Biomedical Image Computing: the development and application of mathematical and computational models

A thesis submitted to the University of Manchester for the degree of Doctor of Science in the Faculty of Medical and Human Sciences

James Graham BSc, PhD

2016

Dedicated to the memory of Thomas Graham, who started my educational journey, but didn't see where it took me.

Table of contents

Abstract

Title: Biomedical Image Computing: the development and application of mathematical and computational models

Submitted to: The University of Manchester

by James Graham

for the degree of Doctor of Science

June 2016

Biomedical images contain a great deal of information that is useful and a great deal that is not. Computational analysis and interpretation of biomedical images involves extraction of some or all of the useful information. The useless information can take the form of unwanted clutter or noise that can obscure the useful information or inhibit the interpretation. Various mathematical and computational processes may be applied to reduce the effects of noise and distracting content. The most successful approaches involve the use of mathematical or computational models that express the properties of the required information. Interpretation of images involves finding objects or structures in the image that match the properties of the model.

This dissertation describes the development and application of different models required for the interpretation of a variety of different image types arising from clinical medicine or biomedical research. These include:

- neural network models,
- Point Distribution Models, and the associated Active Shape Models, which have become part of the research toolkit of many academic and commercial organisations,
- models of the appearance of nerve fibres in noisy confocal microscope images,
- models of pose changes in carpal bones during wrist motion,

A number of different application problem are described, in which variants of these methods have been developed and used:

- cytogenetics,
- proteomics,
- assessing bone quality,
- segmentation of magnetic resonance images,
- measuring nerve fibres
- inferring 3D motion from 2D cinefluoroscopy sequences.

The methods and applications represented here encompass the progression of biomedical image analysis from early developments, where computational power became adequate to the challenges posed by biomedical image data, to recent, highly computationally-intensive methods.

Declaration

The University of Manchester Higher Doctorate Candidate Declaration

Candidate Name: James Graham

Faculty: Medical and Human Sciences

Higher Doctorate title: Doctor of Science

Contribution to Publications

The following is a brief description of my contribution, and that of other co-authors, to each of the publications listed on pages following page 16.

Image Analysis Software Architecture

1. **An architecture for integrating symbolic and numeric image processing.** C.J. Taylor, R.N. Dixon, P.J. Gregory and J. Graham (1986).

   Taylor led the development of "Magiscan" hardware and software. Dixon contributed to both hardware and system software development. My contribution was in designing the practical application and contributing to the design of the data structures and processes that formed the interface between high-level software and machine code. Gregory was the engineer at Joyce-Loebl responsible for the commercial development of the instrument and contributed to the hardware design.

2. **A compact set of image processing primitives and their role in a successful application program.** J. Graham, C.J. Taylor, D.H. Cooper and R.N. Dixon (1986).

   Similar to reference 1, concentrating on the software architecture. Cooper was a software engineer employed at Joyce-Loebl, who contributed to the implementation of the overall software architecture.

3. **System architectures for interactive knowledge-based image interpretation.** C.J. Taylor, J. Graham and D. Cooper (1988).

   Similar to references 1 and 2, this, paper emphasises the requirement for software support for user interaction in biomedical image analysis applications.

4. **Boundary cue operators for model-based image processing.** J. Graham and C.J. Taylor (1988).

   Taylor was PI on the overall project designing model-based approaches to biomedical image analysis of which this work formed a part (also references 5). The reported research was entirely my own.

5. **DEMOB: an object oriented application generator for image processing.** N. Bryson, D. Cooper, J. Graham, D. Pycock, C.J. Taylor and P.W. Woods (1988).

   Part of the same project as 4. Bryson and Cooper developed the object-oriented programming environment. Pycock, Woods and I contributed equally to the system implementation (authors presented in alphabetical order).

6. **User Programmable Visual Inspection.** J. J. Hunter, J. Graham and C. J. Taylor (1995).

   I was PI on this project, which built on the use of models to design a framework for building image analysis applications without the necessity for writing and compiling code. I supervised Hunter's research, with additional input from Taylor. References 21 to 26 also arose from this project.

Chromosome Analysis and Neural Network Models

7. **Automation of routine clinical chromosome analysis I. Karyotyping by machine.** J. Graham (1987).

   This was my own research.

8. **Automation of routine clinical chromosome analysis II: Metaphase finding.** J. Graham and D. Pycock (1987).

   As reference 7. Pycock assisted in coding and performance testing.

9. **The transportation algorithm as an aid to chromosome classification.** MKS Tso and J. Graham (1983).

   This was part of my development of a chromosome analysis system. Tso had expertise in operations research. Approximately equal intellectual contributions from Tso and myself to the algorithm development.

10. **An efficient transportation algorithm for automatic chromosome karyotyping.** M. Tso, P. Kleinschmidt, I. Mitterreiter and J. Graham (1991).

    Kleinschmidt had developed an efficient solution to the bipartite matching problem. Mitterreiter adapted this code to the chromosome classification problem under my direction, with input from Tso.

11. **Resolution of composites in interactive karyotyping.** J. Graham (1989).

    My own research.

12. **Automatic karyotype analysis.** J Graham and J Piper (1994).

    Joint review with Piper (equal contributions) of the methods applied in automated karyotype analysis. Piper was a member of another research group working in this field.

13. **A neural network approach to automatic chromosome classification.** A.M. Jennings and J. Graham (1993).

    Supervised research, contributing to Jennings' dissertation for the MSc by research.

14. **Application of artificial neural networks to chromosome classification.** P.A. Errington and J. Graham (1993).

    Supervised research, contributing to Errington's PhD dissertation.

15. **Classification of chromosomes using a combination of neural networks.** P. A. Errington and J. Graham (1993).

    As reference 14.

16. **Classification of Chromosomes: A comparative study of neural network and statistical approaches.** J Graham and P.A. Errington (2000).

    As references 14 and 15.

17. **A Neural Network Classifier for Chromosome Analysis**. J. Graham (1996).

    Invited contribution to the Handbook on Neural Computing.

18. T**rainable Grey-Level Models for Disentangling Overlapping Chromosomes.** G.C. Charters and J. Graham (1999).

    Supervised research, contributing to Charters' PhD dissertation.

19. **Disentangling Chromosome Overlaps by Combining Trainable Shape Models with Classification Evidence.** G.C. Charters and J. Graham (2002).

    As reference 18.

20. **The application of artificial neural networks to Doppler ultrasound waveforms for the classification of arterial disease.** J. H. Smith, J. Graham and R. J. Taylor (1996).

    Supervised research, contributing to Smith's MSc dissertation. Taylor provided clinical data and input.
    Statistical Models of Shape and Appearance

Statistical Models of Shape and Appearance

21. **Locating overlapping flexible shapes using geometric constraints.** D.H. Cooper, C.J. Taylor, J. Graham and T.F. Cootes (1991).

    I was PI on this project (as references 6, 22 – 26), supervising research. Cooper conducted the study with input from Taylor and Cootes.

22. **Trainable method of parametric shape description.** T.F. Cootes, D.H. Cooper, C.J. Taylor and J. Graham (1992).

    As reference 21. Cootes conducted the study with input from Taylor and Cooper.

23. **Training models of shape from sets of examples.** T.F. Cootes, C.J. Taylor , D.H. Cooper and J. Graham (1992).

    As reference 22.

24. **Active Shape Models - Their training and application.** T.F. Cootes , D.H. Cooper, C.J. Taylor and J. Graham (1995).

    As references 22 and 23.

25. **Building and using flexible models incorporating grey level information.** T.F. Cootes, C.J. Taylor, A. Lanitis, D.H. Cooper and J. Graham (1993).

    As references 22, 23 and 24. Additional input from Lanitis on face analysis.

26. **Image search using trained flexible shape models.** T.F. Cootes, D.H.Cooper, C.J. Taylor and J. Graham (1994).

    As references 22, 23 and 24.

27. **Structured point distribution models: modelling intermittently present features.** M. Rogers and J. Graham (2001).

    Supervised research, contributing to Rogers' PhD dissertation.

28. **Robust Active Shape Model search**. M. Rogers and J. Graham (2002).

    As reference 27.

29. **Detecting asymmetries in hippocampal shape and receptor distribution using statistical appearance models and linear discriminant analysis.** D. Poxton, J. Graham and J.F.W. Deakin, *1998.*

    Supervised Research. Poxton was a PhD student under my supervision. Deakin provided clinical data and input.

30. **An Investigation of morphometric changes in the lateral ventricles of schizophrenic subjects.** K.O. Babalola, J. Graham, W. Honer, L. Kopala, D. Lang and R. Vandorpe (2003.)

    Supervised research, contributing to Babalola's PhD dissertation. Honer, Kopala, Lang and Vandorpe provided clinical input.

31. **Lateral asymmetry in the shape of brain ventricles in control and schizophrenia groups.** J. Graham, K.O. Babalola, W. Honer, L. Kopala, D. Lang and R. Vandorpe (2006).

    As reference 30.

Applications of Image Analysis: Proteomics

32. **Statistical models of shape for the analysis of protein spots in 2-D electrophoresis gel images.** M.D. Rogers, J. Graham and R.P. Tonge (2003).

    Supervised research. Rogers was a postdoctoral researcher who conducted the study. Tonge provided input on Gel Electrophoresis.

33. **Automatic construction of statistical shape models for protein spot analysis in electrophoresis gels.** M. Rogers, J. Graham and R.P. Tonge (2003).

    As reference 32.

34. **Using statistical image models for objective evaluation of 2D gel image analysis.** M.D. Rogers, J. Graham and R.P. Tonge (2003).

    As references 32 and 33.

35. **Robust and accurate registration of 2-D electrophoresis gels using point matching.** M. Rogers and J. Graham (2007).

    Supervised research. Rogers was a postdoctoral researcher who conducted the study under my supervision.

36. **A new paradigm for clinical biomarker discovery and screening with mass spectroscopy through biomedical image analysis principles.** H Liao, E. Moschidis, I Riba-Garcia, Y Zhang, R.D. Unwin, J.S. Morris, J. Graham and A.W. Dowsey (2014).

    Contribution to multi-disciplinary research. Moschidis conducted the image analysis development under my supervision. Liao, Riba-Carcia, Zhang, Unwin and Dowsey provided input on liquid chromatography/mass spectrometry. Morris provided input on machine learning.

Applications of Image Analysis: Assessing Bone Quality

37. **Detecting reduced bone mineral density from dental panoramic radiographs using statistical shape models.** P.D. Allen, J. Graham, D.J.J. Farnell, E. Harrison, R. Jacobs, K. Karayianni, C. Lindh, P.F. van der Stelt, K. Horner and H. Devlin (2007).

    Supervised research . Allen and Farnell were postdoctoral researchers who contributed components of the study. Other authors contributed clinical input.

38. **Automated osteoporosis risk assessment by dentists: a new pathway to diagnosis.** H. Devlin, P.D. Allen, J. Graham, R. Jacobs, K. Karayianni, C. Lindh, P.F. van der Stelt, E. Harrison, J.E. Adams, S. Pavitt and K. Horner (2007).

    As reference 37. Pavitt's contribution was largely organisational

39. **The role of the dental surgeon in detecting osteoporosis: the Osteodent study.** H. Devln, P.D. Allen, J. Graham, R. Jacobs, K. Karayianni, C. Lindh, E. Marjanovic, P.F. van der Stelt, J.E Adams, S. Pavitt and K. Horner (2008).

    As reference 38.

40. **The relationship between the OSTEODENT index and hip fracture risk assessment using FRAX.** K. Horner, P. Allen, J. Graham, R. Jacobs, S. Boonen, S. Pavit, O. Naeckerts, E. Marjanovic, J.E Adams, K. Karayianni, C. Lindh, P. van der Stelt and H. Devlin (2010).

    As references 38, and 39.

41. **Improving the detection of osteoporosis from dental radiographs using active appearance models.** M.G. Roberts, J. Graham and H. Devlin (2010).

    Supervised research . Roberts was a postdoctoral researcher who conducted the study under my supervision. Devlin contributed clinical input.

42. **Changes in mandibular cortical width measurements with age in men and women.** M. Roberts, J. Yuan, J. Graham, R. Jacobs and H. Devlin (2011).

    As reference 41. Jacobs provided Image data. Yuan contributed statistical analysis.

43. **Image texture in dental panoramic radiographs as a potential biomarker of osteoporosis.** M.G. Roberts, J. Graham, H. Devlin (2013).

    As reference 41.

44. **Multi-scale rigid registration to detect damage in micro-CT images of progressively loaded bones.** R. Green, J. Graham and H. Devlin (2011).

    Supervised research, contributing to Green's PhD dissertation. Devlin provided clinical input.

Applications of Image Analysis: Segmentation of the Prostate

45. **Differential segmentation of the prostate in MR images using tissue modelling and 3D Active Shape Models**. P.D. Allen, D. Williamson, J. Graham, and C.E. Hutchinson (2006).

    Supervised research . Allen and Williamson were postdoctoral researchers who contributed components of the study; Hutchinson contributed clinical input.

46. **Automatic differential segmentation of the prostate in 3-D MRI using random forest classification and graph-cuts optimisation.** E. Moschidis and J. Graham (2012).

    Supervised research, contributing to Moschidis' PhD dissertation.

47. **The accuracy of prostate volume measurement from ultrasound images: A quasi-Monte Carlo simulation study using magnetic resonance imaging**. D.-O. Azulay, P. Murphy and J. Graham (2013).

    Supervised research. Azulay conducted the study under my supervision. Murphy contributed to writing the manuscript.

48. **A systematic performance evaluation of interactive image segmentation methods based on simulated user interaction.** E. Moschidis and J. Graham (2010).

    Supervised research, contributing to Moschidis' PhD dissertation.

49. **Propagating segmentation of a single example to similar images: Differential segmentation of the prostate in 3D MRI**. E. Moschidis and J. Graham (2013).

    As reference 48.

Applications of Image Analysis: Diabetic Neuropathy

50. **Application of model based image interpretation methods to diabetic neuropathy.** M. J. Byrne and J. Graham (1996).

    Supervised research, contributing to Byrne's PhD dissertation.

51. **Exploiting weak shape constraints to segment capillary images in microangiopathy.** M. Rogers, J. Graham and R.A. Malik (2000).

    Supervised research, contributing to Rogers' PhD dissertation (as references 27 and 28). Malik provided clinical input.

52. **Dual-model detection of nerve fibres in corneal confocal microscopy images.** M.A. Dabbah, J. Graham, I Petropoulos, M. Tavakoli and R.A. Malik (2010).

    Supervised research. Dabbah was a postdoctoral researcher who conducted the study. Petropoulos, Tavakoli and Malik provided clinical input.

53. **Automatic analysis of diabetic peripheral neuropathy using multi-scale quantitative morphology of nerve fibres in corneal confocal microscopy imaging**. M.A. Dabbah, J. Graham, I.N. Petropoulos, M. Tavakoli, R.A. Malik (2011).

    As reference 52.

54. **An automatic tool for quantification of nerve fibres in corneal confocal microscopy images.** X. Chen, J Graham, M.A. Dabbah, I.N. Petropoulos, M. Tavokoli, R.A. Malik (2016).

    Supervised research. Chen and Dabbah were postdoctoral researchers who developed image analysis methods under my supervision. Other authors provided clinical data and input.

55. **Rapid automated diagnosis of diabetic peripheral neuropathy with in vivo corneal confocal microscopy**. I.N. Petropoulos, U. Alam, H. Fadavi, A. Marshall, O. Asghar, M.A. Dabbah, X. Chen, J. Graham, G. Ponikaris, A.J.M. Boulton, M. Tavakol1, R,A, Malik (2014).

    Supervised research. Chen and Dabbah were postdoctoral researchers who contributed software under my supervision for this clinical study. Other authors provided clinical data and input.

56. **Small nerve fiber quantification in the diagnosis of sensorimotor polyneuropathy: comparing corneal confocal microscopy with intraepidermal nerve fiber density.** X. Chen, J Graham, M.A. Dabbah, I.N. Petropoulos, G. Ponikaris, O. Ashgar, U. Alam. A. Marshall, H. Favadi, M. Ferdousi, S. Azmi, M. Tavokoli, N. Efron, M. Jeziorka, R.A. Malik (2015).

    Supervised research. As reference 55; Jeziorka carried out the intraepidermal nerve-fibre density analysis. This comparative study was conducted by Chen under my supervision.

Applications of Image Analysis: Carpal Kinematics

57. **Inferring 3D kinematics of carpal bones from single-view fluoroscopic sequences.** X. Chen, J. Graham, C.E. Hutchinson, L. Muir (2011).

    Supervised research. Chen was a postdoctoral researcher who conducted the study under my supervision. Hutchinson and Muir provided clinical input.

58. **Integrated framework for simultaneous segmentation and registration of carpal bones.** X Chen, J. Graham, C.E. Hutchinson (2011).

    As reference 57.

59. **Automatic inference and measurement of 3D carpal bone kinematics from single view fluoroscopic sequences.** X Chen, J. Graham, C.E. Hutchinson, L. Muir (2013).

    As references 57 and 58.

60. **Automatic generation of statistical pose and shape models for articulated joints.** X Chen, J. Graham, C.E. Hutchinson, L. Muir (2014).

    As references 57, 58 and 59.

None of the above work has been presented in support of an application for any other degree or qualification at The University of Manchester or any other University or professional or learned body.

I confirm that this is a true statement and that, subject to any comments above, the submission is my own original work.

Signed:  **Date: 25 July 2016**

Copyright Statement

I. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

II. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

III. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

IV. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property University IP Policy (see http://documents.manchester.ac.uk/display.aspx?DocID=24420), in any relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see http://www.library.manchester.ac.uk/about/regulations/) and in The University's policy on Presentation of Theses

Statement

The Candidate

I received the BSc in Physics with first class honours from the University of Edinburgh in 1974. I was awarded the PhD from the University of Cambridge in 1978 following work at the MRC Laboratory of Molecular Biology on the crystal structure of Tobacco Mosaic Virus protein.

I joined the University of Manchester in 1978 as a research associate in the Wolfson Image Analysis Unit, which was part of the department of Medical Biophysics. The remit of the Wolfson Image Analysis Unit (WIAU) was the development of practical applications of computer image analysis that would have sufficient commercial value to allow the group to become self-funding. The WIAU software development was based on a specific hardware architecture (known as Magiscan) that allowed flexible and efficient implementation of image processing and analysis methods at the pixel level. My role in this was the development of image analysis software for clinical cytogenetics. This was one of several software applications under development at the WIAU and the most commercially successful, being marketed by Joyce-Loebl Ltd., who also manufactured the Magiscan. The design constraints in developing practical image analysis software contributed to the design of the underlying software architecture (general purpose data structures and algorithms) that was part of the core technology of the Magiscan instrument. Commercial income from the clinical cytogenetics application resulted in my research being supported by Joyce-Loebl till 1988.

In 1988 I was appointed lecturer in Medical Biophysics with an honorary lectureship in Computer Science. I continued to work on the analysis of cytogenetic images, particularly investigating the use of artificial neural networks for classification of chromosomes and model driven methods for resolving segmentation issues in metaphase and prometaphase images, in particular the resolution of clusters of touching or overlapping chromosomes. I was PI on a project funded by SERC (as it was at the time) to develop a software framework that allowed applications to be generated without the need to write explicit programs. The project (named Visual Inspection System Application Generation Environment – VISAGE) was specifically aimed at industrial inspection. The WIAU had had some experience in industrial inspection, having developed a brake inspection system for Volkswagen AG, to which I had contributed. A central

requirement for application generation was an intuitive method for modelling the shape of objects to be recognised, segmented and measured. It was required to model not only the shape, but also the variation in shape observed across example images. The method that arose from this became known as Active Shape Modelling (ASM). This method proved to be extremely useful in describing the shapes and variations in shapes of components of biomedical images, and the method has been used and developed in a number of directions by my colleagues and others in research groups throughout the world.

My research subsequently explored a variety of applications in image analysis in medicine, biology and in inspection of manufactured objects. This often involved developing and applying the statistical shape modelling method behind ASM, where appropriate and adopting appropriate machine learning methods. In the list of publications that forms the basis of this thesis, these applications are divided into groups: Proteomics, Assessing Bone Quality, Segmentation of the Prostate, Diabetic Neuropathy and Carpal Kinematics.

I was appointed Senior Lecturer in 1992 and Reader in 2011.

List of Publications

Image Analysis Software Architecture

1. **An architecture for integrating symbolic and numeric image processing.** C.J. Taylor, R.N. Dixon, P.J. Gregory and J. Graham, in *"Intermediate - Level Image Processing", M.J.B. Duff (ed.), 1986, Academic Press, London, pp 19 - 34.*
2. **A compact set of image processing primitives and their role in a successful application program.** J. Graham, C.J. Taylor, D.H. Cooper and R.N. Dixon, *Patt. Recog. Lett., 4: 325 - 333, 1986.* doi:10.1016/0167-8655(86)90053-X.
3. **System architectures for interactive knowledge-based image interpretation.** C.J. Taylor, J. Graham and D. Cooper, *Phil. Trans. Roy. Soc., Lond. A324, 451 – 465, 1988.* doi: 10.1098/rsta.1988.0033
4. **Boundary cue operators for model-based image processing.** J. Graham and C.J. Taylor, *Proceedings of the fourth Alvey Vision Conference, Manchester, 1988, pp 59 - 64.* doi:10.5244/C.2.10
5. **DEMOB: an object oriented application generator for image processing.** N. Bryson, D. Cooper, J. Graham, D. Pycock, C.J. Taylor and P.W. Woods, *Proceedings of the fourth Alvey Vision Conference, Manchester, 1988, pp 37 - 44.* doi:10.5244/C.2.7
6. **User Programmable Visual Inspection.** J. J. Hunter, J. Graham and C. J. Taylor, *Image and Vision Computing, 13: 623-628, 1995.* doi:10.1016/0262-8856(95)97287-V

Chromosome Analysis and Neural Network Models

7. **Automation of routine clinical chromosome analysis I. Karyotyping by machine.** J. Graham, *Analyt. Quant. Cytol. Histol., 9: 383 - 390, 1987.*
8. **Automation of routine clinical chromosome analysis II: Metaphase finding.** J. Graham and D. Pycock, *Analyt. Quant. Cytol. Histol., 9: 391 - 397, 1987.*
9. **The transportation algorithm as an aid to chromosome classification.** M. Tso and J. Graham, *Patt. Recog. Lett., 1: 489 - 496, 1983.* doi:10.1016/0167-8655(83)90091-0
10. **An efficient transportation algorithm for automatic chromosome karyotyping.** M. Tso, P. Kleinschmidt, I. Mitterreiter and J. Graham, *Patt. Recog. Lett., 12: 117-126, 1991.* doi:10.1016/0167-8655(91)90057-S
11. **Resolution of composites in interactive karyotyping.** J. Graham, in *"Automation of Cytogenetics", C. Lundsteen and J. Piper (eds.), 1989, Springer-Verlag, Berlin, pp 191 - 203. doi:* 10.3233/978-1-60750-851-9-174
12. **Automatic karyotype analysis.** J. Graham and J. Piper, in *" Chromosome Analysis Protocols", J.R. Gosden (ed), Humana Press inc., Totowa NJ, pp141 - 185, 1994.* doi:10.1385/0-89603-289-2:141
13. **A neural network approach to automatic chromosome classification.** A.M. Jennings and J. Graham, *Phys. Med. Biol. 38: 959-970, 1993.* doi:10.1088/0031-9155/38/7/006

14. **Application of artificial neural networks to chromosome classification.** P.A. Errington and J. Graham, *Cytometry 14: 627-639, 1993.* doi:10.1002/cyto.990140607
15. **Classification of chromosomes using a combination of neural networks.** P. A. Errington and J. Graham, *Proceedings of the IEEE International Conference on Neural Networks, San Francisco, California, 1993, pp 1236-1241.* doi:10.1109/ICNN.1993.298734
16. **Classification of Chromosomes: A comparative study of neural network and statistical approaches.** J. Graham and P.A. Errington in *"Artificial Neural Networks in Biomedicine", P. Lisboa, E. Ifeachor and P. Szczepaniak (eds), Springer, London, pp 259-268, 2000*. doi:10.1007/978-1-4471-0487-2_19
17. **A Neural Network Classifier for Chromosome Analysis**. J. Graham, in *"Handbook of Neural Computation" Chapter G4.3. E. Fiesler and R. Beale (eds.) Oxford University Press and IOP Publishing, 1996.*
18. T**rainable Grey-Level Models for Disentangling Overlapping Chromosomes.** G.C. Charters and J. Graham, *Pattern Recognition, 32: 1335-1349, 1999.* doi:10.1016/S0031-3203(98)00171-X
19. **Disentangling Chromosome Overlaps by Combining Trainable Shape Models with Classification Evidence.** G.C. Charters and J. Graham. *IEEE Trans. Signal Processing 50: 2080-2085, 2002.* doi: 10.1109/TSP.2002.800421
20. **The application of artificial neural networks to Doppler ultrasound waveforms for the classification of arterial disease.** J. H. Smith, J. Graham and R. J. Taylor, *International Journal of Clinical Monitoring and Computing, 13: 85-91, 1996.* doi:10.1007/BF02915843

Statistical Models of Shape and Appearance

21. **Locating overlapping flexible shapes using geometric constraints.** D.H. Cooper, C.J. Taylor, J. Graham and T.F. Cootes, *Proceedings of the British Machine Vision Conference, Glasgow, 1991. Springer-Verlag. pp 185-192.* doi:10.5244/C.5.24
22. **Trainable method of parametric shape description.** T.F. Cootes, D.H. Cooper, C.J. Taylor and J. Graham, *Image and Vision Computing 10: 289-294, 1992.* doi:10.1016/0262-8856(92)90044-4
23. **Training models of shape from sets of examples.** T.F. Cootes, C.J. Taylor , D.H. Cooper and J. Graham, *Proceedings of the British Machine Vision Conference, Leeds, 1992. Springer-Verlag. pp 9-18.* doi:10.5244/C.6.2
24. **Active Shape Models - Their training and application.** T.F. Cootes , D.H. Cooper, C.J. Taylor and J. Graham, *Computer Vision and Image Understanding 61: 38-59, 1995.* doi:10.1006/cviu.1995.1004
25. **Building and using flexible models incorporating grey level information.** T.F. Cootes, C.J. Taylor, A. Lanitis, D.H. Cooper and J. Graham, *Proceedings of the fourth International Conference on Computer Vision, Berlin, 1993, pp 242-246.* doi:10.1109/ICCV.1993.378212
26. **Image search using trained flexible shape models.** T.F. Cootes, D.H. Cooper, C.J. Taylor and J. Graham, *Journal of Applied Statistics, 21 (1-2): 111-139, 1994. doi:* 10.1080/757582971

27. **Structured point distribution models: modelling intermittently present features.** M. Rogers and J. Graham, *Proceedings of the British Machine Vision Conference, University of Manchester, 2001. T.F. Cootes and C.J. Taylor (eds.) BMVA Press, pp 33-41.* doi:10.5244/C.15.5

28. **Robust Active Shape Model search**. M. Rogers and J. Graham, *Proceedings of the seventh European Conference on Computer Vision, Copenhagen, 2002 (vol. 4). A. Heyden, G. Sparr, M. Nielsen, P. Johansen (eds.). Lecture Notes in Computer Science 2353, Springer-Verlag, Berlin. pp 517-530.* doi: 10.1007/3-540-47979-1_35

29. **Detecting asymmetries in hippocampal shape and receptor distribution using statistical appearance models and linear discriminant analysis.** D. Poxton, J. Graham and J.F.W. Deakin, *Proceedings of the British Machine Vision Conference, University of Southampton, 1998. P.H. Lewis and M.S. Nixon (eds.) BMVA Press, pp 525-534.* doi: 10.5244/C.12.

30. **An Investigation of morphometric changes in the lateral ventricles of schizophrenic subjects.** K.O. Babalola, J. Graham, W. Honer, L. Kopala, D. Lang and R. Vandorpe, *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI 2003), Montreal, Canada, 2003,R.E. Ellis and T.M. Peters (eds.) (Lecture Notes in Computer Science 2879) Springer Berlin, pp 521-529.* doi: 10.1007/978-3-540-39903-2_64

31. **Lateral asymmetry in the shape of brain ventricles in control and schizophrenia groups.** J. Graham, K.O. Babalola, W. Honer, L. Kopala, D. Lang and R. Vandorpe, *Proceedings of the IEEE International Symposium on Biomedical Imaging Arlington VA. April 2006. J Kovacevic and E. Meijering, eds. IEEE. pp 414-417.* doi: 10.1109/ISBI.2006.1624941

Applications of Image Analysis: Proteomics

32. **Statistical models of shape for the analysis of protein spots in 2-D electrophoresis gel images.** M.D. Rogers, J. Graham and R.P. Tonge, *Proteomics 3: 879-896, 2003.* doi: 10.1002/pmic.200300421

33. **Automatic construction of statistical shape models for protein spot analysis in electrophoresis gels.** M. Rogers, J. Graham and R.P. Tonge, *Proceedings of the British Machine Vision Conference, University of East Anglia, 2003. R. Harvey and J.A. Bangham (eds.) BMVA Press, pp 369-378.* doi:10.5244/C.17.36

34. **Using statistical image models for objective evaluation of 2D gel image analysis.** M.D. Rogers, J. Graham and R.P. Tonge, *Proteomics 3: 879-886, 2003.* doi: 10.1002/pmic.200300420

35. **Robust and accurate registration of 2-D electrophoresis gels using point matching.** M. Rogers and J. Graham, *IEEE Transactions on Image Processing 16: 624-635, 2007.* doi: 10.1109/TIP.2007.891342

36. **A new paradigm for clinical biomarker discovery and screening with mass spectroscopy through biomedical image analysis principles.** H Liao, E. Moschidis, I Riba-Garcia, Y Zhang, R.D. Unwin, J.S. Morris, J. Graham and A.W. Dowsey, *Proceedings of the IEEE International Symposium on Biomedical Imaging Beijing, China. April 2014, , pp 1332-1335.* doi: 10.1109/ISBI.2014.6868123.

Applications of Image Analysis: Assessing Bone Quality

37. **Detecting reduced bone mineral density from dental panoramic radiographs using statistical shape models.** P.D. Allen, J. Graham, D.J.J. Farnell, E. Harrison, R. Jacobs, K. Karayianni, C. Lindh, P.F. van der Stelt, K. Horner and H. Devlin, *IEEE Transactions on Information Technology in Biomedicine 11(6): 601-610, 2007.* Doi: 10.1109/TITB.2006.888704

38. **Automated osteoporosis risk assessment by dentists: a new pathway to diagnosis.** H. Devlin, P.D. Allen, J. Graham, R. Jacobs, K. Karayianni, C. Lindh, P.F. van der Stelt, E. Harrison, J.E. Adams, S. Pavitt and K. Horner, *Bone 40: 835-442, 2007.* doi: 10.1016/j.bone.2006.10.024

39. **The role of the dental surgeon in detecting osteoporosis: the Osteodent study.** H. Devln, P.D. Allen, J. Graham, R. Jacobs, K. Karayianni, C. Lindh, E. Marjanovic, P.F. van der Stelt, J.E Adams, S. Pavitt and K. Horner, *British Dental Journal, 204: E16*, *2008*. doi:10.1038/sj.bdj.2008.317

40. **The relationship between the OSTEODENT index and hip fracture risk assessment using FRAX.** K. Horner, P. Allen, J. Graham, R. Jacobs, S. Boonen, S. Pavit, O. Naeckerts, E. Marjanovic, J.E Adams, K. Karayianni, C. Lindh, P. van der Stelt and H. Devlin, *Oral Surgery Oral Medicine Oral Pathology Oral Radiology and Endodontology*. *110(2): 243-249, 2010.* doi: 10.1016/j.tripleo.2010.03.035

41. **Improving the detection of osteoporosis from dental radiographs using active appearance models.** M.G. Roberts, J. Graham and H. Devlin, *Proceedings of the IEEE International Symposium on Biomedical Imaging Rotterdam, The Netherlands. April 2010. W. Niessen and E. Meijering, eds. IEEE. pp 440-443.* doi: 10.1109/ISBI.2010.5490314

42. **Changes in mandibular cortical width measurements with age in men and women.** M. G. Roberts, J. Yuan, J. Graham, R. Jacobs and H. Devlin, *Osteoporosis International, 22: 1915-1925, 2011.* doi 10.1007/s001 98-010-1401-3.

43. **Image texture in dental panoramic radiographs as a potential biomarker of osteoporosis.** M.G. Roberts, J. Graham, H. Devlin, *IEEE Trans. Biomedical Engineering, 60(9), 2384 – 2392, 2013*. doi: 10.1109/TBME.2013.2256908

44. **Multi-scale rigid registration to detect damage in micro-CT images of progressively loaded bones.** R. Green, J. Graham and H. Devlin, *Proceedings of the IEEE International Symposium on Biomedical Imaging Chicago IL. April 2011. pp 1231-1234.* doi: 10.1109/ISBI.2011.5872624

Applications of Image Analysis: Segmentation of the Prostate

45. **Differential segmentation of the prostate in MR images using tissue modelling and 3D Active Shape Models.** P.D. Allen, D Williamson, J. Graham, and C.E. Hutchinson, *Proceedings of the IEEE International Symposium on Biomedical Imaging Arlington VA. April 2006. J Kovacevic and E. Meijering, eds. IEEE. pp 410-413.* doi:10.1109/ISBI.2006.1624940

46. **Automatic differential segmentation of the prostate in 3-D MRI using random forest classification and graph-cuts optimisation.** E. Moschidis and J. Graham, *Proceedings of the IEEE International Symposium on Biomedical Imaging Barcelona, Spain. April 2012. pp 1727 – 1730.* doi: 10.1109/ISBI.2012.6235913

47. **The accuracy of prostate volume measurement from ultrasound images: A quasi-Monte Carlo simulation study using magnetic resonance imaging**. D.-O. Azulay, P. Murphy and J. Graham, *Computerised Medical Imaging and Graphics, 37(7), 628-636, 2013.* doi: 10.1016/j.compmedimag.2013.09.001.

48. **A systematic performance evaluation of interactive image segmentation methods based on simulated user interaction.** E. Moschidis and J. Graham, *Proceedings of the IEEE International Symposium on Biomedical Imaging Rotterdam, The Netherlands. April 2010. W. Niessen and E. Meijering, eds. IEEE. pp 928-931.* doi: 10.1109/ISBI.2010.5490139

49. **Propagating segmentation of a single example to similar images: Differential segmentation of the prostate in 3D MRI**. E. Moschidis and J. Graham, in *'Abdomen and Thoracic Imaging: An Engineering & Clinical Perspective', A.S. El BAz,, L. Saba, J. Suri (eds.), Springer Science +Business Media, New York, 2013, Chapter 25, pp 657-684. ISBN 978-1-4614-8498-1.* doi 10.1007/978-1-4614-8498-1_25

Applications of Image Analysis: Diabetic Neuropathy

50. **Application of model based image interpretation methods to diabetic neuropathy.** M. J. Byrne and J. Graham, *Proceedings of the fourth European Conference on Computer Vision, Cambridge, 1996 (vol. 2). B. Buxton and R. Cipolla (eds.). Lecture Notes in Computer Science 1065, Springer-Verlag, Berlin. pp 272-282*. doi: 10.1007/3-540-61123-1_146

51. **Exploiting weak shape constraints to segment capillary images in microangiopathy.** M. Rogers, J. Graham and R.A. Malik, *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI 2000), Pittsburgh, PA, USA, 2000, S.L. Delp, A.M. DiGioia and B. Jaramaz (eds.) (Lecture Notes in Computer Science 1935, Springer, Heidelberg) pp 717-726*. doi: 10.1007/978-3-540-40899-4_74

52. **Dual-model detection of nerve fibres in corneal confocal microscopy images.** M.A. Dabbah, J. Graham, I Petropoulos, M. Tavakoli and R.A. Malik, *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI) 2010, Beijing China, Part 1, T. Jiang, N. Navab, J.P.W. Pluim, M. A. Viergever, eds. (Lecture Notes in Computer Science 6361, Springer, Heidelberg) pp300-307, 2010*. doi: 10.1007/978-3-642-15705-9_37

53. **Automatic analysis of diabetic peripheral neuropathy using multi-scale quantitative morphology of nerve fibres in corneal confocal microscopy imaging**. M.A. Dabbah, J. Graham, I.N. Petropoulos, M. Tavakoli, R.A. Malik, *Med. Image Anal 15(5): 738-747, 2011*. doi:10.1016/j.media.2011.05.016

54. **An automatic tool for quantification of nerve fibres in corneal confocal microscopy images.** X. Chen, J Graham, M.A. Dabbah, I.N. Petropoulos, M. Tavokoli, R.A. Malik, *IEEE Trans. Biomedical Engineering (in press).*

55. **Rapid automated diagnosis of diabetic peripheral neuropathy with in vivo corneal confocal microscopy**. I.N. Petropoulos, U. Alam, H. Fadavi, A. Marshall, O. Asghar, M.A. Dabbah, X. Chen, J. Graham, G. Ponikaris, A.J.M. Boulton, M. Tavakol1, R.A. Malik, *Investigational Ophthalmology and Visual Science, 55, 2071-2078, 2014. doi: 10.1167/iovs.13-13787*

56. **Small nerve fiber quantification in the diagnosis of sensorimotor polyneuropathy: comparing corneal confocal microscopy with intraepidermal nerve fiber density.** X. Chen, J Graham, M.A. Dabbah, I.N. Petropoulos, G. Ponikaris, O. Ashgar, U. Alam. A. Marshall, H. Favadi, M. Ferdousi, S. Azmi, M. Tavokoli, N. Efron, M. Jeziorka and R.A. Malik, *Diabetes Care, 38(6), 1138-1144, 2015. doi: 10.2337/dc14-2422*

Applications of Image Analysis: Carpal Kinematics

57. **Inferring 3D kinematics of carpal bones from single-view fluoroscopic sequences.** X. Chen, J. Graham, C.E. Hutchinson and L. Muir, *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI) 2011 Toronto, Canada. Part 2 (Lecture Notes in Computer Science 6892 Springer, Heidelberg) G. Fichtinger, A. Martel, T. Peters Eds, pp 680-687.* doi: 10.1007/978-3-642-23629-7_83

58. **Integrated framework for simultaneous segmentation and registration of carpal bones.** X. Chen, J. Graham and C.E. Hutchinson *Proceedings of IEEE International Conference on Image Processing, Brussels, September 2011 pp 441-444*. doi: 10.1109/ICIP.2011.6116543

59. **Automatic inference and measurement of 3D carpal bone kinematics from single view fluoroscopic sequences.** X. Chen, J. Graham, C.E. Hutchinson and L. Muir, *IEEE Trans. Medical Imaging, 32(2), 317-328, 2013.* doi: 10.1109/TMI.2012.2226740

60. **Automatic generation of statistical pose and shape models for articulated joints.** X. Chen, J. Graham, C.E. Hutchinson and L. Muir. *IEEE Trans. Medical Imaging, 33(2), 372 – 383, 2014*. doi: 10.1109/TMI.2013.2285503

Summary statement

Introduction

The publications listed in this dissertation are contributions to research in the application of computational image analysis in a number of areas of clinical medicine and biology, with emphasis on the development and use of mathematical or computational models. Among these methods are statistically based models of shape and appearance, now widely used both in academic research and commercial product development. The listed publications represent a selection from my full publication list, intended to present a coherent body of work, notwithstanding the fact that a diverse set of application areas are addressed.

A few papers are invited contributions in edited volumes; otherwise, all items in the list have appeared in peer-reviewed outlets: journals or published proceedings of major international conferences. The conference outlets include Medical Image Computing and Computer-Assisted Intervention (MICCAI), International Conference on Computer Vision (ICCV), European Conference on Computer Vision (ECCV), the IEEE International Symposium on Medical Imaging (ISBI) and the British Machine Vision Conference (BMVC). The last conference began in 1989 to support a growing activity in computer vision in the UK. Acceptance has always been competitive and the conference attracted an increasing international attendance, achieving a status similar to ECCV.

The papers are grouped according to methodology and application area, while following a roughly identifiable chronology. There is considerable crossover between methodology and application, and it would have been possible to divide the list differently. To provide an indication of the reception of these papers, citation counts have been given based on Google Scholar. Taking the Google Scholar "i10 index" as a precedent, citation counts greater than 10 are noted.

Citations of papers forming part of this dissertation (listed above) are cited numerically (e.g. [1]). Citations of other work are cited as Author (date).

Image Analysis Software Architecture.

These papers present my contribution to the early work as a member of the Wolfson Image Analysis Unit (WIAU) in the department of Medical Biophysics at Manchester. The objective of the WIAU was to be self-funding by developing

commercially viable software to address challenging image analysis problems in biomedicine. At the time (late 1970s, early 1980s), image analysis applied to practical problems in biomedicine, and in other fields, generally came in the form of instruments such as the Quantimet series from Cambridge Instruments (Pingel and Jenkinson (2016)), where the image processing and analysis were implemented as hardware modules. Image analysis software running on general-purpose hardware was a relatively new notion. While there was an early body of research in Pattern Recognition applied to several fields (e.g. Ledley *et al.* (1965), Rutovitz (1966), Mayall (1974)) the limitations of computer systems meant that none of these was close to providing practical contributions to clinical, laboratory or industrial procedures. WIAU research, led by C.J. Taylor, sought to develop a computer platform supporting a flexible software architecture, within which a range of specific application solutions could be programmed to run with sufficient efficiency to be realistically useable in a clinical or research environment. The first hardware design (which became – accidentally – known as the "Magiscan") was based around a minicomputer with a small, dedicated image memory. The second generation (Magiscan 2) was microprocessor-based, providing a more realistic vehicle for implementing computationally intensive image processing operations. The software architecture was designed to allow challenging application programmes, such as the analysis of microscope images of chromosomes, coded in high-level computer language, to have efficient access to low-level image processing operations. Papers [1-3] represent collaborative work exploring how the requirements of efficient application software influence generic software architecture. The overall hardware and software architecture is described in [1], while [2] describes the intermediate level data structures and processes, forming the interface between high-level (application) and low-level (pixel) processing. User interaction is an important requirement in practical image analysis systems, and [3] considers the architectural issues of combining user interaction with application knowledge (in the form of high-level programmes). Papers [2] and [3] , addressing a fairly narrow topic, have 10 citations each.

The commercial aims of the WIAU were conducted via a collaboration with Joyce-Loebl Ltd (JL), who became the vehicle for marketing the hardware and image analysis software. One such development was the installation of a brake assembly inspection system for Volkswagen AG (unpublished, though some aspects are

described in Woods *et al.* (1987). Development of this system highlighted the need for explicit representation of domain knowledge in the form of mathematical and computational models. The remaining papers in this section relate to investigations of how far the use of explicit models, encapsulating the expected spatial relationships between image components, can be used to build application solutions without the need for writing and compiling code in high-level computer languages. The prototype system described in [5] was an experimental demonstrator using Object-Oriented Programming (OOP), which was attractive for image analysis because of the facility in representing and displaying image components, particularly for interaction. OOP, although discussed in the Artificial Intelligence community since the 1960s, became more widely known as a programming methodology in the 1980s. Language support for OOP was not well developed and the object-oriented language used in [5] was developed in-house, based on C.

Image analysis often requires the location of edges or boundaries between objects, which was (and still is) usually achieved by the use of linear differential operators (e.g. Canny (1986)). The use of specific models of geometrical relationships provides the ability to use spatially targeted non-linear operators as described in [4], which was shown to be significantly more sensitive than Canny edge detection. A further development of the application generator is described in [6]. Here C++ replaced the in-house language of [5]. The paper demonstrates that a high-level description of a complex application can be achieved using generic models (described later in [22-26]) and a small set of geometric data structures and processes.

Papers [4] and [5] were published in the proceedings of the fourth Alvey Vision Conference. These conferences were precursors of the British Machine Vision Conference (BMVC), which became an important international conference rivalling more clearly established events such as the European Computer Vision Conference (ECCV). A number of later papers were published in BMVC proceedings. Papers [4] and [6] have respectively 15 and 10 citations.

The technology transfer process between WIAU and JL was used as a case study by Benneworth (2001), exploring the nature of academic "commercialisation" activities. While containing some errors of detail, this analysis captures the main issues regarding the technology transfer relationship. It notes that during the

collaboration Joyce-Loebl moved from being a company focussed on analog instrumentation to being an entirely digital organisation. JL was later purchased on the basis mainly of the chromosome analysis application, along with a competitor, Image Recognition Systems, by Applied Imaging International of Sunderland (see below, under Chromosome Analysis). Subsequently the digital technology transferred to JL formed the basis of several new digital companies in the North East of England.

Chromosome Analysis and Neural Network Models

Analysis of chromosomes, which become visible in high magnification microscope images of cells at metaphase or prophase, is important in a number of clinical and research areas. The most widely known task is karyotyping, in which certain genetic disorders can be identified by visual inspection of the chromosomes. Such disorders may be manifest as an abnormal number of chromosomes, or insertions or deletions of genetic material from specific chromosomes. The 46 human chromosomes can be assigned to 22 homologous pairs plus the sex chromosomes: XX (female), XY (male). Chromosomes are made visible by staining. With appropriate pre-treatment the stain can produce a sequence of dark and light bands along the chromosome, which, together with the size and the position of the centromere (a characteristic constriction in the chromosome's width), allows each of the 24 classes of chromosomes to be identified. Prior to the advent of banding the size and centromere position could be used to assign the chromosomes to seven groups (A-G) with different numbers of chromosomes (defined by the Denver Conference (1960). The classification rules for identifying banded chromosomes were specified in the Paris conference (1975).

Karyotyping for pre-natal genetic screening requires metaphase cells to be found in amniotic fluid samples. Other application areas are in post-natal diagnosis of genetic disorders, cancer diagnosis and aberration scoring, where chromosomal structural and numerical abnormalities are used to quantify radiation exposure. Most cells at metaphase are not suitable for visual analysis due to poor separation of chromosomes, poor staining, or other difficulties. In all of these applications there is a requirement for metaphase finding: location of the position of dividing (metaphase or prophase) cells for analysis. In amniotic fluid samples, cells of analysable quality are sparse; this is even more the case in tumour samples. In

aberration scoring, sufficient numbers of cells are required for appropriate statistical analysis.

Automation of chromosome analysis became a challenging target application of the growing field of Pattern Recognition in the 1970s (Ledley *et al.* (1972), Castleman *et al.* (1976), Granlund *et al.* (1976), Brenner *et al.* (1976)). Early work by Castleman and his colleagues at NASA's Jet Propulsion Laboratory (Castleman and Melnyck (1976)) sought to develop a practical system for pre-natal karyotyping. In addition, automated metaphase finding was seen as an important requirement (e.g. Johnson and Goforth (1974), Wald *et al.* (1976), Schoevaert-Brossault *et al.* (1983), Shippey *et al.* (1986)). The early developments did not result in practical instantiations of the proposed systems. The available technology did not permit sufficiently flexible algorithms to be applied or to cope with the image digitisation, storage and display requirements. Automated karyotyping became one of the first target applications of the WIAU, and was my responsibility.

The early papers in this section [7, 8] describe the components of the automated chromosome analysis system (karyotyping and metaphase finding) developed by me as a commercial product for routine use in a clinical laboratory. The first installation of the system was in Rigshospitalet, Copenhagen. While the installation was a commercial contract with our partners, JL, the working system resulted from my close collaboration with clinicians in that group, who had been interested in automation of chromosome analysis for some time (Lundsteen *et al.* (1976)). This involved the development of appropriate modes of interaction to enable efficient resolution of segmentation and classification problems in the system, and adaptation of laboratory routines to make best use of the automated analysis.

The key components in the analysis of chromosome images are segmentation of individual chromosomes, representation of the banding pattern and other features, and classification. The details of these components are outlined in some detail in [12], which was written for the education of clinical cytogeneticists, as automated systems were becoming available, though not yet widely used, at the time. Papers [7], [8] appeared in Analytical and Quantitative Cytology and Histology, which was one of the main outlets for image analysis applied to microscope images at the time. They have respectively 42 and 16 citations. Paper [12] was an invited contribution to an edited book and has 41 citations.

The segmentation of chromosomes is made challenging in the case of G-banded chromosomes, as the local contrast between chromosome pixels and background pixels can be very low or vanish completely. Chromosomes can also touch or overlap, and some form of user interaction is ultimately required to resolve some situations. Several authors proposed methods for separating touching chromosomes based on outline curvature (Vossepoel (1989), Wu *et al.* (1989), Agam and Dinstein (1997)). An efficient method for resolving groups of touching chromosomes, based on region growing following a minimal and natural user interaction is described in [11] (10 citations).

Analysis of chromosomes at prometaphase or prophase has the advantage that the chromosomes are longer and the banding pattern provides much higher resolution detail. However chromosome overlaps occur more frequently. Ji (1989) described a heuristic contour analysis method that sought to separate overlaps as well as touching chromosomes. A model-based approach to resolving clusters of overlapping chromosomes is described in [18 and 19], developed from the statistical shape models described below [22-26]. The shape descriptors in [19] model the shape of the chromosome centre lines, rather than object boundaries, as in [22-26]. Similar models of partial chromosome density profiles, inspired by the notion of "unique band sequences" (Lockwood *et al.* (1988)), were used to provide classification evidence for disambiguating the components of an overlapping cluster in [18]. Papers [18] and [19] have 30 and 26 citations respectively.

The features used for classification are, for "Denver" classification, the chromosome length and centromeric index (the fractional distance along the chromosome of the centromere position), and the banding pattern for "Paris" classification. The representation of the banding pattern is an important issue and a number of proposals were made for this (e.g. Groen *et al.* (1989), Granum (1982), Granlund (1976), Habbema (1979), Lundsteen and Granum (1979) and others, reviewed by Carothers and Piper (1994)).

A useful contextual constraint on the assignment of chromosomes to classes arises from the predetermined class sizes (between two and twelve in the A-G "Denver" groups and pairs, with the exception of the male sex chromosomes, in the "Paris" classification). Piper (1986) evaluated several variants of a heuristic method of applying this constraint, originally proposed by Rutovitz (1977), later developing a method using a genetic algorithm (Piper (1995)). In [9] it is shown that a globally

optimum classification subject to the constraint can be found directly by casting it as a "transportation" problem, for which a solution is known in Operations Research. This was applied in [9] to unbanded ("Denver") classification. In the case of banded chromosomes, where the group size is at most two, the problem can be expressed as an "assignment" problem. Kleinschmidt *et al.* (1987) had developed an efficient solution to this problem, which is applied to banded chromosome classification in [10]. (Papers [9] and [10] have respectively 28 and 40 citations). This method was later used for matching points on the surface of brain ventricles [30].

Artificial Neural Network (ANN) models were adopted in a number of applications in the late 1980s and early 1990s. A preliminary study [13] explored the classification of the chromosome banding pattern using a multi-layer perceptron (MLP) and a Kohonen self-organising map. Useful classification using an MLP led to a more extensive study [14], which also proposed a decomposition of the network that allowed fusion of the different feature types (banding density features and morphology features). A further development introduced a network model to implement the class size constraint to achieve improved performance [15]. In [16] the performance of the MLP classifiers is compared directly with the maximum likelihood classifier, and shown to deliver some real improvement. The availability in this study of a very large set of banding data (collected in Copenhagen using the automated karyotyping system installed there) allowed a further experiment exploring the effect of training and test set size on classifier performance. Piper (1992) also used this data in a study of classifier bias. A summary description of the MLP approach was invited to appear in the Handbook of Neural Networks [17], which surveyed a wide range of techniques and applications. Papers [13], [14] and [15] have respectively 20, 52 and 16 citations.

The installation by JL of automated karyotyping and metaphase finding at Righospitalet, Copenhagen, was the first such system to be used routinely in a clinical laboratory. Reports on the clinical experience of its use in two early installations in Copenhagen and Chicago, can be found in Lundsteen *et al.* (1987) and Lundsteen and Martin (1989), including considerations of the economic benefit of the automated system. Subsequently other companies provided systems for installation in cytogenetic laboratories, making use of the increasing computational capacity of desktop computers. These often arose from academic

research, such as Image Recognition Systems from the work at the Medical Research Council, led by D. Rutovitz, and Perceptive Systems International from the early work of Castleman at NASA's Jet Propulsion Laboratory. The Athena system (van Vliet *et al.* (1990) Mayall *et al.* (1990)) was based on the relatively new Macintosh computer and later commercialised by Amoco. Joce-Loebl and Image Recognition Systems were both bought be Applied Imaging International, who produced a PC-based karyotyping system. While desktop computers had sufficient power, particularly in image display, for the karyotyping task, the intensive processing necessary for efficient metaphase finding was more problematic, and this was not initially offered. I witnessed the original Magiscan metaphase finder working alongside Applied Imaging's PC-based karyotyping system in Copenhagen in 2002, twenty years after its initial installation. Automated interactive karyotyping systems (and metaphase finders) are now provided as standard components of cytogenetics imaging systems provided by several suppliers, along with imaging support for more recent staining techniques based on DNA hybridisation.

Classification of chromosomes and separation of clusters of overlapping banded chromosomes continued, and continues, to be a challenging pattern recognition problem. Lerner *et al.* (1995) also described an MLP network for classification, subsequently proposing a neural network approach to segmentation as well (Lerner (1998)). A number of recent publications address these areas (e.g. Moradi and Setarehdan (2006), Kao *et al.* (2008), Wang *et al.* (2009), Vaidyanathan *et al.* (2009)). Indeed some of the images from my original publications in this area [18, 19] still appear among this literature!

Apart from the chromosome classification problem, the MLP model was also applied to classification of Doppler ultrasound signals [20] (29 citations). Neural network models also appear in later publications in a different application area [52-54].

Statistical Models of Shape and Appearance

The use of geometric models for identification of rigid objects in predictable locations had been well recognised (e.g. Chin and Dyer (1986)). Even for man-made objects, appearance and location are subject to some variability, and robustness can be enhanced by incorporating statistics of the observed variation

among examples (Woods et al. (1987)). Paper [6] showed that, for highly geometrically constrained systems, such as mechanical assemblies, it was possible to describe complex inspection tasks using a very high level syntax and a fairly limited number of image processing and analysis routines under the guidance of overall geometric models. I was Principal Investigator on a project to investigate the modelling and syntax required of such a system in the field of industrial inspection. Paper [6] was an output of this project. Since even man-made objects exhibit variation in shape and appearance, a key requirement was flexible statistical geometric modelling. A number of approaches to flexible modelling had been described (e.g. Bookstein (1989), Kass *et al.* (1987), Yuille *et al.* (1992), Staib and Duncan (1992); for a comprehensive review at the time, see McInerney and Terzopoulos (1996)). These approaches allowed flexibility, but did not allow the deformations to be made specific to a particular class of objects. Papers [21-26] show how this was achieved by making the representation trainable on a set of example images, resulting in the Point Distribution Model (PDM), which formed the basis of an image segmentation technique known as Active Shape Modelling (ASM). Paper [21] describes how a precursor to the PDM – the Chord Length Distribution could be used for locating trained shapes in a cluttered environment, including overlapping instances. This method involved finding a maximum overall probability of a set of boundary points subject to a distribution of the lengths of chords joining these points observed during training. Though computationally highly expensive, this approach indicated the representational power of a set of boundary points and the statistics of their relative positions in describing flexible shapes. In [22] a much more compact description of shape was described, using principal component analysis (PCA) of the covariance matrix relating pairs of chords. Due to non-linear correlations among the chords, the method was capable of generating unrealistic shapes, and the computational complexity of the shape reconstruction was still $O(n^2)$, n being the number of points on the boundary. Point Distribution Models were first described in [23], where the representation was changed to the positions of the points themselves, rather than the chords joining them. This had the effect of reducing the complexity to $O(n)$, and also provided a much more direct representation of the shape. The covariances now represented the variations in the relative positions of individual points and the "modes of variation", represented by the eigenvectors of the covariance matrix, conformed much more to the requirement of being independent. The examples

used in [23] reflect the overall context to the development of these models in industrial inspection. Training of PDMs, along with the limitations of the linear modelling framework are described in detail in [24], which also introduces the idea of using the model to conduct image search for the boundary of modelled objects in unseen images. As described in this paper, the search algorithm seeks local points of high image gradient. The final shape is constrained to lie within the distribution of shapes seen in the original data set, reducing the chance of poor segmentations due to finding spurious high-gradient points. The term "Active Shape Model" (ASM) for the use of PDMs in image search is introduced in [24] expressing the similarity to the "Active Contour Models" of Kass et al. (1987). The only shape constraint in this latter technique is an "internal energy" term, optimised alongside the image gradient on the boundary to encourage local boundary smoothness.

In [25], and more fully in [26], the eigenvector-based modelling technique was extended to include local grey-level descriptions. These took the form of one-dimensional profiles of grey-level derivatives at each point, normal to the local boundary. The local search now found the position where the observed profile in the image best matched the stored local model. A more careful search strategy was also introduced, in which large movements at each iteration were penalised. Three components of the ASM made it, by design, more robust than the Active Contour Model, i.e. less likely to be trapped on incorrect shapes because of locally confusing image evidence. These were: an explicit grey level model of the local image region to be identified; disallowing excessively large movements in the direction proposed by the profile search; and having a highly constrained shape description that prevents poor local fits from generating badly distorted shapes. Active Contour Models were often referred to as "snakes". The fact that the ASM acted in a similar way, but constrained to produce appropriate shapes led us (with perhaps a touch of hubris) to refer to ASMs as "smart snakes".

The remaining papers in this section represent developments of the PDM/ASM method in response to the requirements of specific problems in biomedical image analysis. Paper [27] arises from issues in the analysis of electron micrographs of capillaries in peripheral nerves. Further discussion of this particular application can be found below, referring to papers [50] and [51]. In brief the requirement is to segment images of the cross-section of the capillaries into three regions: the

basement membrane, the endothelial cell layer and the lumen, through which the blood passes. In cases of interest, the lumen becomes constricted and can close completely. To deal with the problem of image components that may be absent in some instances, a modification of the shape modelling procedure was proposed in [27] to include a binary present/not present condition, extending the method to other examples. It was also important for some applications to make the ASM search more robust. The implicit assumption in "standard" ASM fitting that residuals follow a Gaussian distribution is only at best approximately true. In [28] we investigated the use of robust fitting methods, including RANSAC (Fischler and Bolles (1981)) and M-estimators (Huber (1981)), which showed improved accuracy and robustness in several image segmentation problems (though not, as it happened, in the capillary images).

Point Distribution models are built by manual annotation of a consistent set of points around the boundaries of the objects to be modelled in a training set of images. It is important to maintain correspondence between equivalent points across the training set. The ordering constraint on points around a 2D boundary makes correspondence between points straightforward. In principle, the shape modelling and search algorithms developed on 2D images in papers [22-26] can be applied, more or less unaltered, to images of higher dimension. However, in 3D shapes, the annotation problem is significantly more challenging because important "landmark" points are much more difficult to locate accurately and the ordering constraint no longer applies. Papers [30, 31] describe an application in using the parameters of the trained shape model as shape descriptors to investigate possible changes in the shape of brain ventricles between control and schizophrenia groups. This was an early application of shape modelling in 3D structures, and it was necessary to solve the annotation/correspondence problem, as manually annotating the ventricle surface on the 69 3D Magnetic Resonance images was an impossible task. We made use of the fact that a number of consistent ridges occur on the ventricle surface. The correspondence between equivalent points on these ridges was established by expressing the spatial correspondence as a bipartite graph-matching problem, making use of the same matching algorithm that had been used in the chromosome-matching problem [10]. The shape parameters derived from the resulting PDMs were used to define a discriminating shape vector in the space of the shape parameters and hence

identify and quantify the shape differences between the groups [30]. In [31] a similar analysis was used to measure lateral differences between right and left ventricles in male and female subgroups. Subsequently, a more generic approach to the problem of finding point correspondences by global optimisation was proposed, initially in Davies *et al.* (2003) and later more fully in Davies *et al.* (2010).

Paper [29] is another example of using the model parameters as features for discriminating between groups. In this case the images were 2D images: autoradiographs of radiolabelled sections of hippocampal tissue. The study was intended to identify differences in both shape and the spatial distribution of 5-HT1A receptors between right and left hippocampi. The 2D shape parameters were used in the same way as the 3D parameters in the brain ventricle study. By warping all of the shapes onto a mean shape to achieve correspondence between pixels in different images, the spatial distribution of grey-levels, and hence receptors, could be modelled using PCA. The use of PCA on grey-level distributions was taken further by Cootes *et al.* (2001), where a further PCA of a combined shape and grey-level vector was used as the basis for Active Appearance Models: a segmentation technique similar to ASM search, making use of the much richer image description in the complete grey-level model.

ASMs, with a variety of modifications have been used in many applications of biomedical image analysis. (For a review of 3D applications, see Heimann and Meinzer (2009).) Paper [24] is among the most highly cited papers in computer vision (6900), [22], [23] and [25] also have high citation counts (227, 620, 148). [26] has a more modest citation count of 21, despite being the most complete description of the ASM method, while [28] has 149.

Applications of Image Analysis

The following sections describe a number of image analysis application areas in which I have applied statistical and computational models, including ASM and variants, neural network models and more recent machine learning methods, such as random forests (Breiman (2001)).

Proteomics

Papers [32-36] are related mainly to the analysis of 2D electrophoresis gels, which

have been a standard tool in proteomics for many years. In this method proteins or peptides in a mixture are separated according to their molecular weight in one dimension and surface charge on the other to form "spots" of varying density on the gel, visualised using either radioactive or fluorescent ligands. Software tools, developed by academic groups or commercial organisations, were widely used for measurement of the positions and intensities of spots. These methods involved data-driven segmentation techniques (e.g. Lemkin and Lipkin (1981), Garrels (1989), Smilansky (2001)) or modelling individual protein spots as bivariate Gaussian density distributions (Garrels (1989), Anderson *et al.* (1981)). Spot segmentation often results in missing, or failing to separate, faint spots. Paper [32] approached the spot description and measurement problem using a PDM-based shape descriptor to take account of the fact that non-elliptical spots with non-Gaussian density profiles can occur.

A method for automatically generating the PDMs from the hundreds (or thousands) of protein spots present on a gel is described in [33], involving an automatic segmentation of spots in training images, using robust PCA to reject unrealistic spot shapes resulting from mis-segmentations.

Quantitative analysis and comparison of tools for analysing electrophoresis gels requires reliable "ground-truth" against which their measurements could be assessed. While it is possible to produce synthetic gel images with known density distributions, these do not exhibit the complexity of overlapping spots and distorted spot shapes that arise in real gels. The spot modelling method of [32] allowed generation of spots from a distribution of spot parameters, placing these at positions determined from real gels to generate synthetic gels with precisely known characteristics [34]. By varying parameters such as noise level and spot overlap the properties of gel analysis software packages could be investigated. Papers [32] and [34], appearing in Proteomics, one of the major journals dealing with gel electrophoresis, have 63 and 52 citations respectively.

Gel analysis often consists of comparison of the protein compositions of different sample groups. Typically several gels would be run in each group and specific spots identified and compared within and between groups. This process of identifying corresponding spots between gels is a registration problem, made difficult by sometimes severe spatial distortions introduced during the electrophoresis process. Such "non-rigid" registration is a problem commonly

addressed in medical image analysis (see review by Maintz and Viergever (1998)). The gel registration problem has its own challenges that are different from those in the medical image registration case. There are many potential individual matches; the geometric distortions between gels can be rather large; there may be many tens of images to be registered as a group; there are many "unmatchable" spots, which may appear in some images and not in others, either because the proteins they represent are reduced or absent in some samples, or because they are not detected at the segmentation stage. The method described in [35] is based on the Iterated Closest Point (ICP) algorithm (Besl and Mckay (1992)), an algorithm widely reviled because of its lack of theoretical foundation, and widely used because of its reliable convergence properties in many cases. In this case ICP was augmented by a non-Euclidean distance metric and robust estimation of transform parameters to produce very reliable and accurate registration. Evaluation in the presence of increasing levels of distortion and noise showed that the method was highly robust and outperformed other, well-regarded point-matching methods such as SoftAssign (Chui *et al.* (2004)). Appearing in IEEE Transactions on Image Processing, [35] has 49 citations.

One of the scientific drawbacks of 2D gel analysis is limited sensitivity. Recently, more sensitive methods, for detection of small peptide changes have become increasingly used. Liquid Chromatography Mass Spectrometry (LCMS) is one such method. The output has similarities to gel electrophoresis: a two-dimensional data distribution with a requirement for non-rigid registration between the very large data sets produced. Paper [36] gives an initial description of the application of the registration method in [35] applied within a system for detection and recognition of peptide fragments within an LCMS analysis.

Assessing bone quality

Osteoporosis is a condition of reduced bone mass and microarchitectural deterioration of the bone, leading to increased fragility and fracture risk. The standard method for assessment of bone mineral density (BMD) is by dual energy X-ray absorptiometry (DXA), which provides an absolute measurement of bone mineral density. It is usually measured at the neck of femur, pelvis, wrist or spine, where there is the greatest risk of fracture due to reduced bone strength. Screening for osteoporosis is not considered cost effective, despite it representing

a large healthcare burden, with significant associated morbidity and mortality (Johnell (1996)). A number of factors may be used as predictors of risk of developing osteoporosis. These include age, body mass index, treatment with hormone replacement therapy and family history of osteoporosis. Several clinical indices have been proposed composed of differing weighted combinations of these factors, e.g. Sedrine *et al.* (2002), Cadarette *et al.* (1999).

Panoramic dental radiographs are tomographic images of the total mandible and maxilla (see images in papers [37 – 41]), which are taken for several purposes in dental care. In addition to the teeth and trabecular bone, the images also show a region of denser cortical bone at the lower border of the mandible (the inferior mandibular cortex). Several authors had observed that the apparent thickness of this cortical bone is related to BMD, not only in the mandible but also at the other BMD measurement sites, e.g. Taguchi *et al.* (1996), Taguchi *et al.* (2006), Klemetti and Kolmakow (1997), Horner and Devlin (1998), White *et al.* (2005), leading to the suggestion that making this measurement on dental panoramic radiographs could be an opportunistic way of identifying individuals at risk of fracture due to reduced BMD. Making this measurement manually would be impractical; to be useful, it would be necessary to implement an automatic method

Papers [37-40] arose from a large European collaborative study, called OSTEODENT, to investigate the potential for using cortical width measurement as a case-finding mechanism for osteoporosis. Images and clinical input came from partners in Stockholm, Leuven, Amsterdam and Athens as well as Manchester. The automatic system for measuring cortical width is described in [37]. The key point is the reliable location of the lower and upper boundaries of the inferior mandibular cortex, which is achieved by ASM search, allowing the cortical thickness to be measured at appropriate locations. The OSTEODENT study collected a very large set of evaluation images (separate from the training images used in [37]) from 670 female patients between 45 and 75 years of age, each with ground-truth measurements of BMD taken at the femoral neck, hip and spine. The analysis demonstrating the effectiveness of cortical width measurement in identifying osteoporosis is presented in [38], concluding that automatic analysis of dental panoramic radiographs could be used as a suitable triage method for identifying patients who should be referred for further DXA investigation.

In paper [39], the effectiveness of the radiographic measurement is compared with

one of the risk indices (OSIRIS, Sedrine et al. (2002)). The paper observes that combing the radiographic index with OSIRIS using logistic regression resulted in improved prediction of osteoporosis over either method alone. The combined index came to be called the "Osteodent index".

The diagnosis of osteoporosis is entirely based on BMD. However, the clinical burden arises from fractures due to weakened bones. More recently the FRAX tool (Kanis *et al.* (2008)) has been developed specifically for the prediction of 10-year fracture risk. In [40] the Osteodent index is compared with FRAX as a means of recommending patients for further investigation by DXA. The two methods were found to be equivalent.

Papers [37 – 40] have citation counts of 44, 68, 34 and 24 respectively.

In [41] some aspects of the ASM search on the cortex were improved by including a larger number of image features in the model, described as a hybrid of Active Shape and Active Appearance modelling, increasing the reliability of the fit.

This improved search was used in [42] to examine the relationship between measured cortical width and age in both men and women. This study made use of a very large (close to 5000) set of images supplied by one of the OSTEODENT collaborators, representing patients between 15 and 94 years old. The final association between cortical width and age was very similar to that between BMD and age.

It is well known that reduced bone quality can be assessed by the appearance of the bone observed in panoramic radiographs. A semi-quantitative index, the Mandibular Cortical Index, has been developed to allow radiologists to report the appearance of "holes" and "residues" in the cortical and trabecular bone (Klemetti *et al.* (1994)). A number of authors had reported associations between texture measures (principally fractal dimension), applied to the cortical and trabecular bone in the mandible, and BMD (e.g. Ruttimann *et al.* (1992), Yasar and Akgunlu (2006), Geraets and van der Stelt (2000)). It seems likely that "fractal dimension" is applied here as a generalised "roughness" measure. In [43] we tried to deal systematically with the issue of relating radiographic texture to BMD. There are many texture measures that can be applied to image regions (see, for example, Petrou and Carcia Sevilla (2006)). We chose to make use of the classic co-occurrence matrices (Haralick (1973)), because a large number of features,

corresponding to a range of texture appearance, can be calculated in a single set of image measurements. We selected a number of co-occurrence features, combined in a Random Forest classifier Breiman (2001). As fractal dimension had appeared in several places in the literature we included an appropriate measure of fractal dimension. Texture measured in the cortical bone was shown to have a similar association with BMD as cortical width and that the combination of texture and cortical width provided better association than either alone. Fractal dimension did not perform as well as some other texture methods. Papers [41 – 43] have 14, 28 and 11 citations respectively.

Paper [44] reports part of a related study that sought to determine trabecular features related to bone strength. Using micro-CT images of progressively loaded rat vertebrae, sites of damage are located by registration of successive images at different scales. Small-scale damage can result in large-scale changes in the shape of the vertebrae. The successive registration allows the damage sites to be identified by measurement of unregistered voxels. The sensitivity and specificity of damage detection could be estimated using synthesized damage in real images.

Segmentation of the prostate

Papers [45-46] are concerned with differential segmentation of the prostate gland in magnetic resonance (MR) images. The clinical context is the diagnosis of benign prostatic hyperplasia (BPH). Clinical investigation of the prostate is usually conducted using trans-rectal ultrasound (TRUS). Knoll *et al.* (1999) described segmentation of the outer boundary in CT and TRUS images using a constrained Active Contour model. However in BPH diagnosis it is important to identify the main anatomical zones of the gland: central zone, transition zone and peripheral zone (Tewari *et al.* (1995)), which do not show clearly in TRUS or CT images. The central and transitional zones are often combined to form the "central gland". Zwiggelaar *et al.* (2003) described segmentation of the outer prostate boundary in MR images in the context of prostate cancer, but did not address the differentiation of zones. Challenges in segmentation of the MR images arise from the similarity in appearance of the different zones, with no clear boundary between them, and the presence of nearby anatomical structures, mainly the bladder and seminal vesicles. In [45] a 3D PDM was used to describe the shapes of the central gland and peripheral zone. Since boundaries between zones are not distinct, standard ASM

search was not appropriate. A Gaussian mixture model was used to perform voxel classification, while the PDM applied shape constraints in a genetic algorithm optimisation. In [46] a similar approach was taken, using random forest classification to classify voxels followed by graph-cut optimisation. Paper [45] has 26 citations.

The trained prostate shape models used in [45], provided the basis for the study in [47] examining the accuracy and reliability of measuring prostate volumes using TRUS. The usual procedure in estimating prostate volume in ultrasound images involves measuring the prostate diameter in approximately orthogonal 2D views and calculating the volume based on an assumption of ellipsoidal shape. The annotated shapes used to build the models in [45] provided ground truth, which was used to quantify errors in the ellipsoidal volume estimation process, based on a quasi Monte Carlo analysis (quasi Monte Carlo because the synthesised images did not come from a truly random distribution). The study was able to propose a linear regression model, based on measured prostate diameters, that is more accurate than using directly calculated ellipsoidal volumes.

The prostate images were used as one of three challenging segmentation problems in a study to examine suitable algorithms to form the basis of interactive segmentation for 3D PDM model building. Image segmentation is a necessary prior step to creating sets of corresponding surface points, using the method in [29] or the more generic method of Davies et al. (2010). This segmentation would normally be interactive, resulting in a potentially long and tedious task involving multiple images. The study sought to determine methods for approaching this task efficiently. Paper [48] compared the use of three algorithms that had previously been proposed for interactive segmentation: graph-cut (Boykov and Jolly (2001)), grow-cut (Vezhnevets and Konouchine (2005)) and random-walker (Grady (2006)), using a simulated interaction framework to segment prostate and brain images. Graph-cut came out best from the comparison, and was used in further studies, including propagation of segmentation results from one image onto further examples [49]. The automated prostate segmentation method in [46] arose from this study.

An overview of subsequent research in image analysis of the prostate in MR, CT and ultrasound images in the context of diagnosis and treatment of prostate cancer can be found in Madabhushi *et al.* (2011) and Ghose *et al.* (2012).

Diabetic neuropathy

Peripheral neuropathy is a serious and widespread complication of diabetes (and some other conditions) with potentially severe clinical consequences. Papers [50-56] represent different studies involving images related to investigation of, or diagnosis of, peripheral neuropathy. Paper [50] describes the analysis of light microscope images of cross-sections of fibre bundles and electron microscope images of capillary vessels associated with nerve fibres. The constriction of the capillaries in diabetes causes a condition called microangiopathy. The segmentation in these two image types made use of active contour models (Kass et al. (1987)). The electron microscope images of capillaries are also the subject of [51]. In this case the segmentation uses a genetic algorithm optimisation of an ASM. One of the identifiable regions of the capillary is the lumen through which the blood cells pass. This space may be entirely closed in microangiopathy, and the complete analysis required the development of an extension to the ASM that allowed for the inclusion of binary presence/absence of specific features (see discussion of [27], on page 32).

Later work concentrated on measurement of nerve fibres in images obtained by corneal confocal microscopy (CCM). This in-vivo microscopic technique allows nerves at a particular layer in the cornea to be visualised and measured. Several authors had proposed that measurement of length and density of the nerve fibres was related to neuropathy (e..g. Hossain *et al.* (2005), Malik *et al.* (2002), Mehra *et al.* (2007), Hertz *et al.* (2011)), and might form a new biomarker. Manual or interactive measurement of the image features was sufficient to demonstrate the association, but suffered from the usual problems of being lengthy, tedious and subjective. Development and evaluation of methods for conducting this analysis automatically are the subjects of papers [52-56]. In [52] a method for detection of the nerve fibres is described which deals with the often low contrast to noise ratio of the CCM images. Detection of the linear nerve structures has something in common with other applications, such as analysis of retinal images and mammograms (for example). Previous approaches to the problem include Scarpa *et al.* (2008), who described a heuristic method of analysing CCM images adapted from the analysis of retinal images and Holmes *et al.* (2010), who based the detection of nerve fibres on ridge points. In [52] a model of fibre appearance is developed, based on a Gabor function representing the fibres and a Gaussian

background model. This "dual model" was compared with several other algorithms designed to detect linear structures and found to perform well in comparison in this application. In [53] the analysis is extended to multiple scales. Pixel classification as fibre/non-fibre on the basis of the dual model output was achieved using an MLP neural network (shown to slightly outperform random forest classification in this application). A complete system for measurement of a number of CCM features is described in [54], which also includes a technical evaluation, while [55] presents a clinical evaluation and comparison between manual and automatic analysis.

An advantage of CCM analysis over other clinical methods for diagnosing peripheral neuropathy, such as electrophysiology, is that the latter focuses on large-fibre deficits, whereas CCM imaging assesses small fibres, where the earliest signs of neuropathy occur (Dyck *et al.* (1993)). Intra-epidermal nerve fibre density (IENFD) by microscopic analysis of skin biopsy samples is the only other technique that seeks to quantify the morphology of small nerve fibres. It is clearly invasive, but constitutes the current "gold standard" in definitive diagnosis of neuropathy. Paper [56] is a direct comparison of automatic CCM image analysis and IENFD, showing that the two methods are equivalent in determining neuropathy, while CCM analysis is totally non-invasive and automatic.

Paper [52] was accepted as an oral presentation at MICCAI 2010. The paper acceptance rate at MICCAI is low, and oral presentations are only a small proportion of accepted papers. This paper was among 10 invited to submit expanded versions for inclusion in Medical Image Analysis, itself one of the most selective journals in the field. The expanded version is [53] (29 and 58 citations, respectively). Paper [54] is in press in IEEE Transactions on Biomedical Engineering, the version included here appearing as a preprint on the IEEE Explore website. The clinical evaluation [55] has 44 citations. While paper [56] was published recently (2015), it now has 15 citations. The software described in [54] is available on free licence and has been licensed to 70 research groups internationally at the time of writing.

Carpal kinematics

There has been an increasing recent interest in modelling the movement of bones in articulated joints (e.g. Baka *et al.* (2012), van de Giessen *et al.* (2012), Abu Anas

et al. (2014)), usually for the purpose of measuring abnormal movement. Papers [57 – 60] investigate the computational approach necessary for inferring the 3D motion of the carpal bones during wrist movement from a single 2D projection cine-fluoroscopy sequence. Such sequences are used by clinical experts in assessing different types of wrist disorder.

The papers explore several challenges. First is the reconstruction from a 2D projection video sequence of 3D shape and 3D pose (position and orientation) of bones that move in a complex, articulated fashion. This is similar to the problem of 2D-3D projection, often used in aligning pre-operative images to intra-operative images in image-guided surgery. In this case, however, the 2D image is of an individual and the 3D counterpart consists of a model of 3D shape and pose of ten bones as they follow a complex 3D trajectory. Registration of a 3D model to a 2D image is also addressed by Baka *et al.* (2011) in the case of the femur, using stereo X-ray images.

The second challenge is the segmentation of a large number of 3D images to act as the basis for the model. In this case the surfaces of ten separate objects (the eight carpal bones and the ends of the radius and ulna) need to be determined in several poses. The approach to conducting this segmentation and model building are described initially in [58] and in more detail in [60]. The method is similar to that developed in [48-49] (see page 39), making use of the grow-cut algorithm (Vezhnevets and Konouchine (2005)) in this case as multiple labels need to be assigned to voxels. The third challenge is the form of the pose model and how it can be invoked for inference. This is described in [57] and in more detail in [59], where an example is given of diagnosis of a specific wrist pathology. Similar pose models are used by van de Giessen *et al.* (2009) and van de Giessen et al. (2012) for segmentation of wrist bones and measuring kinematics. However, their study made use of 4D CT image sets, rather than the 2D fluoroscopy sequences used in [59].

Paper [57] was an oral presentation at MICCAI 2011. The expanded paper [59] and paper [60] appear in IEEE Transactions on Medical Imaging, one of the most selective outlets in the field.

References

1960. A proposed standard system of nomenclature of human mitotic chromosomes (Denver, Colorado). *Ann Hum Genet,* 24**,** 319-25.

1975. Paris Conference (1971), supplement (1975) Standardization in human cytogenetics. *Cytogenet Cell Genet,* 15**,** 203-38.

Abu Anas, E. M., Rasoulian, A., St John, P., Pichora, D., Rohling, R., et al. 2014. A Statistical Shape plus Pose Model for Segmentation of Wrist CT Images. *Medical Imaging 2014: Image Processing,* 9034.

Agam, G. & Dinstein, I. 1997. Geometric separation of partially overlapping nonrigid objects applied to automatic chromosome classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 19**,** 1212-1222.

Anderson, N. L., Taylor, J., Scandora, A. E., Coulter, B. P. & Anderson, N. G. 1981. The Tycho System for Computer-Analysis of Two-Dimensional Gel-Electrophoresis Patterns. *Clinical Chemistry,* 27**,** 1807-1820.

Baka, N., de Bruijne, M., van Walsum, T., Kaptein, B. L., Giphart, J. E., et al. 2012. Statistical Shape Model-Based Femur Kinematics From Biplane Fluoroscopy. *IEEE Transactions on Medical Imaging,* 31**,** 1573-1583.

Baka, N., Kaptein, B. L., de Bruijne, M., van Walsum, T., Giphart, J. E., et al. 2011. 2D-3D shape reconstruction of the distal femur from stereo X-ray imaging using statistical shape models. *Medical Image Analysis,* 15**,** 840-850.

Benneworth, P. 2001. Academic entrepreneurship and long-term business relationships: understanding 'commercialzation' activities. *Enterprise and Innovation Management Studies,* 2**,** 255-237.

Besl, P. J. & Mckay, N. D. 1992. A Method for Registration of 3-D Shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 14**,** 239-256.

Bookstein, F. L. 1989. Principal Warps - Thin-Plate Splines and the Decomposition of Deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 11**,** 567-585.

Boykov, Y. Y. & Jolly, M. P. 2001. Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. *Eighth IEEE International Conference on Computer Vision, Vol I, Proceedings***,** 105-112.

Breiman, L. 2001. Random forests. *Machine Learning,* 45**,** 5-32.

Brenner, J. F., Dew, B. S., Horton, J. B., King, T., Neurath, P. W., et al. 1976. An automated microscope for cytologic research a preliminary evaluation. *J*

Histochem Cytochem, 24**,** 100-11.

Cadarette, S. M., Jaglal, S. B. & Murray, T. M. 1999. Validation of the simple calculated osteoporosis risk estimation (SCORE) for patient selection for bone densitometry. *Osteoporos Int,* 10**,** 85-90.

Canny, J. 1986. A Computational Approach to Edge-Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 8**,** 679-698.

Carothers, A. & Piper, J. 1994. Computer-Aided Classification of Human-Chromosomes - a Review. *Statistics and Computing,* 4**,** 161-171.

Castleman, K. R. & Melnyck, J. H. 1976. An automated system for chromosome analysis. Pasadena, CA: California Institute of Technology.

Castleman, K. R., Melnyk, J., Frieden, H. J., Persinger, G. W. & Wall, R. J. 1976. Computer-Assisted Karyotyping. *Journal of Reproductive Medicine,* 17**,** 53-57.

Chin, R. T. & Dyer, C. R. 1986. Model-Based Recognition in Robot Vision. *Computing Surveys,* 18**,** 67-108.

Chui, H., Rangarajan, A., Zhang, J. & Leonard, C. M. 2004. Unsupervised learning of an Atlas from unlabeled point-sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 26**,** 160-172.

Cootes, T. F., Edwards, G. J. & Taylor, C. J. 2001. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 23**,** 681-685.

Davies, R. H., Twining, C. J., Allen, P. D., Cootes, T. F. & Taylor, C. J. 2003. Building optimal 2D statistical shape models. *Image and Vision Computing,* 21**,** 1171-1182.

Davies, R. H., Twining, C. J., Cootes, T. F. & Taylor, C. J. 2010. Building 3-D Statistical Shape Models by Direct Optimization. *IEEE Transactions on Medical Imaging,* 29**,** 961-981.

Dyck, P. J., Kratz, K. M., Karnes, J. L., Litchy, W. J., Klein, R., et al. 1993. The prevalence by staged severity of various types of diabetic neuropathy, retinopathy, and nephropathy in a population-based cohort: the Rochester Diabetic Neuropathy Study. *Neurology,* 43**,** 817-24.

Fischler, M. A. & Bolles, R. C. 1981. Random Sample Consensus - a Paradigm for Model-Fitting with Applications to Image-Analysis and Automated Cartography. *Communications of the Acm,* 24**,** 381-395.

Garrels, J. I. 1989. The Quest System for Quantitative-Analysis of Two-Dimensional Gels. *Journal of Biological Chemistry,* 264**,** 5269-5282.

Geraets, W. G. & van der Stelt, P. F. 2000. Fractal properties of bone.

Dentomaxillofac Radiol, 29**,** 144-53.

Ghose, S., Oliver, A., Marti, R., Llado, X., Vilanova, J. C., et al. 2012. A survey of prostate segmentation methodologies in ultrasound, magnetic resonance and computed tomography images. *Computer Methods and Programs in Biomedicine,* 108**,** 262-287.

Grady, L. 2006. Random walks for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 28**,** 1768-1783.

Granlund, G. H. 1976. Identification of Human-Chromosomes by Using Integrated Density Profiles. *IEEE Transactions on Biomedical Engineering,* 23**,** 182-192.

Granlund, G. H., Zack, G. W., Young, I. T. & Eden, M. 1976. A technique for multiple-cell chromosome karyotyping. *J Histochem Cytochem,* 24**,** 160-7.

Granum, E. 1982. Application of Statistical and syntactical methods of analysis and classification to chromosome data. *In:* Kittler, J., Fu, K. S. & Pau, L. F. (eds.) *Pattern Recognition Theory and Applications.* London: D. Reidel.

Groen, F. C. A., Tenkate, T. K., Smeulders, A. W. M. & Young, I. T. 1989. Human-Chromosome Classification Based on Local Band Descriptors. *Pattern Recognition Letters,* 9**,** 211-222.

Habbema, J. D. F. 1979. Statistical-Methods for Classification of Human-Chromosomes. *Biometrics,* 35**,** 103-118.

Haralick, R. M., Shanmugam, K., Dinstein I. 1973. Textural features for image classification. *IEEE transactions on systems, man and cybernetics,* 3**,** 610 - 621.

Heimann, T. & Meinzer, H. P. 2009. Statistical shape models for 3D medical image segmentation: A review. *Medical Image Analysis,* 13**,** 543-563.

Hertz, P., Bril, V., Orszag, A., Ahmed, A., Ng, E., et al. 2011. Reproducibility of in vivo corneal confocal microscopy as a novel screening test for early diabetic sensorimotor polyneuropathy. *Diabetic Medicine,* 28**,** 1253-1260.

Holmes, T. J., Pellegrini, M., Miller, C., Epplin-Zapf, T., Larkin, S., et al. 2010. Automated software analysis of corneal micrographs for peripheral neuropathy. *Invest Ophthalmol Vis Sci,* 51**,** 4480-91.

Horner, K. & Devlin, H. 1998. The relationships between two indices of mandibular bone quality and bone mineral density measured by dual energy X-ray absorptiometry. *Dentomaxillofac Radiol,* 27**,** 17-21.

Hossain, P., Sachdev, A. & Malik, R. A. 2005. Early detection of diabetic peripheral neuropathy with corneal confocal microscopy. *Lancet,* 366**,** 1340-1343.

Huber, P. J. 1981. *Robust Statistics,* New York, John Wiley and Sons.

Ji, L. A. 1989. Intelligent Splitting in the Chromosome Domain. *Pattern Recognition,* 22**,** 519-532.

Johnell, O. 1996. Advances in osteoporosis: better identification of risk factors can reduce morbidity and mortality. *J Intern Med,* 239**,** 299-304.

Johnson, E. T. & Goforth, L. J. 1974. Metaphase Spread Detection and Focus Using Closed-Circuit Television. *Journal of Histochemistry & Cytochemistry,* 22**,** 536-545.

Kanis, J. A., Johnell, O., Oden, A., Johansson, H. & McCloskey, E. 2008. FRAX and the assessment of fracture probability in men and women from the UK. *Osteoporos Int,* 19**,** 385-97.

Kao, J. H., Chuang, J. H. & Wang, T. 2008. Chromosome classification based on the band profile similarity along approximate medial axis. *Pattern Recognition,* 41**,** 77-89.

Kass, M., Witkin, A. & Terzopoulos, D. 1987. Snakes - Active Contour Models. *International Journal of Computer Vision,* 1**,** 321-331.

Kleinschmidt, P., Lee, C. W. & Schannath, H. 1987. Transportation Problems Which Can Be Solved by the Use of Hirsch Paths for the Dual Problems. *Mathematical Programming,* 37**,** 153-168.

Klemetti, E., Kolmakov, S. & Kroger, H. 1994. Pantomography in assessment of the osteoporosis risk group. *Scand J Dent Res,* 102**,** 68-72.

Klemetti, E. & Kolmakow, S. 1997. Morphology of the mandibular cortex on panoramic radiographs as an indicator of bone quality. *Dentomaxillofac Radiol,* 26**,** 22-5.

Knoll, C., Alcaniz, M., Grau, V., Monserrat, C. & Juan, M. C. 1999. Outlining of the prostate using snakes with shape restrictions based on the wavelet transform (Doctoral Thesis: Dissertation). *Pattern Recognition,* 32**,** 1767-1781.

Ledley, R. S., Lubs, H. A. & Ruddle, F. H. 1972. Introduction to automatic chromosome analysis. *Computers in Medicine and Biology,* 2.

Ledley, R. S., Rotolo, L. S. & Golab, T. J. e. a. 1965. FIDAC - Film Inpit to Digtal Automatic Computer and associated syntax directed pattern recognition programming system. *Optical and Electro-optical Information Processing.* Cambridge, Mass.: MIT Press.

Lemkin, P. F. & Lipkin, L. E. 1981. Gellab - a Computer-System for Two-Dimensional Gel-Electrophoresis Analysis .3. Multiple Two-Dimensional Gel Analysis. *Computers and Biomedical Research,* 14**,** 407-446.

Lerner, B. 1998. Toward a completely automatic neural-network-based human chromosome analysis. *IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetics,* 28**,** 544-552.

Lerner, B., Guterman, H., Dinstein, I. & Romem, Y. 1995. Human-Chromosome Classification Using Multilayer Perceptron Neural-Network. *International Journal of Neural Systems,* 6**,** 359-370.

Lockwood, D. H., Johnston, D. A., Riccardi, V. M. & Zimmerman, S. O. 1988. The use of subchromosome-length unique band sequences in the analysis of prophase chromosomes. *Am J Hum Genet,* 43**,** 934-47.

Lundsteen, C., Gerdes, T., Maahr, J. & Philip, J. 1987. Clinical performance of a system for semiautomated chromosome analysis. *Am J Hum Genet,* 41**,** 493-502.

Lundsteen, C. & Granum, E. 1979. Description of Chromosome-Banding Patterns by Band Transition Sequences - New Basis for Automated Chromosome Analysis. *Clinical Genetics,* 15**,** 418-429.

Lundsteen, C., Lind, A. M. & Granum, E. 1976. Visual classification of banded human chromosomes. I. Karyotyping compared with classification of isolated chromosomes. *Ann Hum Genet,* 40**,** 87-97.

Lundsteen, C. & Martin, A. O. 1989. On the selection of systems for automated cytogenetic analysis. *Am J Med Genet,* 32**,** 72-80.

Madabhushi, A., Dowling, J., Huisman, H. & Barrat, D. (eds.) 2011. *Prostate Cancer Imaging, Image Analysis and Image Guided Interventions,* Berlin: Springer-Verlag.

Maintz, J. B. & Viergever, M. A. 1998. A survey of medical image registration. *Med Image Anal,* 2**,** 1-36.

Malik, R. A., Kallinikos, P., Abbott, C., Vanschie, C., O'Donnell, C., et al. 2002. Corneal confocal microscopy: A rapid, non-invasive technique to define nerve fibre degeneration in human diabetic neuropathy. *Diabetes,* 51**,** A79-A79.

Mayall, B. H. 1974. Digital Image-Processing at Lawrence-Livermore Laboratory .2. Biomedical Applications. *Computer,* 7**,** 81-&.

Mayall, B. H., Tucker, J. D., Christensen, M. L., van Vliet, L. J. & Young, I. T. 1990. Experience with the Athena semi-automated karyotyping system. *Cytometry,* 11**,** 59-72.

McInerney, T. & Terzopoulos, D. 1996. Deformable models in medical image analysis. *Proceedings of the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis***,** 171-180.

Mehra, S., Tavakoli, M., Kallinikos, P. A., Efron, N., Boulton, A. J. M., et al. 2007.

Corneal confocal microscopy detects early nerve regeneration after pancreas transplantation in patients with type 1 diabetes. *Diabetes Care,* 30**,** 2608-2612.

Moradi, M. & Setarehdan, S. K. 2006. New features for automatic classification of human chromosomes: A feasibility study. *Pattern Recognition Letters,* 27**,** 19-28.

Petrou, M. & Carcia Sevilla, P. 2006. *Image Processing, Dealing with Texture,* Chichester, John Wiley and Sons.

Pingel, K. & Jenkinson, G. 2016. *Fifty years of image analysis* [Online]. Leica microsystems. Available: http://www.leica-microsystems.com/science-lab/history/50-years-of-image-analysis/ [Accessed 2016].

Piper, J. 1986. Classification of Chromosomes Constrained by Expected Class Size. *Pattern Recognition Letters,* 4**,** 391-395.

Piper, J. 1992. Variability and Bias in Experimentally Measured Classifier Error Rates. *Pattern Recognition Letters,* 13**,** 685-692.

Piper, J. 1995. Genetic Algorithm for Applying Constraints in Chromosome Classification. *Pattern Recognition Letters,* 16**,** 857-864.

Rutovitz, D. 1966. Pattern Recognition. *Journal of the Royal Statistical Society Series a-General,* 129**,** 504-530.

Rutovitz, D. 1977. Chromosome classification and segmentation as an exercise in knowing what to expect. *In:* Elcock, E. W. & Michie, D. (eds.) *Machine Intelligence.* London: Ellis Horwood.

Ruttimann, U. E., Webber, R. L. & Hazelrig, J. B. 1992. Fractal dimension from radiographs of peridental alveolar bone. A possible diagnostic indicator of osteoporosis. *Oral Surg Oral Med Oral Pathol,* 74**,** 98-110.

Scarpa, F., Grisan, E. & Ruggeri, A. 2008. Automatic Recognition of Corneal Nerve Structures in Images from Confocal Microscopy. *Investigative Ophthalmology & Visual Science,* 49**,** 4801-4807.

Schoevaert-Brossault, D., Leonard, C. & Selva, J. 1983. A new method for automatic metaphase finding adaptable to different chromosome preparations. *Comput Programs Biomed,* 16**,** 195-201.

Sedrine, W. B., Chevallier, T., Zegels, B., Kvasz, A., Micheletti, M. C., et al. 2002. Development and assessment of the Osteoporosis Index of Risk (OSIRIS) to facilitate selection of women for bone densitometry. *Gynecol Endocrinol,* 16**,** 245-50.

Shippey, G., Carothers, A. D. & Gordon, J. 1986. Operation and performance of an automatic metaphase finder based on the MRC fast interval processor. *J*

Histochem Cytochem, 34**,** 1245-52.

Smilansky, Z. 2001. Automatic registration for images of two-dimensional protein gels. *Electrophoresis,* 22**,** 1616-1626.

Staib, L. H. & Duncan, J. S. 1992. Boundary Finding with Parametrically Deformable Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 14**,** 1061-1075.

Taguchi, A., Tanimoto, K., Suei, Y., Ohama, K. & Wada, T. 1996. Relationship between the mandibular and lumbar vertebral bone mineral density at different postmenopausal stages. *Dentomaxillofac Radiol,* 25**,** 130-5.

Taguchi, A., Tsuda, M., Ohtsuka, M., Kodama, I., Sanada, M., et al. 2006. Use of dental panoramic radiographs in identifying younger postmenopausal women with osteoporosis. *Osteoporos Int,* 17**,** 387-94.

Tewari, A., Shinohara, K. & Narayan, P. 1995. Transition zone volume and transition zone ratio: predictor of uroflow response to finasteride therapy in benign prostatic hyperplasia patients. *Urology,* 45**,** 258-64; discussion 265.

Vaidyanathan, S. G., Kumaravel, N. & Kar, B. 2009. A novel technique for identification of partially hidden metaphase chromosomes. *International Conference on Signal Processing Systems.* Singapore: IEEE

van de Giessen, M., Foumani, M., Vos, F. M., Strackee, S. D., Maas, M., et al. 2012. A 4D Statistical Model of Wrist Bone Motion Patterns. *IEEE Transactions on Medical Imaging,* 31**,** 613-625.

van de Giessen, M., Streekstra, G. J., Strackee, S. D., Maas, M., Grimbergen, K. A., et al. 2009. Constrained Registration of the Wrist Joint. *IEEE Transactions on Medical Imaging,* 28**,** 1861-1869.

van Vliet, L. J., Young, I. T. & Mayall, B. H. 1990. The Athena semi-automated karyotyping system. *Cytometry,* 11**,** 51-8.

Vezhnevets, V. & Konouchine, V. 2005. "Grow--cut" - interactive multi-label N-D image segmentation. *International Conference on Computer Graphics and its Applications (Graphicon).* Novosibirsk Akademgorodok, Russia.

Vossepoel, A. M. 1989. Separation of touching chromosomes. *In:* Lunsteen, C. & Piper, J. (eds.) *Automation of Cytogenetics.* Berlin: Springer.

Wald, N., Fatora, S. R., Herron, J. M., Preston, K., Li, C. C., et al. 1976. Status-Report on Automated Chromosome Aberration Detection. *Journal of Histochemistry & Cytochemistry,* 24**,** 156-159.

Wang, X. W., Zheng, B., Li, S. B., Mulvihill, J. J., Wood, M. C., et al. 2009. Automated

classification of metaphase chromosomes: Optimization of an adaptive computerized scheme. *Journal of Biomedical Informatics,* 42**,** 22-31.

White, S. C., Taguchi, A., Kao, D., Wu, S., Service, S. K., et al. 2005. Clinical and panoramic predictors of femur bone mineral density. *Osteoporos Int,* 16**,** 339-46.

Woods, P. W., Taylor, C. J., Cooper, D. H. & Dixon, R. N. 1987. The Use of Geometric and Gray-Level Models for Industrial Inspection. *Pattern Recognition Letters,* 5**,** 11-17.

Wu, Q., Snellings, J., Amory, L., Suetens, P. & Oosterlink, A. 1989. Model-based contour analysis in a chromosome segmentation system. *In:* Lundsteen, C. & Piper, J. (eds.) *Automation of Cytogenetics.* Berlin: Springer.

Yasar, F. & Akgunlu, F. 2006. The differences in panoramic mandibular indices and fractal dimension between patients with and without spinal osteoporosis. *Dentomaxillofacial Radiology,* 35**,** 1-9.

Yuille, A. L., Hallinan, P. W. & Cohen, D. S. 1992. Feature-Extraction from Faces Using Deformable Templates. *International Journal of Computer Vision,* 8**,** 99-111.

Zwiggelaar, R., Zhu, Y. N. & Williams, S. 2003. Semi-automatic segmentation of the prostate. *Pattern Recognition and Image Analysis, Proceedings,* 2652**,** 1108-1116.

Reproduction of Publications.

Each of the publications listed is reproduced in the following pages.

Image Analysis Software Architecture

1. **An architecture for integrating symbolic and numeric image processing.** C.J. Taylor, R.N. Dixon, P.J. Gregory and J. Graham , in *"Intermediate - Level Image Processing", M.J.B. Duff (ed.), 1986, Academic Press, London, pp 19 - 34.*

Chapter Two

An Architecture for Integrating Symbolic and Numerical Image Processing

C. J. Taylor, R. N. Dixon, P. J. Gregory and J. Graham

1 INTRODUCTION

There are commonly held to be two, qualitatively distinct, types of activity involved in image processing. Low-level tasks require relatively simple processing, are predominantly numerical and involve large quantities of data. High-level processing is predominantly symbolic and involves relatively small quantities of data. It is often suggested that some form of special hardware is required for low-level processing and that this can be interfaced to a fairly conventional processor capable of undertaking high-level processing [1–3]. In this paper we suggest that it is important to take a more integrated approach to the problem. We confine our attention to relatively conventional computational methods and describe how our ideas have led to the design of two similar hardware systems, both of which have been commercially exploited. It must be said that there are more radical approaches to the problem of dealing with both numerical and symbolic processing, such as those proposed by advocates of connectionist machines [4,5]. These ideas are, however, a long way from practical implementation, and we are particularly interested in systems capable of undertaking visual tasks of useful complexity within the constraints of current technology.

2 COMBINING HIGH-LEVEL AND LOW-LEVEL PROCESSING

When an image is first read into an image-processing system it has very

limited symbolic content and consists mainly of unstructured numerical data. In virtually all practical applications of image processing the aim is to generate a symbolic representation (interpretation) of the image. This may often be associated with a relatively small quantity of numerical data. There are normally additional data of both kinds available to the system (prior knowledge) which can be used in generating the desired solution. The processing that is close to the original image and which predominantly involves numerical data is often called low-level. That which is close to the solution and predominantly involves symbolic data is often called high-level. This distinction is useful for the purposes of discussion, although in practice the aim of any well-designed system is to present the programmer with an integrated set of representational methods that avoid any such dichotomy. The question that we address here concerns the relationship that should exist between the two types of processing.

2.1 The relationship between high-level and low-level processing

First we should make a point that is rather obvious but sometimes seems to be ignored. This is simply that when we attempt to automate real visual tasks it is generally the case that both kinds of processing play an important role. It is often recognized that the quantity of numerical data representing an image results in a large computational burden for low-level processes, which are, of necessity, applied to the whole image. It is not often recognized in this context that high-level processing can be considered to require an infinite computational effort. This can be argued by pointing out that there are, for visual tasks of realistic complexity, a virtually infinite number of possible symbolic interpretations of any image. The purpose of high-level processing is to search through these interpretations for one that is in some sense best. In a practical system it is obviously necessary to limit the search space by using heuristics. In the extreme this can lead to an approach where the symbolic element of the processing task is trivial compared with the numerical part. This is, however, an unsound approach, since heuristics that reliably make such dramatic reductions in the solution search space while still guaranteeing to retain the "correct" solution are not known. For these reasons we argue that it is important to apply as much computational effort as possible to high-level processing.

A possible response to this view is to accept the importance of high-level processing and argue that the requirement is thus for two different types of processor, one optimized for low-level processing and one, of conventional design, to undertake high-level processing. This ignores the fact that there is

a rather intimate relationship between symbolic and numerical data which needs to be reflected in considerable commonality between the hardware involved in each type of processing. If we take a bottom-up approach to scene interpretation such as that proposed by Marr and followers [6,7] we will generate a hierarchy of structures, each of which have both symbolic and numerical content. The original image is almost purely numerical, but intrinsic images generated by low-level processing also have significant symbolic content. Primitives extracted from intrinsic images have greater symbolic content but still have significant numerical content and the same is true of the $2\frac{1}{2}$D sketch. The current consensus seems to accept that an arrangement such as this where information flows only in one direction may not be desirable and that the results of higher-level processes probably need to feed back to lower-level processes. This further reinforces the argument that high-level and low-level processing are intimately related and that this should be reflected in hardware architecture.

We have concluded from these arguments that it is the interface between high-level and low-level processing that is crucial. Whatever computing power may be provided independently for the two types of activity, system performance will be dramatically reduced unless the two levels of processing can be efficiently integrated. An obvious approach is to allow data structures to be shared between the two levels. Having made this general assertation, a practical programming paradigm is required. The framework that we have explored in a number of practical applications involves the idea of a moving focus of attention and will be described below. It is by no means the only possible approach, and indeed can be criticised on the grounds that it involves the use of fairly sweeping heuristics to reduce solution search space. The method has, however, proved useful in automating a number of real visual tasks.

2.2 Focus-of-attention paradigm

In any image-processing system it is, by definition, the role of high-level processing to determine which low-level processing ought to be performed. In a simple system such decisions may be predetermined, but we have already argued that this is unsatisfactory. The focus-of-attention method takes a small step away from complete determinism by allowing the data sets on which low-level processing is performed to be selected dynamically. The simple assumption is made that, although it may be impossible, in a single step, to accurately identify the symbolically significant structures in an image, it is possible to do so approximately. The approximate symbolic description can then be used to identify the image regions that require closer

attention in order to obtain a better description. The process can be repeated and provides a controlled method of successively reducing solution search space.

The manner in which a focus-of-attention method can be used to manage the interface between different levels of processing can most easily be explained by briefly describing its application to a real problem. We have chosen to use, as an illustration, a simplified version of the chromosome metaphase analysis software which is part of a routine clinical package developed in this laboratory [8].

An example of a metaphase spread of chromosomes is shown in Fig. 1. The object of the analysis is to recognize individual chromosomes and to label each as one of the 24 unique chromosome types (1–22, X,Y). The main feature used to identify the chromosome type is the banding pattern. Once the chromosomes have been labelled they must be presented in a karyogram, an ordered display sorted by type number (Fig. 2). The simplified analysis sequence is as follows.

1. Get a grey-level histogram of the image and determine a global threshold.

Fig. 1. The chromosomes from a single cell stained to show G-banding.

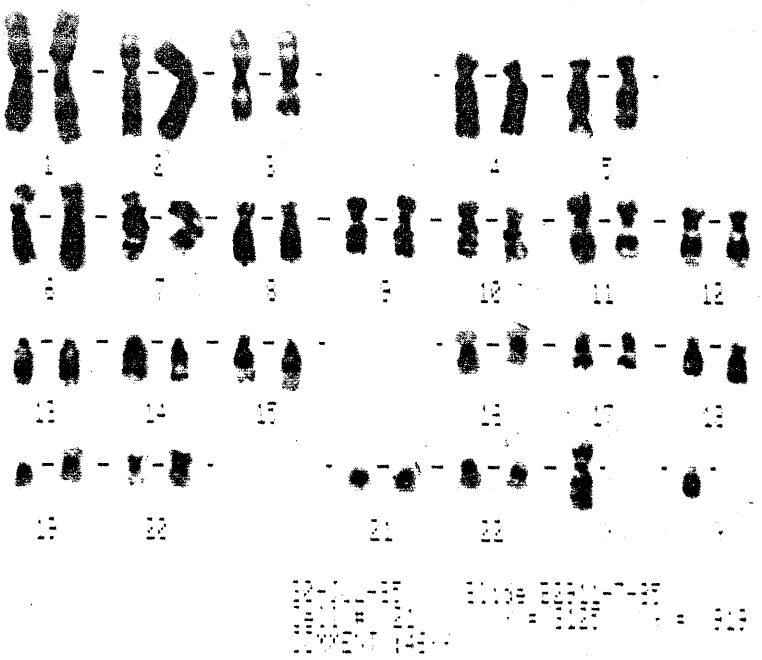


Fig. 2. A karyogram generated from Fig. 1.

2. Threshold the image and find the connected objects in the resulting binary image. These are treated as a rough version of the chromosomes.
3. Get a grey-level histogram of the image in the vicinity of each object and determine a local threshold.
4. Re-threshold the image in the vicinity of each object using the local threshold. Use local connectivity rules to modify each threshold if necessary. If the connectivity rules cannot be satisfied by modifying the threshold in a particular locality then apply a version of the fall set method [9] to segment separate chromosomes.
5. Use curve fitting to obtain a medial axis for each segmented chromosome.
6. For each chromosome, project grey levels from the original image onto the medial axis to obtain an intensity profile. Compare the profile with

a set of standard banding pattern models to select a chromosome type label.

7. Using the type labels, generate a karyogram display. There are pre-assigned slots in the display for two of each chromosome type. Each chromosome is copied into the karyogram image so that its medial axis is vertical. This involves a scale and rotate operation on original image data over the region of the segmented chromosome.

This simple illustration offers many examples of data structures that are involved in both levels of processing. In Step 1 a low-level process generates, from an image, the higher-level histogram structure. High-level processing is involved in syntactically analysing the histogram to obtain a threshold value. In Step 2 low-level processing generates high-level object structures from the image. Step 3 is similar to Step 1 except that the object structures are used to control the low-level process of obtaining a histogram. Step 4 involves an intimate mixture of processing at different levels employing a number of shared data structures. Step 5 is predominantly high-level, and Step 6 once again uses a high-level structure (the medial axis) to control low-level processing. In Step 7 a number of high-level structures generated earlier are used to control the low-level processing involved in generating the karyogram display.

The close coupling between the two levels of processing illustrated by this example suggests that an integrated design methodology ought to be adopted. Particular attention should be paid to the way in which data are stored and manipulated, employing as much shared hardware as possible. This will lead to an architecture where data paths are simplified and overall computational power is significantly enhanced by avoiding unnecessary data transfers.

3 PRACTICAL IMPLEMENTATION

In this section we describe how the ideas outlined above led to the design of two practical image-processing systems which have both been exploited commercially. The Joyce-Loebl Magiscan 2 (M2) was designed and constructed by the authors in 1979/80. The Visual Machines VM1 (also known as CVAS 3000) was designed by the authors in 1983/4 and is architecturally similar. The VM1 offers several enhancements over the M2, but since the differences are not important to this discussion we avoid introducing irrelevant detail and present a description which strictly applies to the M2.

The major goals which we were trying to meet in designing the M2 and VM1 were as follows:

(i) to provide hardware support for the high-level/low-level interface;
(ii) to achieve realistic performance for low-level processing;
(iii) to provide a unified and high-level software development environment;
(iv) to keep within a cost to manufacture ceiling of $20 000.

Although a detailed discussion of these objectives is not appropriate here, the broad aim was to produce self-contained systems, technically and economically suited to undertaking complex visual tasks in the laboratory or factory.

Given these goals, we made the early decision that it was not practical to employ massive parallelism in the system. There has been considerable work showing that goal (ii) can be realized quite effectively using specialized hardware architectures onto which low-level processing tasks map well [1,2,10–15], but we found it difficult to see how goal (i) could easily be achieved by such a system. The dedicated paths that enable efficient low-level processing in cellular arrays constitute a significant problem when communication between low-level and high-level processing is required. We also believed that it would be difficult to achieve goals (iii) and (iv) with a system employing large-scale parallelism.

3.1 Support for high-level processing

The problem of providing hardware support for shared data structures was approached by first proposing an architecture to support high-level processing and then considering how it could be modified to also support low-level processing. PASCAL was chosen as the language in which high-level software would be written, particularly because of its ability to handle complex data structures. The ready availability of transportable compilers and development software was also an important factor, since this significantly reduced the software effort involved in supporting the new architecture. The most straightforward way of supporting PASCAL is to use a compiler generating machine-independent p-code and to design hardware able to emulate a p-machine. This can be achieved using the type of structure shown in Fig. 3. The program and data memory is used to store both p-code and data. During program execution the CPU, controlled by a microprogram, can both interpret p-code instructions and act upon data in memory to generate the appropriate action. The addition of an I/O controller extends such an arrangement into a basic minicomputer.

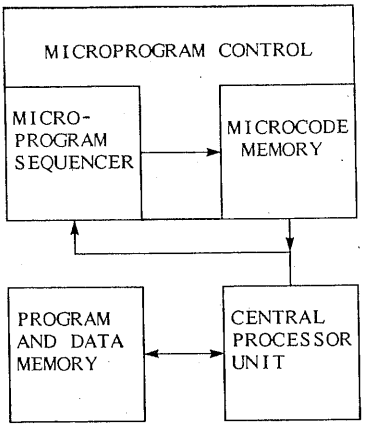


Fig. 3. A microprogrammed high-level processor.

The organization described above is used as the basis for high-level processing in the M2. At the time the system was designed it was possible to achieve a microinstruction cycle time of around 150 ns, making conservative use of standard technology. Given a reasonably powerful microinstruction set, this leads to p-code execution times of around 1 μs, which compares very favourably with other implementations of PASCAL [16]. The M2 uses the UCSD PASCAL compiler and development environment. Modifications to this arrangement, which are not relevant to the discussion, allow the VM1 to use a 68000 coprocessor running UNIX with support for both PASCAL and C. We are actively involved in investigating the application of high-level languages which allow more sophisticated symbolic processing to be performed, and hope in the near future to also provide VM1 support for POPLOG.

3.2 Support for low-level processing

Given the high-level processor outlined above, the next problem was to provide low-level processing of adequate performance and to support shared data structures. Previous applications work [8,17–19] suggested that the ability to perform neighbourhood operations on whole images in times of approximately one second would make the use of such methods a practical proposition. A 3×3 convolution of a 512×512 image in less than one second was adopted as a benchmark for low-level performance, although the same applications experience led us to believe that it would be much too restrictive if this simple class of neighbourhood operation alone were

supported. The aim was to provide complete programmability so that arbitrary nonlinear neighbourhood operations could be performed at comparable speed to linear convolution.

The goal of intimately sharing data structures suggested that the CPU, provided for high-level processing, should also be employed for low-level processing. This represented an ideal arrangement since the CPU was already intended to manipulate high-level structures under control of a microprogram emulating the p-machine. The use of the CPU was a practical proposition since, given a suitable instruction set, a multiply and add could be performed in 300 ns (two instruction times). As long as a memory access time of 300 ns could be achieved, a 3×3 convolution involving nine memory reads and one write could in principle be performed on a 512×512 image in 780 ms. To do this it would be necessary to generate image memory addresses with no time penalty and to effect program flow control, again with no time penalty. This was achieved by adding a memory address processor (MAP) and microprogram control processor (MPC) to act in parallel with the CPU. The three processors operate synchronously and are controlled by separate fields in the microinstruction word. The computing power of both the MAP and MPC is comparable to that of the CPU, so that in a tightly coded loop such as the kernel of a neighbourhood operator an effective instruction time of 50 ns is achieved.

3.3 Hardware description

Figure 4 is a block diagram of the complete M2 hardware showing the main interconnection paths. We can now describe each of the main functional elements in some more detail.

3.3.1 Microprogram control

The MPC consists of a $4k \times 48$-bit microprogram memory and a control sequencer. Each microprogram word is divided into three main fields as shown in Fig. 5. The MPC instruction field allows direct or indirect jumps and subroutine calls and subroutine returns all conditional on CPU status flags. Any branch involves only one instruction and incurs no time penalty. The microprogram memory is loaded via the CPU data path normally from a disk file.

3.3.2 Central processor unit

The CPU is a 16-bit arithmetic and logic processor with a 16-bit barrel

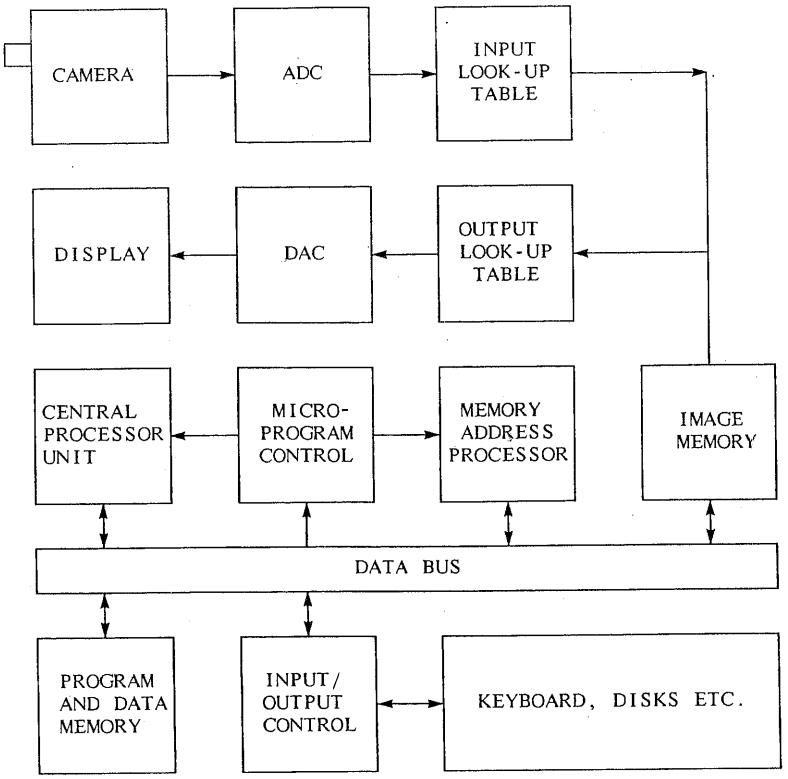


Fig. 4. General arrangement of the Magiscan 2.

shifter, 8 × 8 parallel multiplier, 1024 × 16-bit registers and an accumulator. The general arrangement of the unit is shown in Fig. 6. One operand for the processor is always a register. The same register can also act as the destination. The second operand is one of the other data sources such as the image memory, program memory or MAP. As well as using the selected register as a destination other data destinations such as image memory, program memory or MAP can be addressed.

The barrel shifter appears in the data source input path since its most common use is to extract fields from memory words. In the case of the image memory this allows images of arbitrary word length to be stacked and the data accessed for immediate processing by the CPU. (This also requires bit masking, which is supported in the image memory.) In the case of the program memory the shifter allows efficient field extraction by the p-code interpreter.

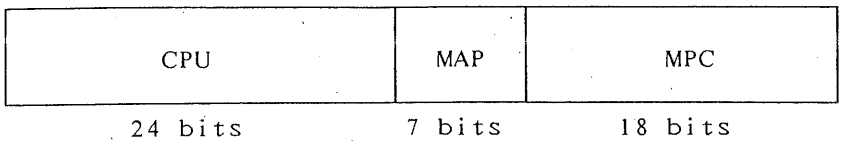


Fig. 5. Microcode instruction fields.

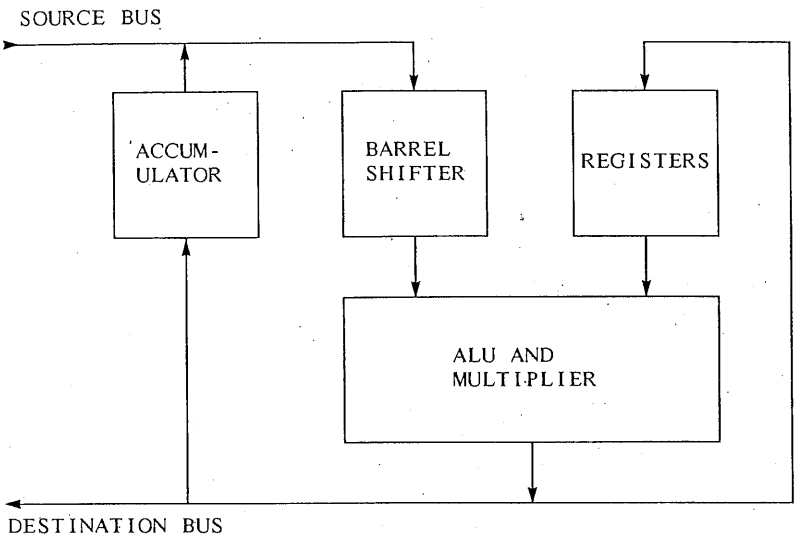


Fig. 6. The central processor unit.

The parallel multiplier allows single-cycle multiplication of operands up to eight bits and is thus suitable for most image data. It operates on the same operands as are presented to the arithmetic and logic processor, and both results are available though normally one will be redundant.

3.3.3 Memory address processor

The MAP is used to compute all access addresses for the image memory. This involves not only microprogram-controlled addressing for the purpose of processing but also memory access for video frame-grabbing and display. The processor also handles light-pen sensing and cursor display and is responsible for refreshing the dynamic memory devices used in the image memory.

The basic arrangement of the MAP is shown in Fig. 7. A 12-bit arithmetic logic processor with sixteen internal registers is used to generate addresses

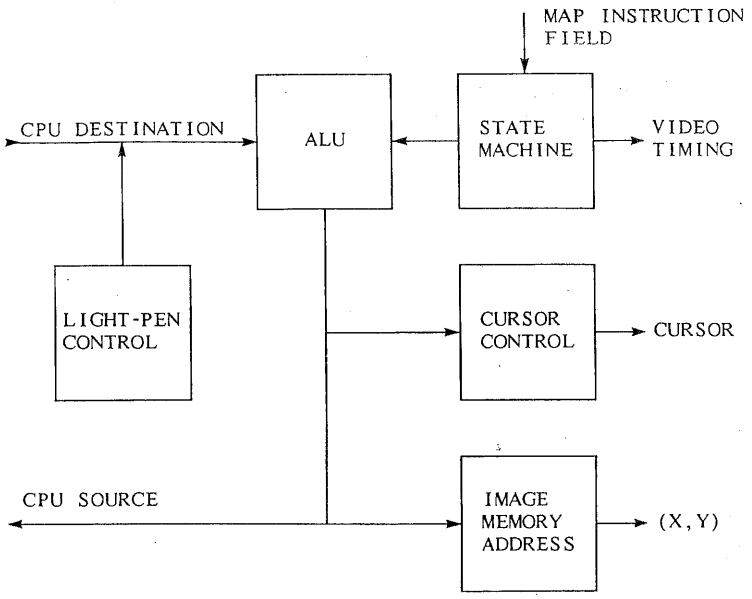


Fig. 7. The memory address processor.

and to carry out other computations necessary to support the light-pen and cursor. A synchronous state machine generates instructions (held in ROM) to control all video related activities and refresh. If an image in the image memory is being displayed, this occupies the processor about 50% of the time. If no stored image is displayed, less than 10% of the time is occupied. Arbitration logic allows microprogram access to the processor whenever it is free. The MAP instruction field allows relative changes in the current X and Y image address pointers, which are held in two registers, and controls the initiation of image memory read or write cycles. Changes in (X, Y) can either be local (± 7 pixels in one instruction) or can use a vector (SX, SY) stored in two other registers to shift between equivalent pixels in two images with different origins. Local changes are used to move around a neighbourhood and through connected regions. The origin shift facility allows for arbitrary allocation of image memory space to source and destination images during image to image transformation.

3.3.4 Image and program memory

There were a number of factors that led to the separation of image memory and program memory. In some ways this was an unfortunate compromise which in any future design should ideally be avoided. The main reasons for

separating the two memories were the limited address range afforded by the processor used in high-level processing (CPU, which is a 16-bit machine), the need to provide a direct video input to image memory and the need to support masking on the image memory. Of these the limited address range was the major problem and was virtually insuperable since it was also the case that UCSD p-code only supported a 16-bit address range. Thus, it was impossible to treat images as PASCAL arrays.

The program memory is a 64k × 16-bit memory with 300 ns cycle time and occupies the full address space supported by PASCAL p-code. Since images themselves cannot be held in this memory they are effectively shared between low-level and high-level processing by using a PASCAL record that purports to be the image but is in fact a specification of the image memory space actually occupied by image data.

The image memory is a 1k × 1k × 16-bit memory with a 300 ns cycle time. It is a hardware option to fit a memory with a word size less than 16 bits since this is seldom required by a single image. Images can be of any number of bits (up to 16 bits) and can be stored at any origin in image memory. Images of 512 × 512 or 256 × 256 can be loaded from video or displayed and some support is provided for 1024 × 1024 images. Data masking is provided on input and output, allowing stacked images to be treated independently. Images can be loaded or displayed at any position in memory.

3.3.5 Video input and output

The video input and output paths contain look-up tables which allow arbitrary transformation of the digital video data. The output table is used in the normal way to affect the way in which data are displayed and allows grey-level display, binary overlays, pseudo-colour display and colour overlays. The input table is used for a number of purposes. First it is used to shift image data to the appropriate bit position within the 16-bit memory data word so that images may be directly loaded into any set of planes. Secondly the table is used to perform single-pixel operations such as contrast manipulation or even grey-level thresholding (at multiple thresholds).

3.4 Software description

An integrated software environment is provided for the development of application programs, which involve both high-level and low-level processing.

Microcode programs are generated by a microcode assembler (written in PASCAL) which accepts symbolic program text and generates object files.

Standard microcode exists to implement a PASCAL p-machine on which compiled UCSD PASCAL will execute. Standard microcode is also provided to support a set of primitive manipulations of shared data structures. The facilities provided are sufficiently comprehensive that it is uncommon for even a sophisticated user to require new microcode. The most common reason for generating new microcode is to implement a new type of neighbourhood operator. In this case standard routines provide support for image memory addressing, result scaling, etc. and only the operator kernel need be written.

PASCAL programs are prepared, compiled and linked in the normal way under the UCSD system. Calls to microcode primitives are indistinguishable from calls to PASCAL procedures and the normal user is unaware which of the routines in the library are implemented in PASCAL and which in microcode. The microcode assembler generates link files which are used to statically bind PASCAL and microcode when an application program is link-loaded. When the application program is executed its corresponding microcode file is automatically loaded into microcode memory.

4 DISCUSSION

In the early part of the chapter we presented some general principles which we believe are important to any discussion of image-processing architecture. These ideas influenced the design of the M2 and VM1 systems, but these machines do not, by any stretch of the imagination, fully embody the philosophy.

From a programmer's viewpoint there are two unnecessarily inconvenient features. First is the fact that high-level and low-level code must be written in different languages. The effect of this is significantly reduced by the provision of a comprehensive set of microcode primitives that are sufficiently general that most applications programs can be written without the need to generate new microcode. When new microcode is required the fact that common data structures can be manipulated is important, but the arbitrary distinction between programming methods for the two levels of processing is contrary to the spirit of the system. Similar comments can be made about the arbitrary split between the image memory and the program memory. This creates the additional task for the programmer of managing the allocation of image space. This problem is partially overcome in the VM1 by a software image memory manager, but this is still not an ideal arrangement since it falls outside the normal rules of memory management supported by PASCAL for all other data structures.

Another shortcoming of both the M2 and VM1 is the use of PASCAL (or

C) for high-level processing. Procedural languages are not ideally suited to symbolic processing of any complexity and restrict the choice of analysis strategy. Logic programming or data flow languages would probably be more appropriate, although a multiparadigm programming environment such as LOOPS [20] may be ideal.

The focus-of-attention method is the main framework that we have used in applications programming. As we have already pointed out, this only represents a limited improvement over a completely predetermined solution. The main problem is that the method is basically procedural and, as we have already suggested in the context of languages, this is not entirely desirable. The choice of method was, of course influenced by the languages available. The result of the strictly procedural approach is that it is difficult to generate solutions that are adaptive to the image data that is presented, either at run-time or beforehand during some form of training. In fact the situation is not quite as bad as it seems, since it is possible to embed data driven methods into such a framework. For industrial inspection we have, for instance, used a model matching technique to identify individual mechanical components. We control the application of model matching using the focus-of-attention method. A possible generalization of the method that does not involve the same limitations would use visual cues (not necessarily approximately correct symbolic descriptions) to direct the application of model matching, the results of which could be used recursively as cues.

Finally, the systems as described offer significant but limited computational power for both high-level and low-level processing. In common with most workers in the field, we are interested in the use of VLSI technology to achieve high computational power at reasonable cost. Clearly the use of regular parallelism is becoming a more and more attractive way of achieving this. We strongly believe, however, that it is important to adhere to the basic principles that we have discussed earlier in the paper. SIMD machines do not seem to offer a convenient mechanism for handling the interface between different levels of programming. We still approach architecture from the standpoint that the ability to perform efficient low-level processing should be embedded into a powerful high-level processor.

In summary, practical experience of tackling image-processing applications has led us to develop complementary hardware and software that allow high-level and low-level processing to be integrated in an efficient manner. We know there is still a long way to go!

REFERENCES

[1] Duff, M. J. B. (1979). Parallel processors for digital image processing. In *Advances in Digital Image Processing* (ed. P. Stucki), pp. 265–276. Plenum Press, New York.

[2] Hunt, D. J. (1981). The ICL DAP and its application to image processing. In *Languages and Architectures for Image Processing* (ed. M. J. B. Duff and S. Levialdi), pp. 275–282. Academic Press, London.

[3] Shippey, G., Bayley, R., Farrow, S., Lutz, R. and Rutovitz, D. (1980). A fast interval processor (FIP) for cervical pre-screening. *Anal. Quant. Cytol.* **3,** 9–16.

[4] Ackley, D. H., Hinton, G. E. and Sejnowski, T. J. (1985). A learning algorithm for Boltzmann Machines. *Cognitive Sci.* **9,** 147–169.

[5] Feldman, J. A. and Ballard, D. H. (1982). Connectionist models and their properties. *Cognitive Sci.* **6,** 205–254.

[6] Marr, D. (1982). *Vision.* Freeman, New York.

[7] Mayhew, J. E. W. and Frisby, J. P. (1978). Texture discrimination and Fourier analysis in human vision. *Nature* **275,** 438–439.

[8] Graham, J. and Taylor, C. J. (1980). Automated chromosome analysis using the Magiscan image analyser. *Anal. Quant. Cytol.* **2,** 237–242.

[9] Rutovitz, D. (1978). Expanding picture components to natural density boundaries by propagation methods. The notion of fall set and fall distance. In *Proc. 4th IJCPR, Kyoto, Japan,* pp. 657–664.

[10] Batcher, K. E. (1980). Design of a massively parallel processor. *IEEE Trans. Comp.* **29,** 836–840.

[11] Kruse, B. (1976). The PICAP picture processing laboratory. In *Proc. 3rd IJCPR, Coronado, California,* pp. 875–881.

[12] Uhr, L. and Douglass, R. (1979). A parallel-serial recognition cone system for perception. *Patt. Recog.* **11,** 29–40.

[13] Granlund, G. H. (1981). GOP: a fast and flexible processor for image analysis. In *Languages and Architectures for Image Processing* (ed. M. J. B. Duff and S. Levialdi), pp. 179–188. Academic Press, London.

[14] Gerritsen, F.A. and Monhemius, R.D. (1981). Evaluation of the Delft Image Processor DIP-1. In *Languages and Architectures for Image Processing* (ed. M. J. B. Duff and S. Levialdi), pp. 189–203. Academic Press, London.

[15] Gerritsen, F. A. (1983). A comparison of the CLIP4, DAP and MPP processor-array implementations. In *Computing Structures for Image Processing* (ed. M. J. B. Duff), pp. 15–30. Academic Press, London.

[16] Gilbreath, J. and Gilbreath, G. (1983). Erastothenes revisited: once more through the sieve. *Byte* **8,** 283–326.

[17] Dixon, R. N. and Taylor, C. J. (1979). Automated asbestos fibre counting. In *Machine Aided Image Analysis,* pp. 178–185. IOP Conf. Ser. No. 44, Institute of Physics, Bristol.

[18] Pycock, D. and Taylor, C. J. (1980). Use of the Magiscan image analyser in automated uterine cancer cytology. *Anal. Quant. Cytol.* **2,** 195–201.

[19] Brunt, J. N. H., Taylor, C. J., Dixon R. N. and Gregory P. J. (1983). Theory and practice of applying image analysis to angiography. In *Physical Techniques in Cardiological Imaging* (ed. M. D. Short *et al.*), pp. 153–162. Adam Hilger, Bristol.

[20] Bobrow, D. G. and Stefik, M. (1981). The LOOPS manual. *Tech. Rep.* DK-VLSI-81-13, *Knowledge Systems Area, Xerox Palo Alto Research Center.*

2. **A compact set of image processing primitives and their role in a successful application program.** J. Graham, C.J. Taylor, D.H. Cooper and R.N. Dixon, *Patt. Recog. Lett., 4: 325 - 333, 1986.* doi:10.1016/0167-8655(86)90053-X.

Pattern Recognition Letters 4 (1986) 325-333
North-Holland

October 1986

A compact set of image processing primitives and their role in a successful application program

J. GRAHAM, C.J. TAYLOR and D.H. COOPER
Wolfson Image Analysis Unit, University of Manchester, Stopford Building, Manchester M13 9PT, United Kingdom

R.N. DIXON
Visual Machines Ltd., Enterprise House, Manchester Science Park, Lloyd Street North, Manchester M15 4EN, United Kingdom

Abstract: We describe an integrated set of data structures and image processing primitives which enhance not only the run time efficiency of the application program, but also the programmability of the solution. This is illustrated by a program for clinical chromosome analysis.

Key words: Data structures, image representations, image processing, applications program, chromosome analysis.

1. Introduction

Despite three decades of research in image processing and computer vision there are few successful applications of the technology. By 'successful application' we mean a vision system doing a job in the 'real world' of the factory, or the hospital or whatever. The problem has been the relatively small body of workable techniques and the requirement for substantial computing power with its associated cost. The situation is now changing; computing power is now relatively inexpensive and the body of useful techniques is growing. The efficiency of generating solutions to application problems has become a limiting factor in their creation. Because even everyday problems in the 'real world' are highly complex and answers are required in real (or realistic) time, run time efficiency is also of great importance. A central issue in attaining efficiency both at run time and in programming is the selection of an appropriate set of data representations.

For some time we have held the belief that cost efficiency in successful application is achieved by using flexible hardware capable of being applied to a wide range of tasks – a general purpose machine whose market is not limited by target applications and whose production costs can be kept as low as possible. The required flexibility is achieved by making all operations programmable. The necessity for application programs to run in realistic times means that great attention must be paid to run time efficiency of algorithms. This may seem at odds with the need for efficiency of solution generation, but we have found that both can be achieved by the use of a compact set of data structures and primitive procedures integrated to form a 'universal tool kit' for generating application programs. We will illustate the use of this tool kit by referring to a successful application program in the field of chromosome analysis.

The machine on which this application is implemented consists of a fairly fast central processor (cycle time 150 ns) with access to three kinds of memory – image memory, micro memory – where its microcode instructions reside, and macro memory containing high level programs and data. Its architecture has been described in detail by Taylor et al. (1983).

0167-8655/86/$3.50 © 1986, Elsevier Science Publishers B.V. (North-Holland)

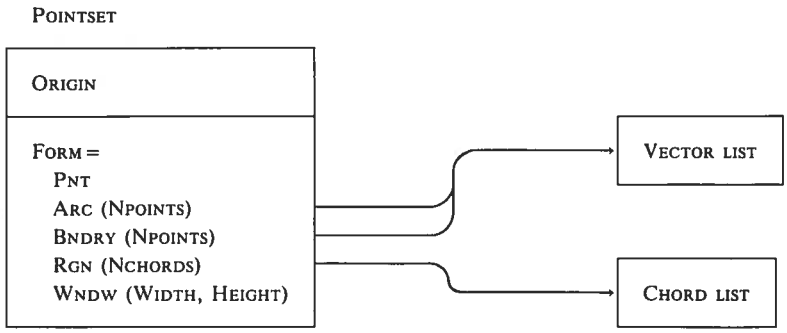


Figure 1. Pointsets are variant records (see text). Line and region pointsets are described by dynamically created vector and chord lists accessed by pointers.

cally. For these dynamic pointsets, the quantity of space occupied is kept as NPOINTS and NCHORDS respectively. A further useful feature is that the dynamically allocated space can be released when it is no longer required. Regions of interest change during the execution of a program, and the creation and disposal of pointsets gives us a way of accomodating these changes while conserving memory.

Many applications of image analysis involve making measurements on 'blobs' – irregularly shaped image components. Chromosome analysis is a good example of this. For such objects it is useful to have a description of both the boundary of the object, for measuring shape for instance, and the region, for things like integrated optical density. For this reason a data type called a BLOB is defined consisting of two pointsets – a boundary and a region.

2.2. *Images*

Pointsets describe regions of interest in images. The record describing an IMAGE is illustrated in Figure 2. An IMAGE allocates a block of image memory by defining its origin, least significant bit plane, number of bit planes and spatial resolution. Sometimes the region of interest is the whole image, although not all of the allocated space may be valid. To account for this, the IMAGE record contains a field called WHOLE – a pointset whose FORM is WNDW and whose width and height would initially be those of the whole image (typically 512×512). The dimension of WHOLE would be

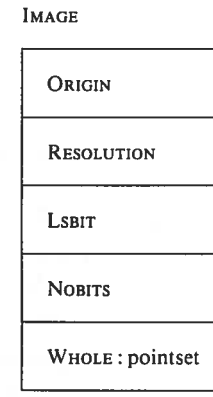


Figure 2. Image records describe physical blocks of image memory, addressable as bit planes. The pointset IMAGE.WHOLE is of form WNDW and represents the current meaningful area of an image.

automatically modified, for example, by the application of neighbourhood operations. The useful image size after application of a 3×3 linear filter is reduced from that of the original image by a one pixel border.

Images can be thought of in a variety of forms: grey level scenes, transformed images containing edge maps, label images, binary mask images, etc. All such images are logically equivalent, their description defining only a region in image memory. Whereas images define absolute positions in image memory (physical co-ordinates), pointsets are logical descriptions of connected sets of points which are applied relative to each image origin. Thus the same pointset referring to two different images will refer to the same geometric region of both images but to different regions of image memory.

2.3. Other data

One other data structure worthy of mention is the dynamic array. Much of the data arising from images is conveniently stored in arrays for the sake of random access, but the size of the arrays required is highly unpredictable. Vector lists and chord lists are examples of this type of data, as are collections of image sample values, histograms and profiles. On this one occasion we have had to create a non-standard Pascal implementation by turning off array bound checking to provide a convenient, memory efficient way of handling such data. Such arrays, of necessity, live in the dynamic data space.

3. Primitive operations

Given a basic set of data representations it is possible to define a repertoire of primitive operations which provide the basis for tackling a wide range of problems. It turns out that facility in defining grey-level and geometrical data via images and pointsets allows a fairly small set of operations to go a long way. A list of the procedures we have found most useful is shown in Table 2. They tend to involve performing the same operation at all points in a pointset. Thus the first essential is a mechanism for driving any operator through a pointset efficiently, whatever the form of the pointset. This is achieved by considering all pointsets as a collection of chords; the difference between the different pointset forms is then reduced to different methods of selecting the next chord. This is highly efficient for regions and windows, slightly less efficient for lines, which have proportionately fewer points, and rather inefficient for isolated points, whose use occurs only rarely.

For many operations it is only necessary to write a kernel for the operator which is then applied at all required points using the pointset driver. Linear filters provide a good example of this. All linear filters work in the same way – they visit all pixels of interest and replace the pixel's grey value with some linear combination of the grey values in its neighbourhood. If the number of pixels of interest is small compared with the whole image then substantial savings in processing time

Table 2
A list of fundamental image processing procedures. Unless otherwise stated the operations apply to any form of pointset. Most of the operations require either one or two image as arguments

IMAGE ARITHMETIC	– add, subtract, multiply
COPY	– image to image
DRAW	– a value into an image
HISTOGRAM	– of image values
JOINT PROBABILITY HISTOGRAM	
SAMPLE	– image values at each pixel
SLICE	– create a binary image by thresholding
SHADE CORRECT	
LINEAR FILTER	– convolution at all pixels
NON LINEAR FILTERS	– line operator, median filter, extremum filter
POINT TRANSFORMATION	– by table
BINARY IMAGE OPERATIONS	– erode, dilate, thin disconnect
GREY LEVEL EROSION, DILATION	
NEXT HIT	– find the next non zero pixel in a pointset
EXTRACT AN ARC	– starting at a hit point
EXTRACT A BLOB	– starting at a hit point
GENERATE A STRAIGHT ARC	– create a pointset
GENERATE A WINDOW	– create a pointset
ROTATE	– the contents of a window
DENSITY PROFILE	– projected onto a straight or curved arc
PROFILE FILTER	– one dimensional derivatives, etc.
CURVE FITTING	– straight line, cubic, circle
MEASURE	– area perimeter etc.

can be achieved by operating over pointsets. The convolution weights can be held in a table, thus a single procedure covers a range of cases from gaussian smoothing to a laplacian edge enhancer. Obviously non-linear filters may also be expressed as a kernel for the pointset driver, but the filter itself must be coded as part of the kernel. Examples include extremum filtering and an operator for detecting line structures (Dixon and Taylor 1979).

Most of the operations listed in Table 2 can take a binary image as their source image as readily as a grey image. The results tend to have particular meanings for binary images. For example multiplying binary images is equivalent to a logical AND operation and histogramming a binary image provides a detected area measure. For the sake of efficiency, certain operations, such as erosion, dilation and thinning are coded differently for binary and grey-value images.

Efficient execution of these fundamental operations is achieved by writing them as microcode procedures which can be called from high level language. Many of them are very compact indeed. The procedure to draw a value into a pointset consists of eleven microcode instructions including argument popping and jumps to the pointset driver. Slicing (creating a binary image) uses up 16 words of micromemory. The whole microcoded library fits within $3K$ words of microprogram memory. It is always possible to write specific kernels but this is seldom necessary. The metaphase finder, described later, contains one instance of specially crafted microcode where processing speed is of overwhelming importance. In general, enough efficiency is inherent in the operations that ease of programming is the dominant consideration and operations are built up of combinations of these primitives in high level language.

A commonly required process in image analysis is that of obtaining blobs, which is a useful illustration of this integrated software scheme. Blobs are obtained from binary images by searching within a pointset (such as IMAGE.WHOLE) for a non-zero pixel (NEXT HIT in Table 2) and extracting all connected pixels in the form of a boundary and chord list. If required, the BLOB.REGION can act as the pointset for another blob finding exercise, thus allowing indefinite nesting of blobs.

4. Application to chromosome analysis

Figures 3 and 4 show cells at metaphase viewed on a microscope slide at high and low magnification respectively. We have developed a chromosome anlaysis package which is designed to assist the clinician in analysing such cells and consists of two parts, a metaphase finder and a metaphase analyser. The sequence of actions performed by these components can be briefly described as follows.

The metaphase finder scans the slide field by field at low magnification in an attempt to identify metaphase cells for subsequent analysis. It does this by analysing local texture over the whole field and then focussing attention on promising areas to measure a figure of merit for potential metaphases

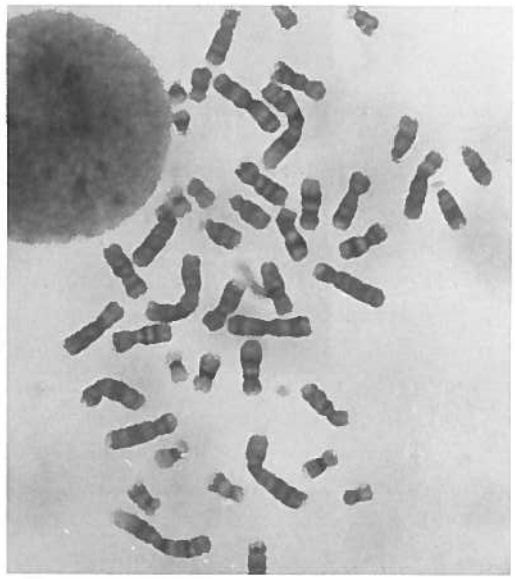


Figure 3. A metaphase cell viewed at high magnification showing individual chromosomes.

so they can be ranked in order of their quality for analysis.

The metaphase analyser examines the selected metaphases at high magnification with the aim of identifying and classifying the chromosomes. In Figure 3 the chromosomes can be seen as a set of darkish 'blobs' within which they grey values are rather variable due to a pattern of bands along the length of the chromosome. The analysis involves finding the positions of chromosomes by identifying dark patches in the image, first rather crudely and subsequently with more accuracy. Each chromosome has a characteristic constriction, called the centromere, which is a useful feature in classification, as is the distribution of density along the length of the chromosome. To both of these features it is necessary to find an axis for the chromosome and measure width and density profiles (Figure 6). Features extracted from these and other measures allow chromosomes to be classified into some 24 groups and presented to the clinician in the form of a karyogram – a tabular array in which each chromosome has a position corresponding to its class. Karyograms are highly useful tools in diagnosing chromosome disease.

What follows is not an attempt to describe the entire operation of the chromosome analysis pro-

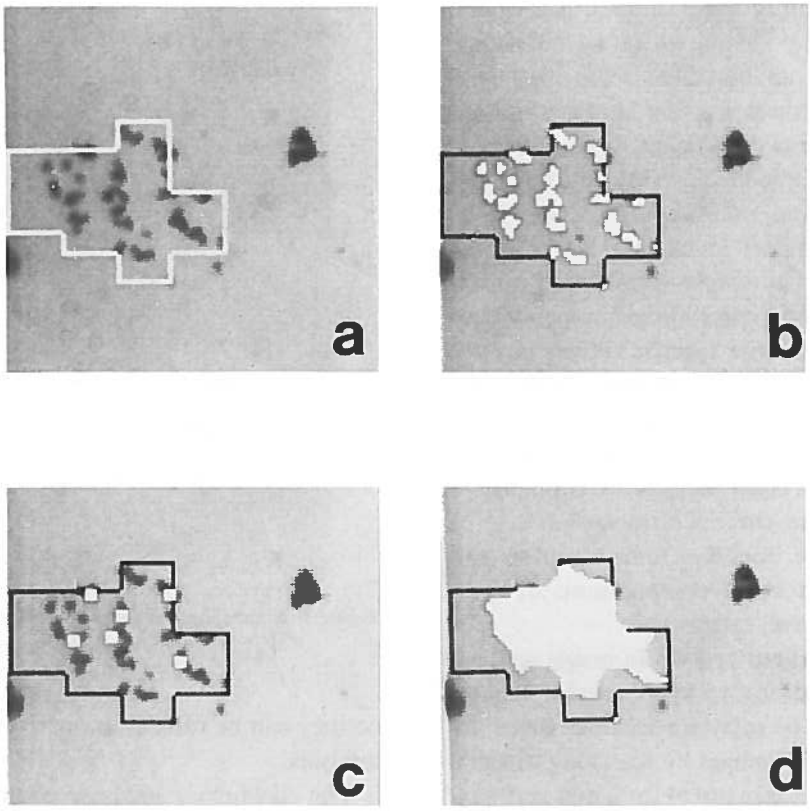


Figure 4. At low magnification an irregular region of interest is used in assessing metaphase quality (a). Thresholding (b) followed by binary opening (c) and closing (d) provide features for assessment. All of these operations take place within the irregular region shown superimposed in black in (b), (c) and (d).

gram, but a selection of operations from that program to illustrate the application and manipulation of the representations described earlier. Chromosomes themselves are very naturally described as blobs, their axes are linear pointsets. Irregularly shaped regions are a natural and efficient representation of the region of interest in metaphase finding.

4.1. Metaphase finder

The use of nested blob finding is particularly useful in the chromosome analysis software. In the metaphase finder, for instance, each field is first scanned by a specially microcoded texture analyser, the details of which we need not consider here. This results in the value 1 being drawn over window pointsets into a binary image at places where texture values are high. Clusters of these windows form irregularly shaped objects (Figure 4a) which are detected as blobs. Thresholding and slicing within the REGION pointset of these blobs allows the use of binary processing to evaluate the quality of the object as a metaphase. The binary image within the region of interest is opened and closed into separate binary images. Further blob finding within the irregular region allows a count of separate objects. The detected areas in the opened and closed images, determined by histogramming, give size and 'clumpiness' measures, and the resulting blob in the closed image provides an accurate size for the metaphase as a whole (Figure 4b,c,d).

The ability to restrict the binary processing to very small region of the searched area not only allows non-metaphase objects to be largely excluded from the processing, so improving accuracy, but keeps the processing time down to reasonable values (1–2 s per field), which is essential in this ap-

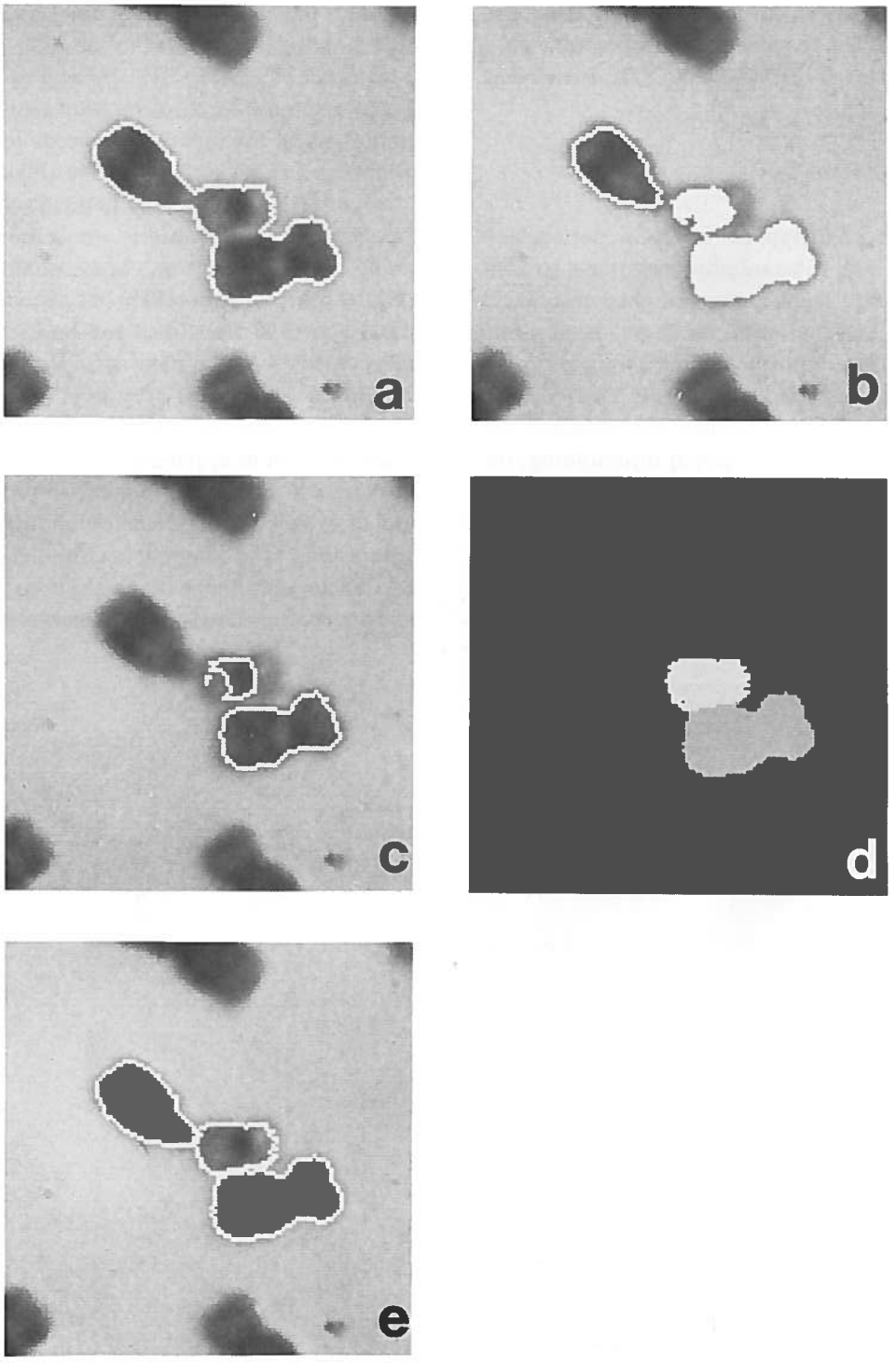


Figure 5. Composite chromosomes (a) can be split up using nested thresholding and blob finding (b,c). The solid region in (b) is further rethresholded and split to obtain final separation as in (c). The objects thus found can be labelled and re-expanded to their natural boundaries (d) to provide a correct segmentation (e).

plication. Piper and Rutovitz (1985) have also noted the efficiency gained from operating on a small, irregular region rather than a large enclosing window.

4.2. *Metaphase analysis*

In the case of metaphase analysis, nested blob finding can also substantially contribute to efficiency. Accurate segmentation of chromosomes is made difficult by the variations in grey level within the chromosome. Thresholds are obtained on the basis of local grey level histograms. The localities within which these histograms are obtained are determined from previous global thresholding; the presence of dark objects acts as a cue to direct the detailed processing. Even these locally obtained thresholds have a tendency to produce composite objects (Figure 5a). One approach to separating these objects is to rethreshold at successively lower (darker) thresholds until the object breaks up. The thresholding is done within the blob found at the previous level, making the splitting by rethresholding a recursive procedure; the area searched is minimal and the splitting proceeds only as far as necessary for each chromosome (Figures 5b,c).

The thresholds necessary to divide chromosomes frequently produce objects whose boundaries are highly unsatisfactory as chromosome boundaries (Figure 5c). Rutovitz (1978) has shown how a two-pass sequential transform can be used to expand these 'seeds' back to their natural boundaries, the expansion stopping at grey level minima between objects. In our software scheme this 'fall-set' transform can be applied over the pointset defining the original thresholded region, to produce an image of labelled objects corresponding to the best separation of the component chromosomes (Figure 5d). Slicing this image at thresholds corresponding to each of these labels in turn enables extraction of

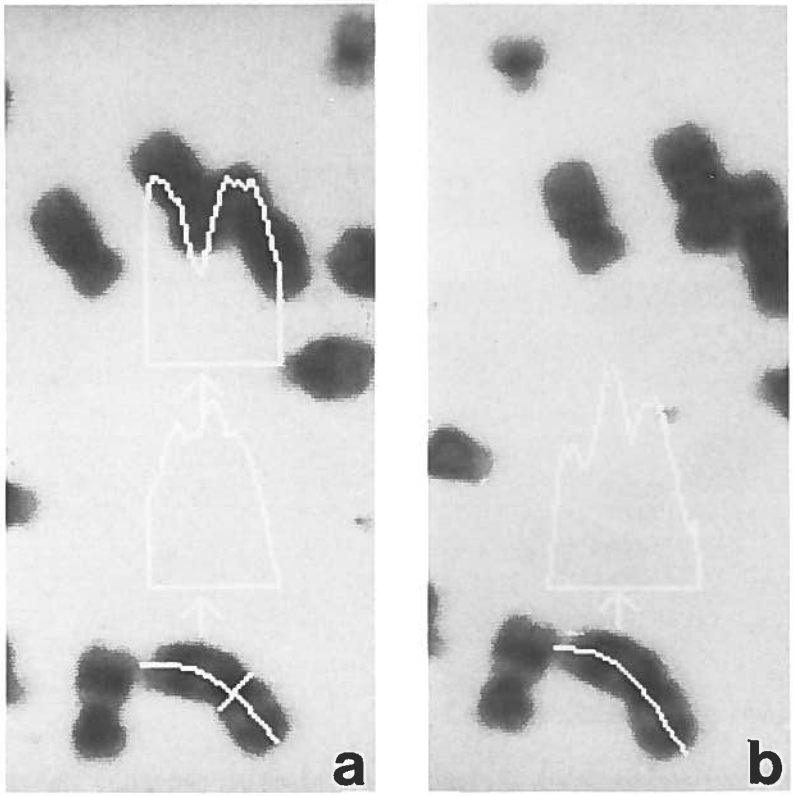


Figure 6. The chromosome axis is a pointset of form ARC. Profiles of the chromosome width on either side of the axis are used to find the centromere (a). Profiles of the density projected onto the axis (b) are used in classification. In each of the profiles the arrows indicate the centromere position.

blobs corresponding to each chromosome (Figure 5e).

A large part of the job of analysing a metaphase image is devoted to obtaining profiles of chromosomes along the axis. These may be width profiles for centromere location (Figure 6a) or density profiles for obtaining classification features (Figure 6b). They are obtained by projecting normals to the chromosome axis at each point and obtaining the appropriate value. The normals themselves are pointsets (ARCS).

Obtaining density profiles is such a commonly required facility that it is a standard library procedure (Table 2). The quantity measured in the width profiles is the squared euclidean distance from the axis point to the boundary of the object. The appropriate boundary point is found by searching along the normal in a masked image until a non zero pixel is hit (NEXT HIT in Table 2). This is the same procedure that is used to find the first pixel of a blob in a window or region pointset during blob finding. In this case the search takes place along an arc pointset.

It is worth noting that pointsets can be used in ways other than as means of addressing image data. They can be data objects themselves: the chromosome axes are calculated by fitting least squares straight lines or cubics to the boundary pointsets and chromosome size is measured as the area of the region pointset. They are also very useful as graphics objects during operator interaction to display regions of interest and graphical constructs.

5. Discussion

We have described a small set of low level image processing algorithms integrated with a few high level data structures for describing images and their components. This combination provides a framework within which efficient solutions to complex image analysis problems can be obtained using modest (and moderately priced) computing resources. The important point is that the high level symbolic representation of the image is integrated with the low level operations on the image values, mainly via the linkage of the pointset data structure to an efficient mechanism for using that data structure to access images. Thus efficiency of programming is enhanced by programming largely in symbolic terms while run time efficiency is maintained by restricting processing to those areas of the image which are of interest.

Chromosome analysis is a topic with a long and noble history in computer vision. The program used as an illustration here fulfills our criteria as a successful application insofar as it is in routine use or clinical trial in a number of clinical and commercial laboratories (e.g. Philip and Lundsteen, 1985). In arriving at a solution to the chromosome analysis problem, efficiency in generating the software has been as important as runtime efficiency. The software regime described here enabled almost all of the programming to be done in a highly structured high level language without serious sacrifice of execution speed. It should be stressed that programming in a highly structured language with the constraints of strong data typing has never hindered flexibility of the software. On the contrary, it has been a distinct advantage.

In the design of image processing hardware, it is all too often the case that little attention is paid to programmability. We believe we have shown that attention to this topic pays dividends.

References

Dixon, R.N. and C.J. Taylor (1979). Automated asbestos fibre counting. In: *IOP Conference Series No.* 44, Machine Aided Image Analysis, Institute of Physics, Bristol.

Marr, D. (1982). *Vision.* Freeman, San Francisco.

Philip, J. and C. Lundsteen (1985). Semiautomated chromosome analysis – A clinical test. *Clinical Genetics* 27, 140–146.

Piper, J. and D. Rutovitz (1985). Data structures for image processing in a C language and Unix environment. *Pattern Recognition Letters* 3, 119–129.

Rutovitz, D. (1978). Expanding picture components to their natural boundaries by propagation methods. The notion of the fall set and fall distance. *Proc. IV IJCPR*, Kyoto Japan pp. 657–664.

Taylor, C.J., J.N.H. Brunt, R.N. Dixon and P.J. Gregory (1983). Designing and implementing an algorithm to extract motion from images. In: O. Bradick and A. Sleigh, Eds., *Physical and Biological Processing of Images*, Springer, Berlin.

3. **System architectures for interactive knowledge-based image interpretation.** C.J. Taylor, J. Graham and D. Cooper, *Phil. Trans. Roy. Soc., Lond. A324, 451 – 465, 1988.* doi: 10.1098/rsta.1988.0033

Downloaded from http://rsta.royalsocietypublishing.org/ on February 19, 2015

Phil. Trans. R. Soc. Lond. A **324**, 457–465 (1988)
Printed in Great Britain

System architectures for interactive knowledge-based image-interpretation

By C. J. Taylor, J. Graham and D. Cooper
Department of Medical Biophysics, University of Manchester, Stopford Building, Oxford Road, Manchester M13 9PL, U.K.

[Plates 1 and 2]

We discuss hardware and software architecture for automated image-interpretation. The importance of considering the complete system is emphasized leading in particular to the conclusion that high-level and low-level processing are intimately linked. We present arguments to support the idea that automated image-interpretation systems should be knowledge-based and interactive. We attempt to identify the main architectural problems which such systems must address and outline a systematic strategy for acquiring, structuring and using knowledge.

Introduction

In this paper we consider the architectural issues raised by systems for automated image-interpretation. System architecture involves both hardware and software and we attempt to discuss their interrelation, arguing in the end that software problems are the more crucial. We begin by considering the broad characteristics of the image-interpretation task and identify high-level and low-level processing as qualitatively different activities. We review some of the specialized hardware configurations that have been developed to deal with the large quantities of data contained in images, highlighting the tendency for low-level processing to be considered in isolation. Against this background we discuss the desirability of a knowledge-based approach and suggest that interaction between the user and the system is also an important issue. We attempt to identify the central architectural problems and argue that they are not amenable to solution simply by the application of massive computing power but rather are truly architectural in nature and require a strategy for selecting, structuring and using information. Finally we propose a systematic approach to some of the fundamental problems that have been identified. We describe this as a software architecture to emphasize that architecture is not concerned solely with hardware. The architecture makes use of explicit image models and is the subject of current research.

Where possible we have used, as illustrations, practical examples from our own work in medical and industrial image-interpretation. We believe, however, that the ideas are perfectly general and have attempted to relate them to remotely sensed imagery where possible.

The first illustrative problem is chromosome analysis. Figure 1*a* shows the genetic material from a single human cell arranged, as it is during cell division, into bodies of characteristic size and shape called chromosomes. When such images are used clinically to detect genetic abnormalities, a technician uniquely identifies each of the individual chromosomes by its size, shape and pattern of stain uptake (banding pattern). They are normally arranged in a regular

The Royal Society is collaborating with JSTOR to digitize, preserve, and extend access to *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences.* www.jstor.org

Downloaded from http://rsta.royalsocietypublishing.org/ on February 19, 2015

display called a karyogram (figure 1*d*). A machine that could take such cells and automatically generate a karyogram would be ideal. In practice, a semi-automated, interactive system is clinically useful (Graham 1988).

The second illustrative problem is automated industrial inspection, specifically the inspection of motor-car drum-brake assemblies (figure 2). In this case the task is to recognize and locate each of the component parts (which are moveable) and check that they are present, correctly fitted and undamaged (Woods *et al.* 1987).

Levels of processing

It is convenient to identify two qualitatively different types of processing that are required for automated image-interpretation. Low-level processing is primarily numerical and acts directly on image data. Examples are geometric correction, stereo ranging, homogeneous region extraction and edge-detection. In general, low-level processing takes one or more images as input and produces a different image as output. Typically, the result at a point in the output image depends on the intensity values in a small neighbourhood surrounding the corresponding point in the input image. An example is edge-detection which is shown in figure 2*b*, *c*.

High-level processing is largely symbolic and involves recognizing and describing image structures. Typically this involves matching the observed data to some model of the expected appearance of known structures. For instance, for the brake assembly we might store an idealized edge map for each subcomponent and attempt to match these to the edges detected in an observed image. In remote sensing a comparable example might involve the use of map data as a model to which observed data must be related.

High-level and low-level processing present very different characteristics. Low-level processing typically involves very simple repetitive operations performed on large data sets, whereas high-level processing is essentially symbolic and involves complex processing on relatively small data sets. It is important to recognize that both types of processing are involved in automated image-interpretation and to reflect this in system architecture.

Specialized hardware

In this section we consider some of the hardware arrangements that have been developed to address specific problems posed by image-processing. Much of the effort has concentrated on low-level processing where significant computing power must be brought to bear, though we argue later that these problems should not be considered in isolation.

Coprocessor systems

Figure 3*a* shows the arrangement of a coprocessor system (see, for example, Taylor *et al.* 1986). In such a system a relatively small number of specialized processors co-operate to act upon image data and other types of data held in memory. Operations such as data-processing, memory-address calculation and program-flow management can be performed simultaneously by processors specializing in these activities. The coprocessors can be designed and programmed to make selected primitive operations very efficient and can thus achieve realistically high performance for low-level processing. Such an arrangement is, however, extremely flexible and can also be designed to perform high-level processing efficiently within the same structure.

Downloaded from http://rsta.royalsocietypublishing.org/ on February 19, 2015

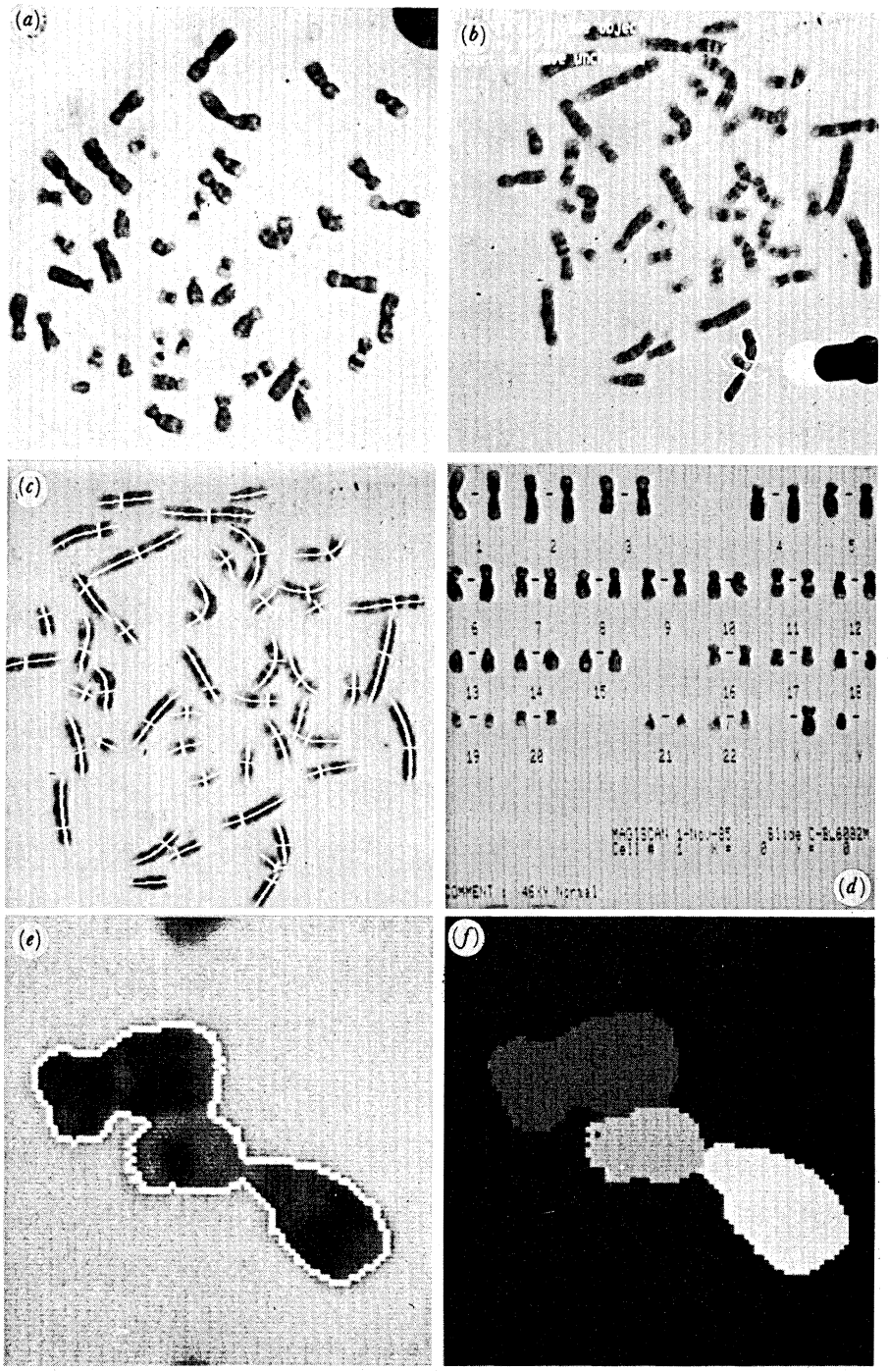


FIGURE 1. Chromosome analysis. (*a*) Microscope image of a dividing human cell; (*b*) user interacting with an automated analysis system to separate overlapping chromosomes; (*c*) axis and centromere automatically located for each chromosome; (*d*) automatically generated karyogram; (*e*) erroneous initial hypothesis for a chromosome boundary; (*f*) modified hypothesis generated as a result of obtaining further low-level evidence.

(*Facing p.* 458)

Downloaded from http://rsta.royalsocietypublishing.org/ on February 19, 2015

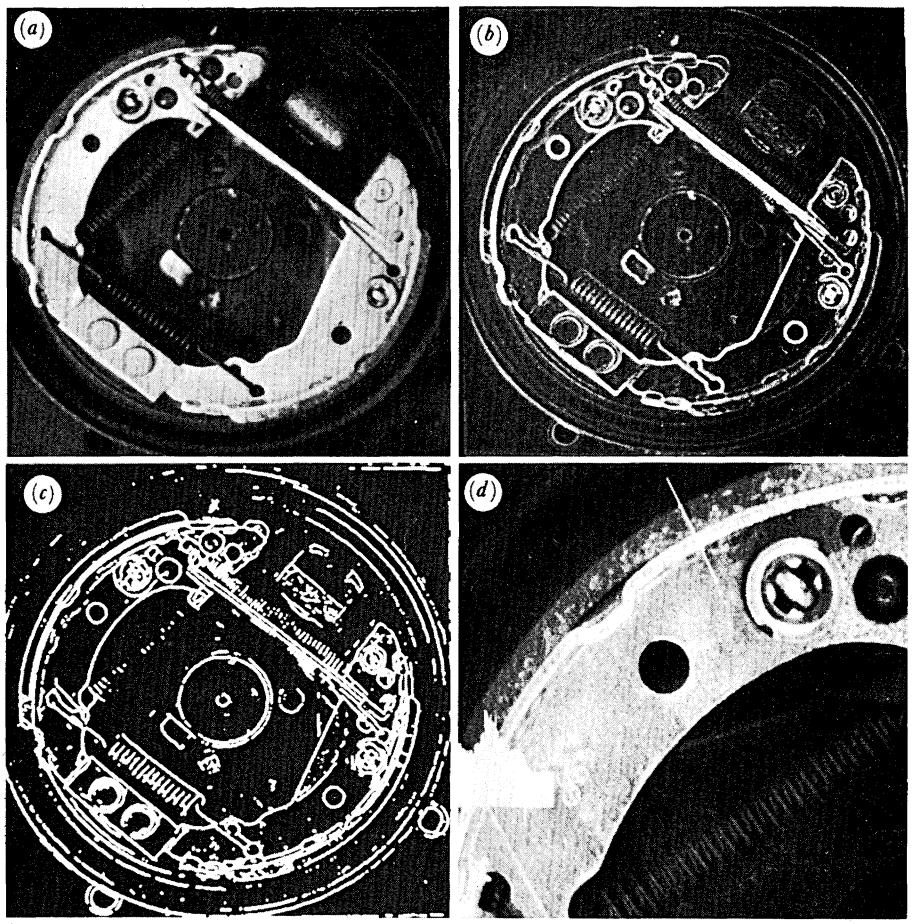


FIGURE 2. Brake inspection. (*a*) Plan view of a motor-car drum brake assembly; (*b*) edge strength image of (*a*); (*c*) detected edges from (*b*); (*d*) line which defines the position at which brake lining thickness should be measured. The intensity profile along this line is displayed and markers on the line indicate the points between which the system intends to make the measurement.

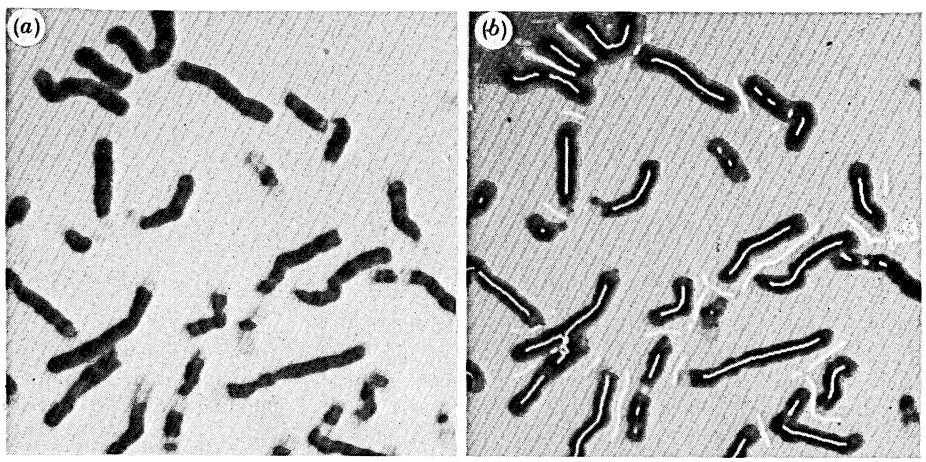


FIGURE 6. Centre of symmetry cues. (*a*) Original image-containing chromosomes; (*b*) loci of intensity symmetry detected directly from the intensity image.

Downloaded from http://rsta.royalsocietypublishing.org/ on February 19, 2015

Pipeline processors

A simple arrangement, which achieves high performance for low-level processing tasks, is shown in figure 3*b* (see, for example, Gerritsen & Monhemius 1981). The pipeline processor makes use of the fact that in low-level processing an identical sequence of operations must often be performed on a large number of pixel values. Each processor in the pipeline undertakes one step in the sequence of operations. Each pixel value is sent down the line in turn so that if there are n stages in the pipeline n pixels are operated on at once. If all the processors have the same speed then the system is n times as fast as an individual processor. Once high-level processing is involved the arrangement offers virtually no assistance.

Array processors

The arrangement of an array processor is shown in figure 3*c* (see, for example, Duff, 1979). These systems make use of the characteristic of many low-level algorithms that they involve operations on small neighbourhoods of pixels. In this radically different arrangement there is a processor and memory element for each pixel. Each processor–memory element is connected to each of its neighbours so that it can act upon its own and neighbouring pixel values. Thus by sending the same instructions to each processor an operation such as edge-detection can be performed for each pixel in the image simultaneously. Again, although this configuration is well suited to low-level processing it is inappropriate for high-level processing because the connections between processors are insufficiently general.

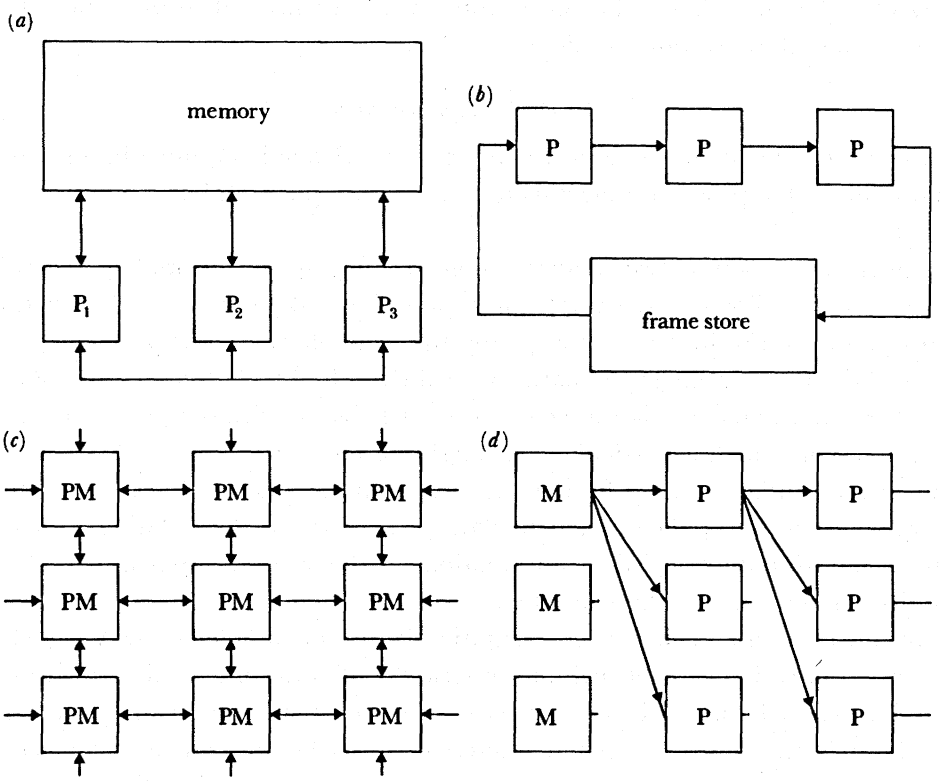


FIGURE 3. Specialized hardware for image processing. (*a*) Coprocessor system; (*b*) pipeline processor; (*c*) array processor; (*d*) adaptive network. P, processor; P_t, specialized processor; M, memory.

Downloaded from http://rsta.royalsocietypublishing.org/ on February 19, 2015

Adaptive networks

Adaptive networks (figure 3*d*) bear a superficial resemblance to the organization of the brain and represent a more radical architectural approach (see, for example, Hopfield 1982). Here the processors only ever execute one type of operation, combining the set of inputs according to a predetermined rule to produce an output. The nature of the processing that takes place is thus determined not by the processors but by the pattern and strength of the connections between processors. In principle, the final output of such a system can be either another image or a symbolic interpretation. Adaptive networks are very interesting but so far no-one knows much about how to set up the connections to achieve the desired result or indeed about their generality. A particular concern is that knowledge of the problem domain must necessarily be incorporated implicitly into such systems, a characteristic which, as we argue in the next section, is undesirable.

Interactive knowledge-based interpretation

Images from a particular domain, such as remote sensing, are normally interpreted by making use of specialized knowledge. This knowledge includes the nature of the interpretation task together with the identity and expected appearance of key objects and structures. Automated systems have been developed to interpret complex images (Aleksander 1983; Brunt *et al.* 1983; Dixon & Taylor 1979; Graham 1987; Pycock & Taylor 1980; Thomason 1986; Tucker & Shippey 1983; Woods *et al.* 1987) but generally the prior knowledge involved has not been easily identifiable; it has rather been implicit and embedded in particular algorithmic approaches to the interpretation problem.

Knowledge representation

A knowledge-based system is one in which prior knowledge is supplied in an explicit form, quite separate from the program that puts it to use. This approach has a number of potential advantages:

(i) the clear separation between application-specific knowledge and a general-purpose interpretation engine allows more complex interpretation tasks to be tackled reducing the cost and engineering expertise involved in applying the technology to new problems;

(ii) knowledge can be applied systematically as the result of an automated reasoning process whereas systems that make implicit use of prior knowledge require the programmer to foresee the circumstances in which a particular fact might be relevant;

(iii) knowledge can be acquired from a number of sources and used in a coherent manner; thus in a remote-sensing application, map data, expert knowledge relating to feasible configurations of land-use and statistical information on the size and appearance of regions of known land-use might be usefully combined.

An important issue in knowledge representation is that of completeness. A model is a form of representation in which geometrical and intensity configurations are described in sufficient detail that images of feasible objects and structures (or their significant features) may be generated. Models are thus of particular interest because they provide a means by which an automated image interpretation system may arrive at a complete explanation of each observed image (Ayache & Faugeras 1986; Brooks 1981; Hanson & Riseman 1978; Pollard *et al.* 1987).

Downloaded from http://rsta.royalsocietypublishing.org/ on February 19, 2015

In addition to a method of knowledge representation a practical system requires a means by which the user can supply relevant knowledge. Although some knowledge can be presented to the system formally (in a manner analogous to conventional programming) there is also a need for interactive dialogue between the user and the machine. Two types of interaction may be considered. The fact that both involve the user gaining rather direct access to the geometrical and intensity structure of an image provides additional support for a model-based knowledge representation that can form a natural link between the user interface and the internal-processing régime.

Offline interaction

Although some knowledge can be stated simply, a great deal of what is used to interpret images can only be conveyed conveniently by showing examples. This is indeed the method we often use when explaining a complex visual task to another person. The requirement is that the user can present example images and interact with the system in such a way that it extracts salient features. We show this with an example from the interpretation of brake assembly images.

In figure 2*d* the system has identified the position at which the thickness of the brake lining is to be measured. The intensity profile observed along the measuring line does not, however, allow the inner and outer edges of the lining to be easily identified. The figure shows the system making an educated guess at the location of the inner and outer boundaries. The user accepts the guess if it is accurate or corrects it with a lightpen if it is erroneous. This is repeated for a number of examples and because, in each case, the system has both the observed intensity profile and the correct interpretation it can, by building an internal model, learn how to guess correctly so that no interaction is required at run-time.

Online interaction

In many circumstances it may be desirable for the user to provide additional knowledge at run-time. This may be appropriate because complete automation is too difficult or because a user-guided system is more appropriate for the task in hand. We show this with an example from chromosome analysis (figure 1). Here the generation of a karyogram can be significantly automated but occasionally configurations arise that are difficult for the system to resolve. In any case, because the system is used diagnostically, skilled supervision is desirable. The system starts by determining a boundary for each chromosome. Figure 1*b* shows how the user can separate overlapping chromosomes with the lightpen. To decide which chromosome is which, an axis of symmetry must be obtained for each and used to define a path along which the banding pattern should be measured and the position of the centromere located. Figure 1*c* shows a display of the results, which can again be corrected with the lightpen. Finally the chromosomes are classified and displayed in a karyogram, which can be modified if necessary by moving chromosomes with the lightpen (figure 1*d*).

The architectural problem

Having identified some of the important characteristics of an ideal image-interpretation system we can consider some of the architectural issues which are raised. First let us look at the question of computational complexity. It is often argued (Sternberg 1980) that this is a particularly important issue in low-level processing because of the large quantities of data

Downloaded from http://rsta.royalsocietypublishing.org/ on February 19, 2015

involved. Figure 4 shows with a simple example that there is also a problem with high-level processing. The figure shows a line-drawing model of a house, composed of straight-line segments and an example image from which we may assume all the line segments can be perfectly recovered. To recognize the house we must match the model and the example line by line. The number of ways to do this goes as the factorial of the number of line segments and in this example there are approximately 10^{20} ways of matching. To recognize the house in one second by using brute force would require computing power of at least 10^{14} Mips (million instructions per second). To put this in perspective NASA's massively parallel processor provides approximately 10^4 Mips (Batcher 1980). Thus it is unrealistic to assume that the application of parallelism will solve the problem, neither can it be avoided because any of the matches might be the best one. What is required is a framework within which the system can focus its attention on likely interpretations avoiding computational effort being wasted on those which are unlikely.

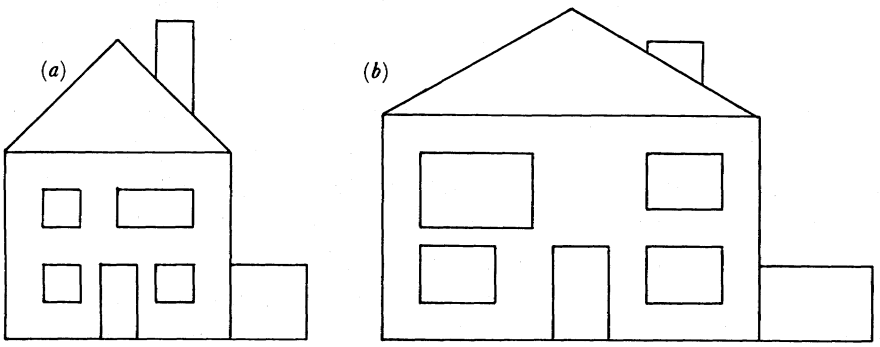


FIGURE 4. (*a*) Idealized house model composed of straight line segments; (*b*) the line segments which might be obtained for an example of a house image.

To appreciate fully the nature of the architectural problem with which we are faced we must address a further complication. In our discussion of model matching we have assumed that low-level processing may be used to extract evidence of structure (e.g. line segments) which may subsequently be interpreted by matching to a model. In practice this is unrealistic because the evidence of structure obtained by unguided low-level processing is subject to errors and will not in general represent the best evidence that could be sought in support of an emerging interpretation hypothesis. A more realistic approach is to undertake some low-level processing to provide sufficient evidence to make an initial interpretation which can then be used as an organizing hypothesis to guide the collection and interpretation of further low-level evidence. This further evidence may lead to the rejection, modification or refinement of the original hypothesis. Figure 1*e*, *f* show an example in chromosome analysis where initial low-level processing identifies a composite object. From its shape this is recognized as unlikely to be a single chromosome and further low-level evidence is sought which generates a modified hypothesis of three objects (Graham *et al.* 1986).

In summary there are two main conclusions which we draw from this discussion:

(i) high-level and low-level processing are intimately linked and the interface between them is crucial;

(ii) real visual tasks involve massive computational complexity and the central issue is that of deploying the available computing resources to best effect rather than increasing them.

Downloaded from http://rsta.royalsocietypublishing.org/ on February 19, 2015

We (and others) have designed coprocessor hardware which optimizes the interface between different levels of processing (Taylor *et al.* 1986; Graham *et al.* 1986) and have investigated strategies for focusing computing resources in a number of problem domains (Dixon & Taylor 1979; Brunt *et al.* 1983; Pycock & Taylor 1980; Woods *et al.* 1987; Graham 1988).

A SOFTWARE ARCHITECTURE

In previous sections we have argued that the problem of image-interpretation should be posed as that of explaining an observed image in terms of an explicit image model. The model will describe the objects and structures that may appear in the image and the relations between them. It will be parametrized so that a particular set of parameter values uniquely defines the geometric and (ideally) intensity configuration of a particular feasible image. Typical parameter values and measures of variability will have been obtained by offline interaction with example images. Image-interpretation involves a search for that set of model parameters that are consistent with those observed in example images and that define an image most similar to the observed image.

Hierarchical models

We showed in the previous section that the cost of establishing a correspondence between model elements and image structures rises exponentially with the number of model elements. To model visual worlds of realistic complexity requires that the matching problem be divided into a number of subproblems of manageable cost. This can be achieved in a natural way by organizing the model as a hierarchy. Figure 5 shows the manner in which the house model of figure 4 might be broken down into a number of submodels. In this example a partition of the model such as BODY will contain information describing the expected geometrical relations between submodels such as WINDOW and WALLS. Each submodel at the lowest level describes the relations between the line segments of which it is composed. If we can match each submodel independently the number of segment to segment matches which must be considered is reduced from approximately 10^{20} to approximately 10^3.

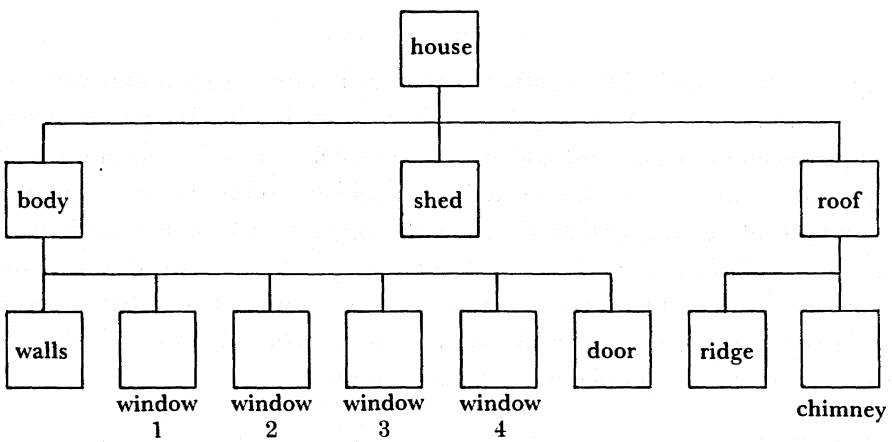


FIGURE 5. The house model of figure 4 partitioned into a hierarchy of submodels.

Downloaded from http://rsta.royalsocietypublishing.org/ on February 19, 2015

Cues and control

A cue is a feature that suggests that a particular model element should be instantiated. Given a hierarchially organized model and an observed image it is generally necessary to apply low-level processing to generate cues that provide evidence with which to initiate the matching process. In the house example we might detect intensity discontinuities (edges) in the observed image and starting from the bottom work upwards. Once a submodel has been matched its relation to other submodels can be used to limit the search space. For example, once WALLS has been recognized the approximate position of each WINDOW is defined and may be used as a cue which propagates the matching process.

It is important to recognize that it is not necessary to start matching at the lowest level of the model. In the house example we might, for instance, use a cue-generator that found dark blobs of about the size of a DOOR. On finding such a cue the system would try to recognize and locate the rectangle of the DOOR perhaps seeking edge information in the vicinity of the cue. If the DOOR were located the system would move up a level in the hierarchy and use the known relation between DOOR and WALLS to generate a WALLS cue. If the WALLS were successfully located then the WINDOWS could be cued. This process of cue-driven matching can continue until a complete match has been established. An important feature of this organization is that it is possible for the system both to infer higher-level models from details and details from higher-level models. Ultimately, however, the interpretation must be supported by direct evidence. It is also important to note that the pattern of control is determined, as seems appropriate, by the nature of the images to be interpreted.

Figure 6, plate 2 shows a practical example of a cue generator which operates robustly at a higher level than edges. The figure shows a chromosome image from which loci of intensity symmetry have been extracted directly without first detecting edges or separating objects from background. These lines represent candidate chromosome axes which can be refined and tested by appealing to other levels of a model to recognize chromosome boundaries, banding pattern and so on. Cues such as this which detect reasonably high levels of organization in the image are important because they will, as a result, tend to be less ambiguous.

CONCLUSIONS

In the paper we have argued that software is more important than hardware in automated image-interpretation systems. Most specialized hardware that has been built provides support for low-level processing but does not address the complete image-interpretation problem. The real issues are the manner in which knowledge is acquired, represented and used. The need is to develop methods of dealing systematically with application-specific knowledge. We suggest that explicit models of image structure offer a good means of representing knowledge internally and allow the user to interact with the system in a powerful way. The software architecture that we have outlined is the subject of current research and is believed to offer the basis of a solution to some of the architectural problems we have identified.

We thank P. W. Woods for providing pictures of the brake inspection application, and S. A. Thornham and P. J. Azzopardi for their assistance in preparing other figures.

Downloaded from http://rsta.royalsocietypublishing.org/ on February 19, 2015

References

Aleksander, I. 1983 *Patt. Recog. Lett.* **1**, 375–384.
Ayache, N. & Faugeras, O. D. 1986 *IEEE Trans.* **PAMI 8**, 44–54.
Batcher, K. 1980 *IEEE Trans Comput.* **29**, 836–840.
Brooks, R. A. 1981 *Artif. Intell.* **17**, 285–348.
Brunt, J. N. H., Taylor, C. J. & Dixon, R. N. 1983 In *Physical techniques in cardiological imaging* (ed. M. D. Short), pp. 153–162. Bristol: Hilger.
Dixon, R. N. & Taylor, C. J. 1979 In *Machine aided image analysis, 1978 (IOP Conference Series no.* 44) (ed. W. E. Gardner), pp. 178–185.
Duff, M. J. B. 1979 In *Advances in digital image processing* (ed. P. Stucki), pp. 265–276. New York: Plenum Press.
Gerritsen, F. A. & Monhemius, R. D. 1981 In *Languages and architectures for image processing* (ed. M. J. B. Duff & S. Levialdi), pp. 189–203. London: Academic Press.
Graham, J. 1988 *Analyt. Quant. Cytol. Histol.* (In the press.)
Graham, J., Taylor, C. J., Cooper, D. H. & Dixon, R. N. 1986 *Patt. Recog. Lett.* **4**, 325–333.
Hanson, A. R. & Riseman, E. M. 1978 In *Computer vision systems* (ed. A. R. Hanson & E. M. Riseman), pp. 303–333. Orlando, Florida: Academic Press.
Hopfield, J. J. 1982 *Proc. Natn. Acad. Sci. U.S.A.* **79**, 2554–2558.
Pollard, S. B., Porrill, J., Mayhew, J. E. W. & Frisby, J. P. 1987 *Image Vision Comput.* **5**, 73–78.
Pycock, D. & Taylor, C. J. 1980 *Analyt. Quant. Cytol.* **2**, 195–202.
Sternberg, S. R. 1980 In *Real-time medical image processing* (ed. M. Onoe, K. Preston & A. Rosenfeld), pp. 11–22. New York: Plenum.
Taylor, C. J., Dixon, R. N., Gregory, P. J. & Graham, J. 1986 In *Intermediate-level image processing* (ed. M. J. B. Duff), pp. 19–34. London: Academic Press.
Thomason, R. L. 1986 In *Vision Conference Proceedings, Detroit, 1986*, pp. 5:51–5:61.
Tucker, J. H. & Shippey, G. 1983 *Analyt. Quant. Cytol.* **5**, 129–137.
Woods, P. W., Taylor, C. J., Cooper, D. H. & Dixon, R. N. 1987 *Patt. Recog. Lett.* **5**, 11–17.

Discussion

D. Lane (*Intelligent Automation Laboratory, Department of Electrical and Electronic Engineering, Heriot-Watt University, Edinburgh, U.K.*). In Dr Taylor's presentation he mentioned the subject of feedback, and described the way that high- and low-level processes may interact. Feedback and the ensuing issue of system stability have been much studied by mathematicians and control engineers for a number of years. Mathematical tools (pole-zero diagrams, nyquist plots, bode diagrams) have been developed to enable a designer to predict the stability margins of a system. Is the stability issue relevant in the context of knowledge-based image-interpretation, and if so, how may it be approached?

C. J. Taylor. The issue of stability is relevant, but the system with which we are dealing is much more complicated than those for which the mathematical tools mentioned were developed. Feedback is involved in knowledge-based image interpretation in the sense that high-level hypotheses may be used to guide the search for low-level supporting evidence which in turn may be used to modify the high-level hypotheses; a high-level interpretation is stable but it is not necessarily so that such a system will converge to that interpretation. The interaction between high-level and low-level processing is, however, sufficiently complex that it is difficult to see how the methods of control engineering can easily be applied.

4. **Boundary cue operators for model-based image processing.** J. Graham and C.J. Taylor, *Proceedings of the fourth Alvey Vision Conference, Manchester, 1988, pp 59 - 64.* doi:10.5244/C.2.10

Boundary Cue Operators for Model Based Image Processing

J. Graham and C.J. Taylor
Wolfson Image Analysis Unit
Department of Medical Biophysics
University of Manchester
Manchester M13 9PT
U.K.

An efficient method of using explicit shape models of objects in boundary instantiation is to apply one dimensional edge searches in locations where boundaries are likely to occur. In many important cases, linear edge operators produce at best only weak responses. We investigate here the use of three different statistical measures applied over a sliding 'dipole' as candidates for detecting weak boundaries. Their performance is compared with an implementation of the Canny operator as a benchmark on synthetic images of step edges in random noise and on certain difficult real images. In the former case their performance compares favourably with the Canny operator, while in the latter case they can produce significant responses where the Canny operator detects only weakly or not at all.

INTRODUCTION

In most applications of computer vision and image processing, the correct location of the boundaries, between different objects, or between object and background is of central importance in achieving a correct image interpretation. The literature abounds with methods for detecting these boundaries, which make use either of the different properties of the regions on either side of the boundary, or the fact that the boundary is characterised by a pronounced grey-level discontinuity or edge.

The edge based approach is much favoured in interpretation of unconstrained three dimensional scenes, where the properties of regions may not easily be predicted. Region based approaches are often used in cases when the image is more constrained, and may be considered to be two dimensional, e.g. in remote sensing or microscopy. Both approaches are based on models of the world which are acknowledged to be flawed. Region properties tend to be less well-defined near the very boundaries they are used to detect, and edges are often weaker on true boundaries than at other, semantically irrelevant, points. Both region and boundary methods tend to be applied without reference to high level knowledge concerning the likely location and properties of boundaries.

Recent experience in our group has shown that considerable improvements in boundary detection can be made by directed, one-dimensional edge detection. The direction comes from a model of what is expected in the image, providing a prediction of the positions and orientations of expected boundaries. The exact boundary locations are determined by one dimensional edge searches across the predicted boundary. The confidence in a detected edge point can be assessed by reference to local and global models of the expected edge. This approach has produced very encouraging results in application fields as disparate as industrial inspection[1] and histology of muscle sections[2] . In order to make this type of analysis applicable to a wide range of applications, we require a robust boundary cue locator which operates by one dimensional search, avoids the problem of ill defined edges and which does not depend critically on the nature of the boundary. In this paper we describe some operators which approach this requirement by measuring properties of the distribution of image values on either side of the boundary. We show that using this approach, boundary detection performance can be as good as or better than optimal methods of edge primitive detection in terms of sensitivity and accuracy, while allowing the flexibility of being adapted to local models of the image.

BOUNDARY DETECTION OPERATORS

Given a prediction of where to look for a boundary, its correct position is located by searching along a line perpendicular to its putative orientation. To increase signal to noise ratio, it is best to integrate the response across some width perpendicular to this line. The search therefore takes place within an elongated rectangle, and we are seeking a partition of the rectangle along its length which produces the two most distinct distributions of image values. In order to make appropriate comparison of the distributions of image values on either side of the boundary, it is important that equivalent areas are sampled. It is also necessary to avoid confusion due to the inclusion in the sampling of nearby boundaries with other regions. The detector we have used is a "dipole" consisting of a rectangular box partitioned

AVC 1988 doi:10.5244/C.2.10

into two poles, whose length and width can be varied according to the grey level and geometrical model of the expected edge. This dipole is scanned across the edge; at each point on the scan the distributions in either pole are sampled and compared in ways described below.

Three different statistics have been implemented for the comparison of the two poles: the entropy, the standard deviation and the mean of the distributions.

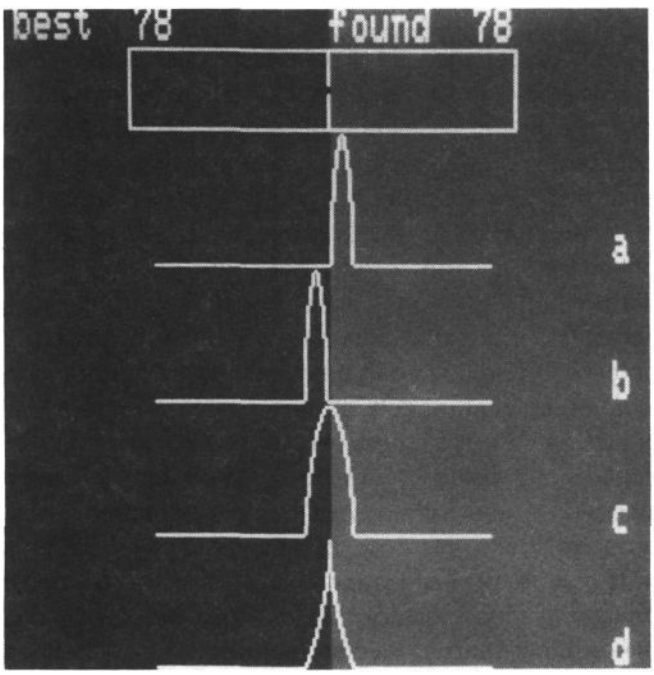


Figure 1. The entropy dipole response to a perfect step edge. The dipole (in this case of half length 10 pixels) is scanned across the window at the top (width 30 pixels). The mark at the top of the window indicates the true edge position ("best"), and that at the bottom the edge position located by the dipole ("found"). The entropy value in pole A (E_A) as it crosses the edge is shown in trace a, that of pole B (E_B) in trace b, and that of the whole dipole (E_t) in trace c. Trace d shows the response of the operator measure $2 \times E_t - (E_A + E_B)$.

Entropy

The entropy of a probability density function is given by $E = -\Sigma\ p_i\ \ln(p_i)$ where p_i is the probability of occurrence of state i. It has frequently been used as a threshold selection measure in region-based segmentation, where its usefulness lies in the fact that it acts as a measure of "peakiness" or compactness of a distribution. A very narrow distribution of states gives a low value for E, whereas as broad distribution of roughly equally populated states gives a high value. When a distribution is being divided into two distributions on either side of a threshold, for example, the division which minimises the sum of the two entropies produces the intuitively optimal result. In our case we are dividing a distribution not by a threshold, but by spatially partitioning the area from which it is sampled. However, the principle of producing the most compact distributions from the region on either side of the partition is still a useful one.

Figure 1 shows the behaviour of E_t, E_A and E_B as the dipole is scanned across a simple step edge, where E_A and E_B are the entropies in poles A and B, and E_t is the total entropy in the window. Fortunately, entropy is a self-normalising measure. The entropies in each of the poles rises as that pole crosses the edge; that is E_A has a maximum on the right hand side of the edge, E_B has a maximum on the left hand side of the edge. Both have low values (0 in this ideal case) when the partition is on the edge. E_t on the other hand has a maximum when the partition is on the edge. We calculate the signal $2 \times E_t - (E_A + E_B)$ which rises sharply to a maximum on the edge. Figure 2a shows the response of the entropy dipole to a noisy edge

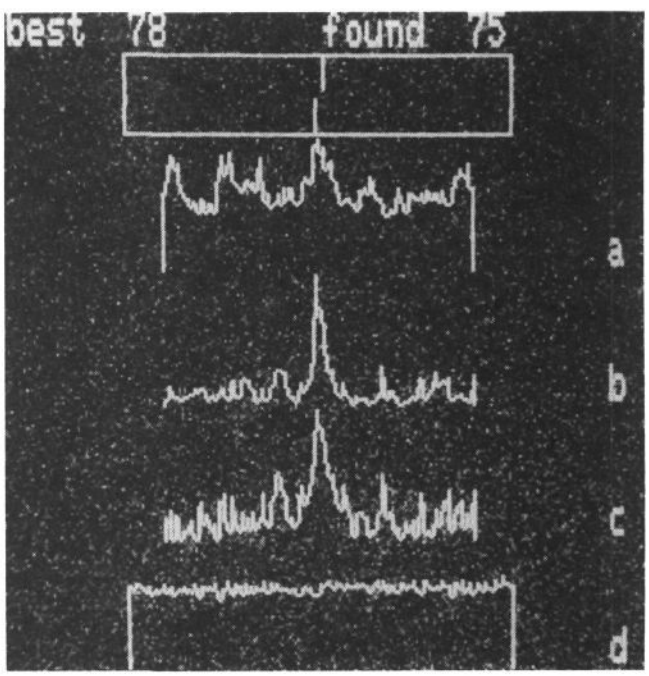


Figure 2. The response of the three dipole operators to a step edge in gaussian noise. The edge amplitude is 1 grey level and the noise standard deviation is 7 grey levels. Trace a is the entropy response, trace b is the SD, and trace c the significance of means. Trace d is a density profile along the window integrated across its width.

Standard Deviation

The entropy measure responds to the shape of the distribution but is costly to calculate. A cheap alternative, which also responds to the shape of the distribution and which is also self normalising, is the standard deviation. In similar vein to the entropy, the measure is $2 \times SD_t - (SD_A + SD_B)$ and has a similar response.

Mean

One way of looking at our approach is to say that we are examining two distributions to determine whether they appear significantly different. A straightforward method of doing this is to examine the significance of the difference in means given by

$$z = | \mu_A - \mu_B | / (\sigma_A^2 - \sigma_B^2)^{\frac{1}{2}} .$$

Figure 2 shows the responses of each of these three operators in locating a step edge with superimposed noise in a case where the image signal to noise ratio is low (0.14). All of these operators show the capability of locating step edges in noise.

PERFORMANCE

To determine which of these operators has the best properties of sensitivity and accuracy, we have undertaken a systematic test of their responses to a step edge in noise, varying the step size, noise standard deviation and dipole width and length. As a benchmark by which the responses could be measured we included the response of the Canny edge detector in the test.

The Canny Operator

The edge operator due to Canny[3] is widely regarded as the best compromise between sensitivity and accuracy in the detection of edge primitives. Indeed it was designed to provide the optimum response to a step edge amongst gaussian noise. It consists in essence of a one dimensional gaussian smoothing of the raw image in the direction parallel to the edge, followed by a one dimensional derivative of gaussian convolution across the edge. The widths of the gaussians in the two directions are typically equal, and the edge response is integrated across some sampling width. Canny's implementation provides for detection of edges at different scales and combination of the responses at different scales to produce an edge map. Our requirement is not for an edge primitive detector, but an edge locator. The different scales at which the Canny operator can be applied correspond roughly to the varying dipole size of our detectors. We do not need to track the response to the edge through scale space, we are merely interested in the sensitivity and localisation accuracy at a particular scale.

The Test

The images used consisted of 256 x 256 pixels with a single vertical step edge extending the height of the image, on which had been superimposed gaussian random noise. The signal to noise ratio (edge amplitude divided by the noise standard deviation) was varied from 0.14 to 1.33. The dipole widths and (half) lengths were varied from 10 to 50 pixels in steps of 10. In the case of the Canny operator gaussian smoothing of standard deviation 1, 3, 5, 7 and 9 pixels was applied.

For each point in parameter space 10 measurements of edge position were made at completely separate positions along the edge. (Thus for some of the broader supports more than one noise image was required.) The measurement consisted of scanning the dipole across the whole width of the image and measuring the responses of the dipole and Canny operators. In the case of the Canny operator the output of the smoothed one dimensional derivative was integrated along the edge direction. In all cases, the detected edge position was taken to be the position of absolute maximum response.

The measurements made on each scan were:

The distance of the located edge from the true edge.
The response at the position of the true edge.
The response at positions distant from the true edge.

From these measures at each point in parameter space we have:

A (sparse) histogram of localisation error.
A distribution of response signal.
A distribution of response noise.

From which we derive :

Sensitivity	:	(signal mean – noise mean) / (noise standard deviation)
Localisation accuracy	:	The mean localisation error
Localisation precision	:	The standard deviation of the localisation error.

Results

Derived performance values were obtained over a range of dipole widths and half-lengths. In each case the Canny smoothing standard deviation used as an equivalent to the detector length is such that two s.d.'s is about equal to the dipole half-length. The two cases are not directly comparable in terms of the contributions of their support regions, and the decision to adopt a particular combination of support sizes as being equivalent is a fairly subjective one. The two standard deviation cut off was selected, since the gaussian weighting is certainly significant within this boundary, and indeed for some distance beyond it. The Canny results are included to give some idea of the scale of values.

Figures 3 to 5 show examples of some results at a particular scale. They show how the derived values

vary with the signal to noise ratio of the image using a dipole width of 30 pixels and a half length of 20. An s.d. of 9 was used for the corresponding Canny operator.

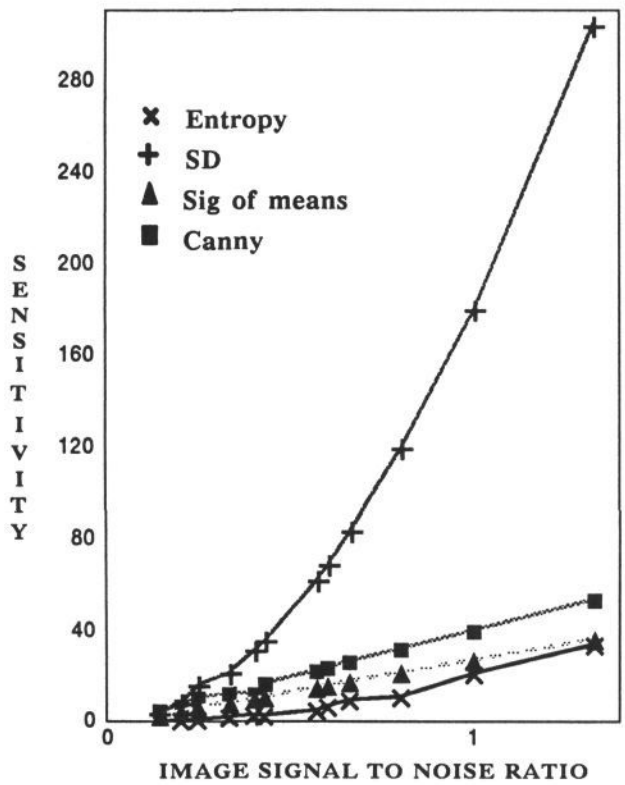


Figure 3. Sensitivity of the dipole operators compared with the Canny operator as a function of the image signal to noise ratio (edge amplitude divided by noise s.d.). The dipole width and edge integration width is 30, the dipole half length is 20 and the s.d. of the smoothing gaussian in the Canny case is 9.

Figure 3 shows the variation in sensitivity of the different operators. Not surprisingly, the sensitivity of all the operators increases steadily with the signal to noise ratio of the image. Against the Canny benchmark, the performance of the entropy dipole is poor, particularly at low signal to noise values. The significance of means dipole is better at low signal to noise, having about 60% of the Canny response. The SD dipole has very similar sensitivity to Canny at very low signal to noise ratios, becoming increasingly better as the signal to noise ratio increases above 0.25.

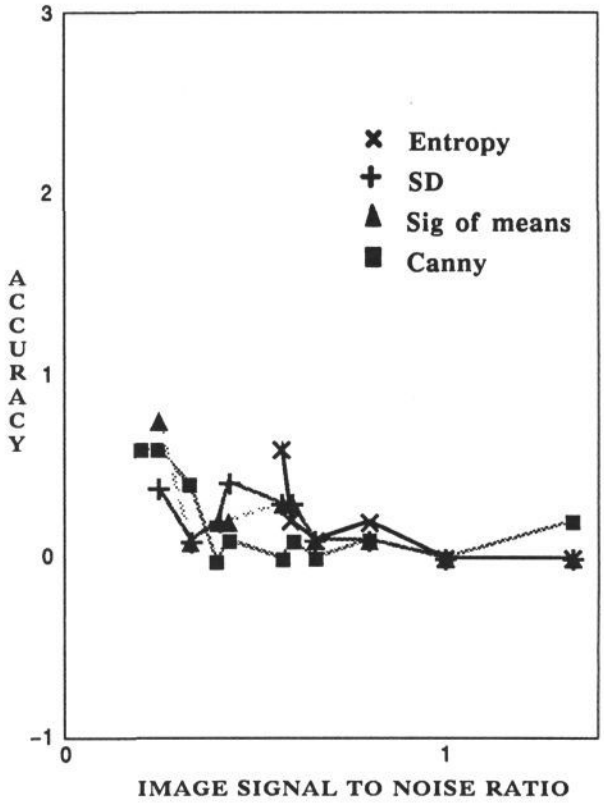


Figure 4. Edge localisation accuracy (mean error) for the 30 x 20 (9) support. (See figure 3).Off scale values at low image signal to noise are not shown.

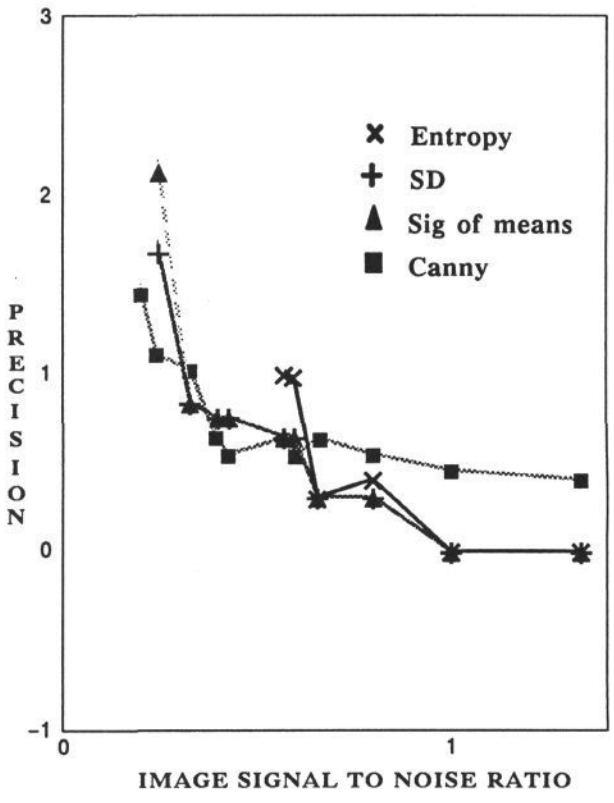


Figure 5. Edge localisation precision (standard deviation of the error) for the 30 x 20 (9) support. (See figure 3).Off scale values at low image signal to noise are not shown.

Figures 4 and 5 show the variation of localisation accuracy and precision using the 20 x 30 pixel support. Both accuracy and precision give a measure of how reliably the edge is located, and the graphs show similar behaviour. Both give good values (< 1 pixel), down to some signal to noise threshold at which the localisation becomes quickly unreliable. In the case of the entropy dipole, the threshold is rather higher than for Canny. SD and significance of means dipoles have slightly higher thresholds than Canny. Notice that this threshold can occur at values of sensitivity which appear fairly high.

With smaller support sizes, similar behaviour is observed. All the sensitivities are reduced, of course, and the threshold at which localisation accuracy becomes unreliable is higher. The sensitivity of the SD dipole at a signal to noise ratio in the image of unity, using a 10 x 10 pixel support, is about twice that of the Canny operator, compared to about five times as in figure 3.

Real Images

Experiments with test images provide confidence that the dipole operators are likely to be reasonable candidates for providing boundary cues. The model of a step edge among random noise, however, is not an ideal one for the cases in which we would like to apply these operators, namely to diffuse or weak edges among structured noise. Experiments with a number of images, particularly of biological material, indicate that one or other of the dipole operators can give a strong response at faint or noisy boundaries where the Canny operator responds only weakly or not at all. Obviously cases can be found in which these operators fail to detect a boundary, but in such cases the Canny operator also fails. There is no clearly best candidate among the three dipole operators. The significance of means response is consistently similar to that of the SD operator, and consistently more noisy, making it clearly the worst. Whether the best results are obtained by the entropy or SD operator depends on the image in question, notwithstanding the poor showing of the former on the noise images. The difficulty in modeling real cases means that it is difficult to make an objective assessment of performance or to demonstrate power in boundary location. Further study may allow us to find methods of determining the most appropriate operator for particular cases. For illustrative purposes we present some examples of edge responses.

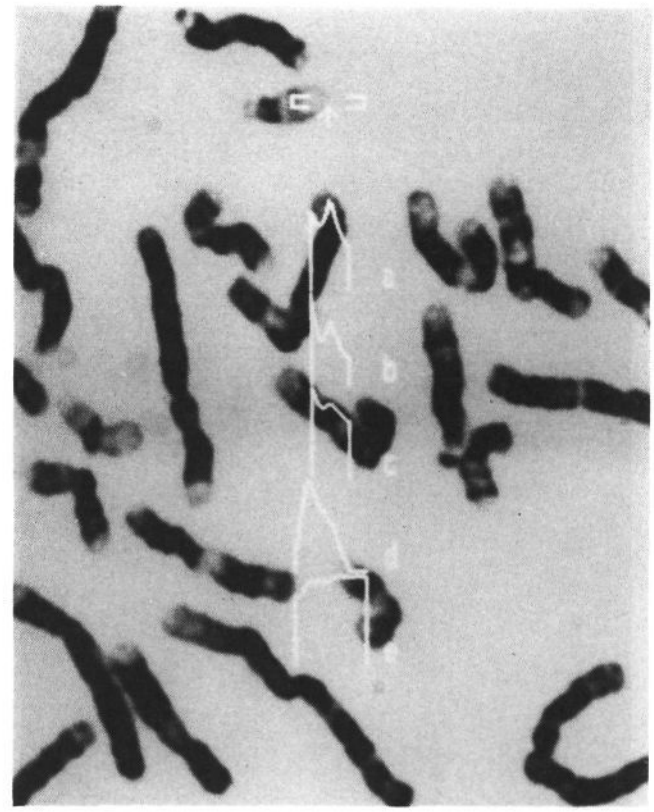


Figure 6. Detection of a weak edge in a chromosome image. The search region is indicated by the bracketed window.
a – Entropy dipole response
b – SD dipole response
c – Significance of means dipole response
d – Canny response
e – Projected density profile along the search line
The arrow indicates the edge position determined from a. The dipole width is 7 pixels and half length is 10 pixels. The corresponding standard deviation for the Canny operator is 5.

Figure 6 shows a search for a difficult edge in a chromosome image. A shape model predicts the existence of a boundary in a certain direction. The search is confused by the existence of strong edges in addition to the weak true edge. The responses of all four operators under test are shown. The dipole operators, including the entropy dipole give significant responses while the Canny operator produces no response. Figure 7 is part of a radiograph of a hip prosthesis. The required boundary is that between the bone and the retaining cement. The edge in this case can be very indistinct, but dipole search with a large support can provide important cues to its position. No case has been observed in which a boundary which can be detected by the Canny operator cannot be detected by one or more of the dipole operators.

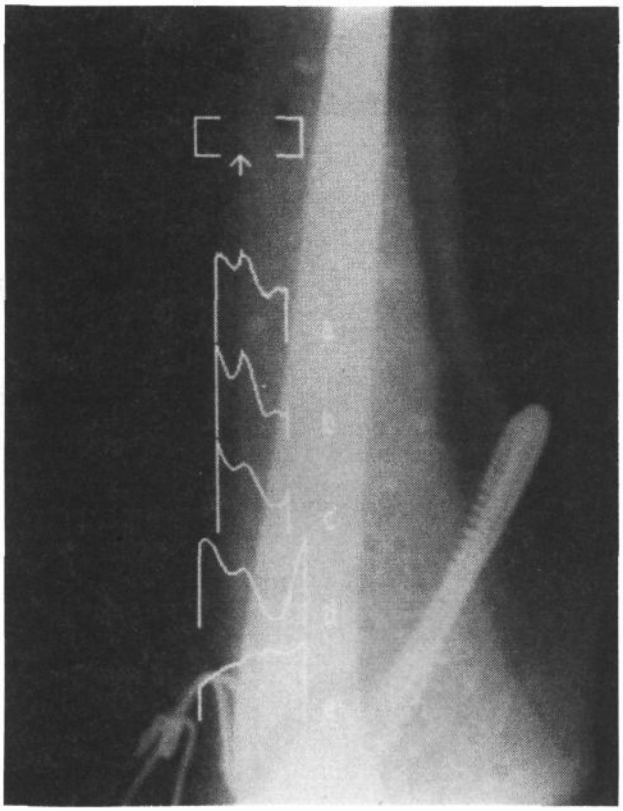


Figure 7. Part of a radiograph of a hip prosthesis. The required boundary is that between the bone and the cement holding the prosthesis in place. Indicated responses etc. as for fig.6.

It is important to notice that the image signal to noise ratio at the edge is quite high in both of these cases : about 2.3 for the chromosome image and about 1.9 for the radiograph. Even if we choose to model the boundary detection using a step edge amongst noise, our working region in real images is likely to be well to the right of, or beyond, the scale of figures 3, 4 and 5.

DISCUSSION

The approach taken in designing the dipole operators described here is that something is known about the location and orientation of a boundary between two regions and that its true position can be determined by measuring some statistic of the distribution of image values on either side of the edge. De Sousa[4] has described a similar application of sliding statistical tests to radiographs and natural texture images. One of his measures was identical to the significance of means dipole described here, which has consistently shown behaviour similar to that of the SD dipole, but more noisy. Several authors [5, 6] have considered using a comparison of medians or other order statistics to detect edges. These methods are all used in the context of edge preserving smoothing to reduce impulse noise. We did not consider this to be an appropriate model for the type of boundary detection we wish to achieve.

Three statistics whose properties seem reasonable for the task have been implemented and tested systematically on an artificial image, with the Canny operator acting as a benchmark. This test of sensitivity and accuracy was an exacting one since the Canny operator is optimised in one sense for the detection of step edges among gaussian noise. The performance of the entropy dipole was disappointing, but that of the significance of means dipole was better. The SD dipole gave encouraging results, being about as accurate as the Canny operator and in many cases much more sensitive.

The application of the dipole detectors to difficult real-world images has shown that one or more of them can provide useful boundary cues in cases where linear edge detection fails – the very cases in which model based instantiation is most necessary. Despite its poor showing in detecting model step edges, the entropy dipole appears to retain some promise as a boundary cue operator in real-world images.

REFERENCES

1. **Woods, P.W., Taylor, C.J., Cooper D.H. and Dixon, R.N.** "The use of geometric and grey level models for industrial inspection." *Patt. Recog. Lett.* Vol 5 (1987) pp11 – 17.

2. **Azzopardi, P.J., Pycock, D., Taylor, C.J. and Wareham, A.C.** "An experiment in model based boundary detection." *These Proceedings.*

3. **Canny, J.F.** "Finding lines and edges in images." *MIT Artificial Intelligence Lab, Cambridge, MA, Technical report AI-TR-720* (1983).

4. **de Sousa, P.** "Edge detection using sliding statistical tests" *Computer Vision Graphics and Image Processing* Vol 23 (1983) pp 1–14.

5. **Bovik A.C. and Munson D.C.** "Edge detection using median comparisons" *Computer Vision Graphics and Image Processing* Vol 33 (1986) pp 377–389.

6. **Pitas I. and Venetsanopoulos A.N.** "Nonlinear order statistic filters for image filtering and edge detection" *Signal Processing* Vol 10 (1986) pp 395 – 413.

5. **DEMOB: an object oriented application generator for image processing.**
N. Bryson, D. Cooper, J. Graham, D. Pycock, C.J. Taylor and P.W. Woods,
Proceedings of the fourth Alvey Vision Conference, Manchester, 1988, pp 37 - 44.
doi:10.5244/C.2.7

DEMOB - An Object Oriented Application Generator for Image Processing

N. Bryson, D.H.Cooper, J.G.Graham, D.P.Pycock, C.J.Taylor, P.W.Woods

Wolfson Image Analysis Unit, Department of Medical Biophysics
The Victoria University of Manchester, Manchester M13 9PT

This paper describes an Object Oriented program generator for image processing applications. Control is represented by a dataflow graph and interaction by "view" objects which update displays and modify domain objects. Progress so far also indicates that Object Oriented Programming for user-defined image processing requires a rich programming support environment.

Many groups have developed libraries of image processing (IP) modules for applications in the medical and industrial fields. The use of these libraries requires considerable programming effort and expert knowledge, which limits the economic viability of the technology. To develop an application, the user must not only have expert knowledge of image processing algorithms, but must also contend with the messy details of programming. Because of the commercial pressure to quickly demonstrate the feasibility of program designs, there is a need for an application generator. This paper describes our attempts to develop such a tool (DEMOB). The primary aim of this tool is not to help with the "art" of image processing, but to help make the construction of an application as direct as possible.

Objectives of DEMOB

One of the objectives in designing DEMOB was to investigate the problems involved in designing an application generator. Prototyping tools implemented using conventional structured programming suffer from having the structure of domain objects, interactions, and control programmed in. For example the implicit control structure may be a pipeline, along which a work image flows, with the user being offered (via menus) a choice of the operations to be performed on the image. Interactions occur in a pre-programmed manner, and tedious re-coding is required when new operations are added.

Our design goals included the requirement that control be separated out explicitly from the operation of the tool and from the domain objects. Interactions should also be factored out, so that interactive tools for different domains could be rapidly built from a set of building blocks. The program generator should be open, so that further domain objects and interactions could be added easily. The user should not be constrained to follow a particular design cycle, but instead should be able to interact with subcomponents of the problem in a relatively unstructured way. The Object Oriented Programming (OOP) technique[1], with its desirable properties of information hiding and run-time binding seemed to offer a viable approach to implementing such a tool. One of the aims of the project was to explore the uses of OOP for IP. Because of the nature of IP, with the need to display and interact with raw and processed images, the tool should have a mouse - driven, window based graphical interface. A brief description of the OOP paradigm is given below. For a fuller description, the reader is referred to reference 2.

APPLICATION REPRESENTATION

Separation of Control

A prototype application is represented internally by a dataflow graph[3]. This representation was chosen because the explicit recording of dependencies between data items provides the potential for automated reasoning about applications. A simple graph, representing "blob" extraction from an image, is shown in figure 1. The dataflow graph consists of: nodes, representing actions applied to data objects; tokens, representing references to data objects; and directed arcs, which transfer references to tokens between nodes.

A node is ready to fire when it has received all its tokens from its input arcs. New tokens are created by the node, and are transferred to further nodes via the

This work is supported by an SERC grant and is part of ALVEY project MMI-093: "Techniques for User Programmable Image Processing (TUPIP)"

AVC 1988 doi:10.5244/C.2.7

output arcs, making them ready to fire. After execution only the "dangling" output arcs of the graph contain references to tokens, all intermediate data objects (and their tokens) having been consumed. Each token contains a counter for references to the data object. When references to tokens are created or consumed, the counter is incremented or decremented, respectively. When the counter reaches zero, the data object, and token, can be deleted.

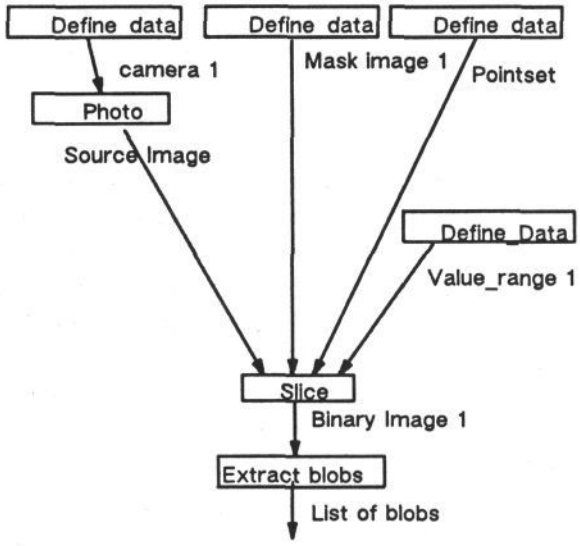


Figure 1. Graph representing blob extraction.

The User Model

DEMOB is designed to present the application and data objects graphically to the user. The user may select any action or data object in his application by selecting the appropriate node or arc of the graph, which will respond by offering an interacton with the appropriate object. Each interaction takes place within a window, which contains a menu, a display of the object, and a prompt line. Selection within the window leads to a more specialised interaction with the object, or an interaction with a new object. Initially the user is offered a general interaction with an empty list object. Typically the user then chooses to load a graph from a disc file, or to create a new example. An interaction with the new graph is entered and the user is presented with a diagram such as that shown in figure 1, and a menu of available interactions. For example, if the user selects a dangling output arc he will enter an interaction with the data object on the arc. The arc itself can be selected, and modified so that, during execution, it will display the data objects passing through it.

During execution the arcs and nodes are highlighted as they become active, and windows open on the screen to show data objects as they are created. The user modifies his program by cutting arcs, creating new nodes, and adding them to the graph. Graphs and objects can also be saved and retrieved from disc file.

IMPLEMENTATION

The Object Oriented Programming Environment

The key features of OOP are data encapsulation and inheritance. The OOP technique consists of identifying objects which are to be manipulated in an application, and in defining the data structures and operations needed for each object. The private data structure representing the object is protected from direct manipulation by the user, and operations on it may only be carried out by sending a message to the object, which uses it to select an appropriate operation. It is important to note that the message only conveys "what" the programmer wants done, but the object itself decides "how" it is done. For example the message AREA sent to a shape object may cause different operations to be carried out, depending on whether the shape is a circle, rectangle, etc. Because the internal representation is hidden, it can be changed without affecting the user, who need only know the messages to which the object responds.

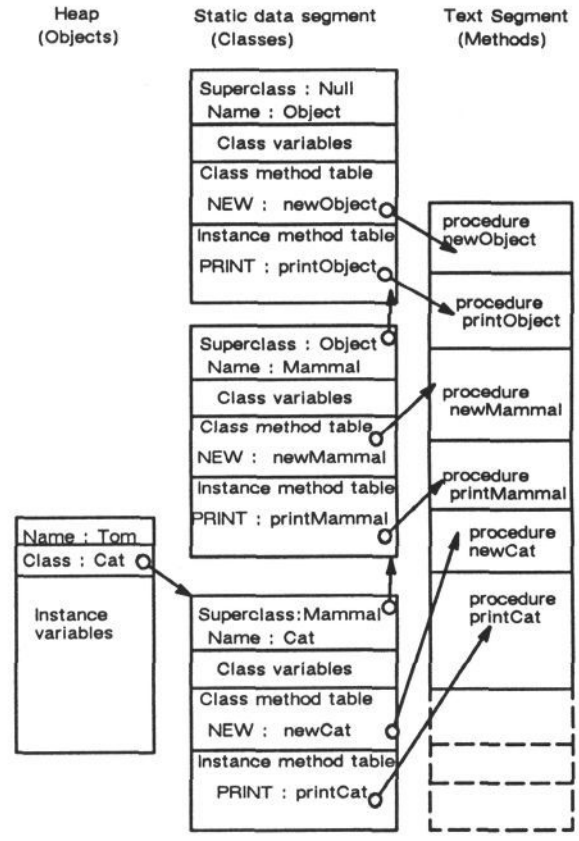


Figure 2. Schematic memory map of OOP system

An individual object is regarded as an instance of a particular class e.g. Tom is an instance of the class

Cat. The class definition defines the data structure ("instance variables") of its instances and the messages (with corresponding operations or "methods") to which it will respond, in the form of a lookup table, indexed by the message selectors. Each object contains a pointer to its class, which is itself an object, with corresponding class methods and "class variables". In particular, each class can respond to the message NEW by returning a new instance (by invoking the newCat method). Figure 2 shows a schematic memory map showing the relationship between objects, classes and methods.

The use of encapsulation helps to protect code from the effects of changes. The use of inheritance allows code to be re-used. The classes are arranged in an inheritance hierarchy, with more general operations and instance variables being defined in classes high in the hierarchy. For simplicity, each class can inherit from one "superclass" only, although in other implementations, multiple inheritance is a useful feature. Each class contains a pointer to its superclass (except for Object, the class at the root of the hierarchy). Methods high in the hierarchy may only manipulate instance variables defined at that level or higher in the hierarchy.

We have developed an OOP environment in-house in which to implement DEMOB. This consists of C with some preprocessor tools. The programmer causes messages to be sent to objects by inserting code of the form:-

```
Tom = NEW$(Cat);
PRINT$(Tom);
```

where Tom is of type "ob_ptr" i.e. a pointer to an object. A special global object pointer, "self", represents the receiver of the current message, and is used within methods to access the instance variables of the object. A preprocessor tool checks that the message selector is represented in a file of valid messages, and then converts this string into a call to a message handling routine:

```
Tom = msg(NEW,Cat);
msg(PRINT,Tom);
```

The message handling routine expects the message selector and the receiver of the message as the first two parameters. Upon execution, the message handling routine searches the method lookup table of the class of the object for the corresponding method (it recognises a class by the fact that its class pointer instance variable is null). The stack is adjusted to simulate a standard C procedure call, and control is passed to the method.

The programmer may also specify that the search for a method should begin in the superclass, rather than in the class, by writing:-

```
PRINT$super(Tom);
```

This powerful facility allows the chaining together of methods. For example, the PRINT message causes the printing of the values of all instance variables of the object, and is implemented at each level of the class hierarchy by methods of the form:-

```
PRINT$super(self);
printf("Instance variables at this level");
```

This causes the instance variables, starting at the root of the hierarchy, to be printed.

The low level image processing routines of our system are microcoded to make use of particular data structures, optimised for IP operations and executed on a slave co-processor. Since a large effort had already been expended in developing existing software and hardware,we were constrained to include these existing data structures in DEMOB. One of the objectives in developing DEMOB was to investigate the effective use of mixed typed data structures and objects.

Image Processing Objects

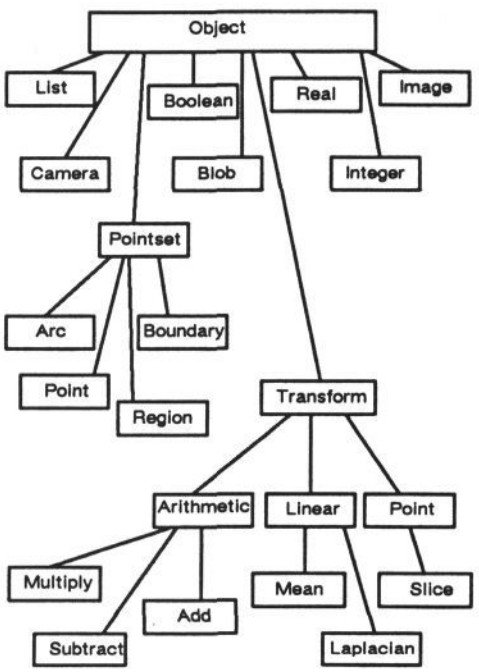


Figure 3. Image Processing Objects Class Hierarchy

Since we wish to be able to add new classes and modify existing classes without too much disturbance to the rest of the system, DEMOB makes no assumptions about the domain classes, other than that they will respond to a limited set of messages, which provide the interface between these classes and DEMOB. These messages allow DEMOB to ascertain the default method of interaction and the operations available for the domain object. Besides general purpose objects such as integers, reals, booleans, lists, etc., a variety of data objects specific to IP are also required. These include images, cameras, pointsets, value ranges, and image transform specifiers[4]. A class hierarchy of these

objects is shown in figure 3.
Associated with each object is a set of messages which are used to process the data e.g. a Camera object responds to the PHOTO message by producing an Image object. The processing steps in the application are built up using these messages, which are inserted into the graph as described below. Also associated with each object is a set of interactions e.g. the Real number object has a set of interactions allowing the number to be modified in a variety of ways. These interactions can be inserted into the graph as described below.

Graph Nodes

The structure of the dataflow graph allows control over the order in which operations are carried out. Further control is provided by supplying specialised nodes which implement further control structures. The class hierarchy of the objects used in constructing graphs is shown in figure 4. The Node class provides the functionality common to all nodes e.g. the ability to receive and transmit tokens. Further, more specialised functionality is provided by the subclasses of Node.

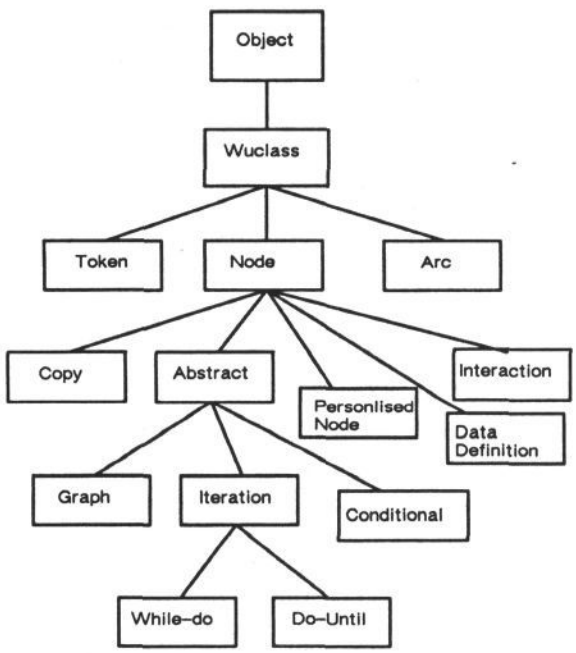


Figure 4 . Inheritance hierarchy of graph objects

One such class is the "Data Definition" class, whose instances act as sources of data for the dataflow graph. When the node is created the user is asked to define the data object, and on subsequent execution a token representing the data object is output by the source node. In the example graph shown in figure 1, four data objects are needed : a camera, a mask image, a pointset (defining the region over which the operation is done) and a value range (which defines the thresholds for the slice operation).

Multiple references to data objects are frequently required, so a "copy" node class is provided. These work by incrementing the reference counter of the input token, and passing out multiple references to it via the output arcs.

A "graph" node, like any other, can receive tokens from its input arcs, and passes tokens to its output arcs. Internally its execution is represented by a collection of nodes and arcs i.e. a sub-graph. A graph can thus have a layered structure, with a "root" graph node controlling the execution of its sub-graph, which itself can contain graph nodes.

Iterative execution is implemented by providing a variety of "iteration" node classes. For example, a specialised type of graph node, the "while-do" node controls the execution of a condition graph and an action graph. During execution the set of input tokens is passed to the condition graph, which executes and returns a boolean object. If it is "true" the set of input tokens is passed to the action graph, which returns a new set of tokens, and the cycle is repeated. If the boolean is "false", the tokens are passed out as output.

Conditional execution is implemented by providing an "if-then-else" node class. This node controls a condition graph, a "true" graph and a "false" graph. The execution cycle is similar to that of the "while-do" graph i.e. the input tokens are passed to the condition graph, and a boolean object is returned. Depending on its value, the input tokens are passed to the "true" or "false" graph.

These nodes introduced so far provide the control structures of conventional structured programming (a subroutining facility could easily be added). A further class of node, the "personalised" node, is included to allow user selected actions on data objects. Actions are carried out by sending a message to an object e.g. to add two numbers together the message is :-

```
C = ADD$(A,B);
```

where A, B,C are references to the number objects, ADD is the message, and A is the receiver of the message. The personalised node has instance variables allowing it to store this message. The convention adopted is that the message is sent to the data object input via the first input arc, with the data objects from the other input arcs as parameters. The returned object from the message is assumed to be a list of output data objects, which are transferred to the output arcs. In this case a single object, C, is returned. During the creation of a personalised node, the user is prompted to define a receiver, which then offers for selection all the available messages for that class of object.

Part of the user's application may involve run-time interaction with the data objects, and a special "interaction node" class is provided to allow this. When creating this node, the user will be prompted to define the class of input, and will be asked to select from all available interactions with this class. An instance of this interaction (see below) will be created and placed in the interaction node. When the data object arrives along the input arc, the interaction will take place, and the modified data object will be output along the output arc.

The Interaction Model

The programming of interaction required a large part of the effort in developing DEMOB. A graphics sub-system was developed, which implements a GKS[5] model of graphics i.e. overlapping rectangular regions of the display device, called "viewports", display "icons" in "worlds" through "windows" opening on the worlds. Viewports are grouped together in frames. These components were implemented by defining classes of objects, with all device-dependent code being confined to the Screen class.

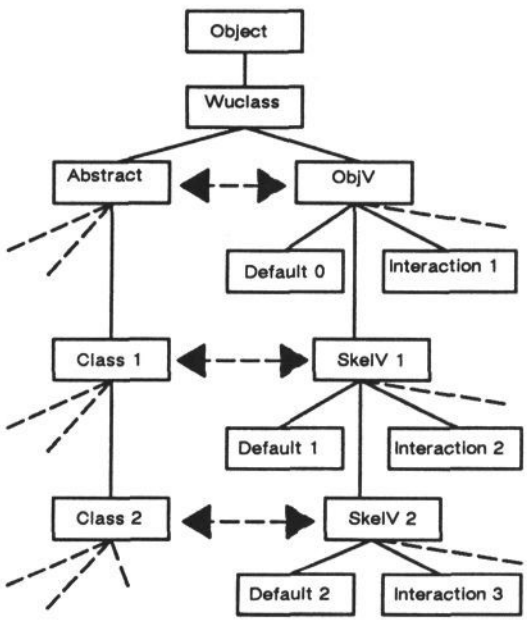


Figure 5. The View Class Hierarchy

Rather than have each domain object controlling its interactions, we decided that each interaction should be defined by a separate class, to increase modularity and to save memory space. When an interaction with an object is required, an instance of the relevant class, known as a "view", is created and is put in control of the object. The view object would inherit the instance variables and most of the methods needed to represent and manage the display from its superclasses, and need only provide locally the code needed to determine the course of this particular interaction.

We thus define a class hierarchy of views to accompany the hierarchy of domain objects, as shown in figure 5 . This view hierarchy contains "skeleton" classes which provide all the instance variables and methods common to interactions, and "view" classes which define particular interactions. Each object has a default view class, which implements the most general interaction possible i.e. to display the object and to offer all interactions available for it. It also allows the user to select sub-components of the object for further interaction. When a new domain class is added a corresponding skeleton class with default view and any further view classes are also added.

RESULTS

A basic version of DEMOB has been implemented on a CVAS 3000 (Visual Machines Ltd.) system. Images of the screen, showing a sample graph during and after execution, are shown in figures 6 and 7. These show a graph interaction window, with associated menu, and displays of a source image and a sliced image.

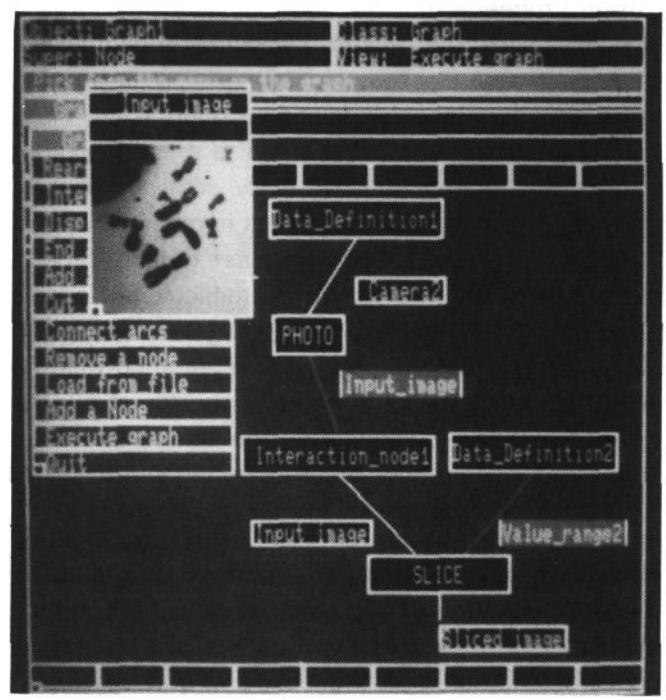


Figure 6. Sample graph during execution

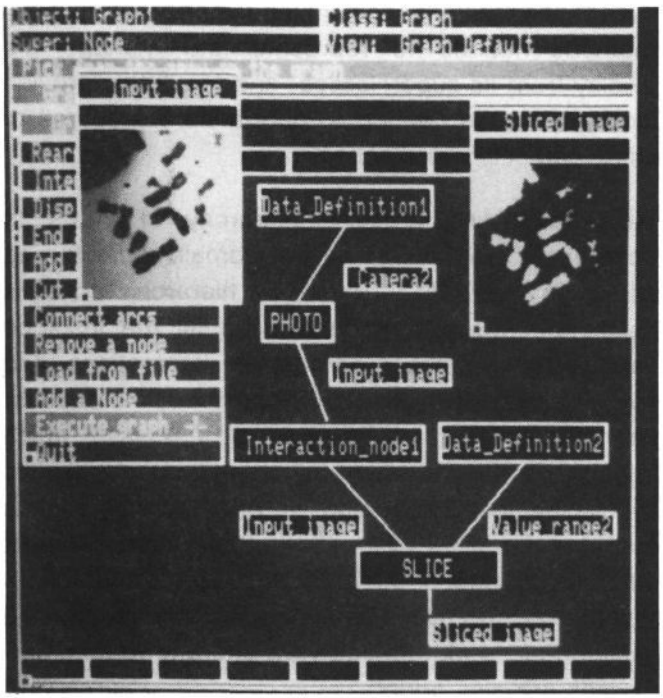


Figure 7 – Sample graph after execution

DISCUSSION

Limitations of the OOP Environment

Designing with objects proceeds by identifying the instance variables and methods of classes. As the design proceeds, the programmer identifies ways in which code can be reused, usually by spotting ways of adding new classes and by splitting or moving methods and instance variables around in the class hierarchy. The final system contains a large number of classes, each with a relatively small number of methods, each containing a small number of lines of code. It is easy for the programmer to lose track of such a system, and it can be difficult for a new programmer to add new functionality to the system, since in order to do so, he must understand what he is inheriting. We thus need tools to allow the programmer to browse around the source code in a structured manner such as are supplied in Smalltalk[6].

In our OOP environment, all class and method definitions are compiled. While this increases run-time efficiency, it also means that the definition of an object is fixed. Further, although the code is compiled, no static type-checking is done.

Design Issues Raised in OOP

Initially, design concentrated on using the "information hiding" aspect of OOP, but as we gained experience, inheritance gained in importance. The derivation of a class hierarchy can be difficult, in part because the only relationships provided are the "is-a" and "is-an-instance-of" relations. These do not, however, describe the relationship between say, views and their objects. For these objects the "information hiding" aspect of OOP is a disadvantage, since the views need to have access to the instance variables of their objects. This problem arises because it is difficult to decompose IP applications into independent static classes of objects i.e. where information can be encapsulated permanently into objects. In practice we wish to be able to merge the information from several classes of object e.g. in applying an operation over a pointset in an image. Where a static encapsulation of information suffices(e.g. the dataflow graph) the model works well. The frame paradigm [7] presents a richer environment for the representation of domain knowledge,allowing object decomposition as well as more complex relationships between objects.

Our experience has shown that the use of mixed typed and object variables reduces the advantages of using OOP, since much code had to be handcrafted to deal with the typed variables. Of course, these are needed to map onto the IP software and hardware. The correct way to deal with them is to implement a basic set as objects, with methods handcoded to implement the functionality provided automatically for other objects. It is instructive to compare our approach with the Eiffel OOP language [8] . This language has a compiler which recognises the use of a limited set of simple types, but any complex data structure must be represented as an object.

Design Issues in the User Interface

The user interface is crucial to the success of a tool such as DEMOB. Users are sensitive to features which appear relatively trivial, such as the particular style of interactions, or the wording of prompts. There are a variety of ways in which an interaction, such as modifying an integer, can be implemented. One solution is to recognise that these interactions have much in common, and thus could be represented by an appropriate view class hierarchy.

Instead of being presented with the dataflow graph, some users felt that it would be better for this to be hidden, and instead to present only the data objects. Design would proceed by creating and selecting objects for processing. The user would then selects an appropriate operation which could take these objects as inputs. The selected action would be invoked, causing new objects to be created.

CONCLUSIONS

The dataflow graph is adequate for representing an application. However the need to fully specify the dataflow is tedious for programmers, since the

normal constructs of conventional programming languages correspond to a lot of dataflow, which the programmer does not normally need to specify explicitly.

While OOP greatly increases the robustness and reuseability of code, and provides a useful paradigm for the development of complex software systems, the behaviour and relationships between the complex data structures used in IP are not adequately modelled by message passing objects arranged in a single inheritance class hierarchy. To use OOP the development environment should be fully integrated with the language to allow rapid modification of an evolving system. A clean interface between typeless data structures (objects) and typed data structures should be maintained. The dataflow graph provides a viable means of representing procedural knowledge, but needs careful design of the user interface in order to gain user acceptability.

REFERENCES

1. *ACM SIGPLAN Notices*, Vol. 21, No. 10, Oct. 1986
2. **Brad C. Cox,** "*Object Oriented Programming – An Evolutionary Approach*", Addison – Wesley, 1986
3. *IEEE Computer*, Special Issue on Dataflow Systems, Vol. 15, No. 2, Feb. 1982
4. **Graham J., Taylor C.J., Dixon R.N.,** "A compact set of image processing primitives and their role in a successful application program", *Patt. Recog. Lett.*, 4, 325 – 333, 1987
5. **Hopgood F.R.A., Duce D.A., Gallop J.R., Sutcliffe D.C.,** "*Introduction to the Graphics Kernel System (GKS)*" A.P.I.C. Studies in Data Processing No. 19, Academic Press, 1983
6. **Goldberg A., Robson D.,** "*Smalltalk – 80 The Language and its Implementation*", Addison – Wesley Series in Computer Science, 1983
7. **Wood P.W., Pycock D.P., Taylor C.J,** "A Frame–based System for Modelling and Executing Visual Tasks", *Alvey Conference 1988.*
8. **Meyer B.,** "Eiffel: Programming for Reusability and Extendability", *SIGPLAN Notices* Vol. 22., No. 2, Feb. 1987

6. **User Programmable Visual Inspection.** J. J. Hunter, J. Graham and C. J. Taylor, *Image and Vision Computing, 13: 623-628, 1995.* doi:10.1016/0262-8856(95)97287-V

User programmable visual inspection

J Jeffrey Hunter, Jim Graham and Chris J Taylor

The usefulness of advanced machine vision techniques for automated inspection is restricted by long development cycles for the inspection software and a lack of general applicability of the resulting system. These difficulties arise because in the majority of cases the particulars of the inspection are embedded in the software structure resulting in severe restrictions on user reconfiguration. This paper describes the structure and major components of an inspection system which is capable of tackling a wide range of inspection problems without reprogramming. User configurability is based around a simple, interpreted geometric description language which provides support for both dimensional and surface quality measurements. The operation of the system is illustrated using the inspection of automotive brake assemblies as an example.

Keywords: visual inspection, machine vision, user configurability

There is considerable potential for the use of automated visual inspection in manufacturing industry. Some inspection problems require specifically tailored algorithms to achieve satisfactory results[1], but there is a large class of problems which can be simply stated in terms of object location, dimensional measurement and surface quality assessment. Members of this class can be loosely termed as 'generic inspection problems'.

In the past, practical inspection systems have typically relied on embedding knowledge of the target inspection task in the algorithms[2]. This approach tends to limit the reusability of the system, and certainly ensures that there is no possibility of the system or its subparts being reused by anybody other than a vision expert. This means that new inspection systems are expensive to develop often precluding on cost grounds applications where, otherwise, visual inspection would be an ideal solution. Most inspection system developers appreciate these difficulties and attempt to surmount them by creating inspection libraries. Libraries go some way to providing for simpler development, but they are generally low-level entities still only of value to the vision expert. Alternatively, a number of visual programming systems exist which allow the user to specify image processing tasks by using an interactive tool. These systems go some way towards removing the necessity for programming expertise, but they are generally limited in image processing capability and still require the operation to be described in terms of the image processing rather than the inspection task. What is required is a system in which the user's description of the task is separated from the image analysis and which provides a language whose syntax and semantics are appropriate to the problem domain. We have achieved this and constructed an inspection system which is flexible and user configurable giving the user the ability to tackle a wide range of generic inspection problems.

We describe the system's components and operation and show how a user-specified task can launch complex image analysis operations. We demonstrate how our system allows generic inspection tasks to be rapidly prototyped by a domain expert and can be used to perform new inspections tasks with minimal effort.

To assist the reader in understanding the material presented we have illustrated the steps in solving a typical inspection problem using our system. *Figure 1* shows part of a brake assembly which is the object to be inspected. One of the inspections is a test for a departure from circularity of the lining on the shoe. To perform this inspection the user would initially provide the system with some example training images with the objects identified by landmark points. Then, using the task description language, he/she would describe the interpretation task. In this case language statements would be used to tell the system that a circle should be fitted to the lining boundary landmark points and that a calculation of the distance between these boundary points and the circle is to be made. Once the description is complete the user can execute the inspection by requesting the value of the distance.

Wolfson Image Analysis Unit, Department of Medical Biophysics, University of Manchester, Manchester MI3 9PT, UK (email: jeff@-wiau.mb.man.ac.uk)
Paper received: 28 July 1994; revised paper received: 1 November 1994

0262-8856/$09.50 © 1995 Elsevier Science B. V. All rights reserved

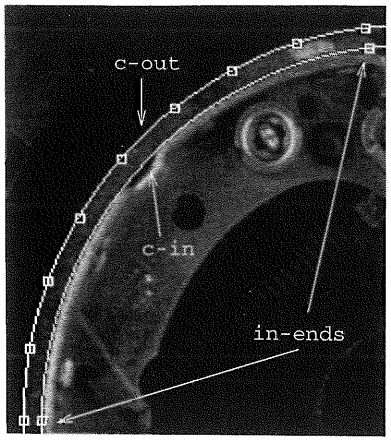


Figure 1 Display obtained from the system during inspection prototyping (with manually added annotation)

THE INSPECTION SYSTEM

Visual inspection consists of a number of stages from image acquisition through to process control. In this work we are interested in that part of the inspection process which is concerned with the calculation of measurements from images of the components to be inspected. In our design this subsystem comprises of three elements: object search, geometric interpretation and surface property evaluation. These components are discussed individually in the following subsections.

Object search

It is clear that to create a generic inspection system we require a generic object search technique. We have used the active shape model (ASM) technique described by Cootes *et al.*[3]. They compare this approach with a number of other methods for image interpretation based on flexible or deformable models and identify the strength of the ASMs[4] as arising from their specificity to the particular object for which they are searching. Specificity gives an ASM robustness in the face of noise and clutter in the image but from our point of view the general applicability of the technique is even more important. This generality arises from three key characteristics of the approach.

Firstly, the method relies on the identification of object landmark points. Landmark points are positions which can be unambiguously identified in all examples of the object. Most objects which are the subject of inspection had identifiable landmark points because these are the points which define the object's dimensions. The landmark points are necessary but not sufficient to determine the shape and are often augmented by a number of intermediate points, for example, boundary points on a straight edge between two corners.

Secondly, the method records statistics of the relationships between the positions of the landmark points in a set of training examples. The statistics which are recorded make no assumptions about the form of the object. Finally, and a major reason why the ASM technique has a fundamental role to play in the creation of a generic inspection system, the technique treats the spatial relationships between multiple objects in an assembly naturally. Many inspection tasks consist of both inter and intra object measurements and the ASM is able to deal with both in a uniform and configuration-independent way.

A part of every inspection task is determining whether the assembly has components missing. In this respect, a further useful capability of an ASM is its ability to provide a measure of confidence that it has located an instance of the target object in an image. These measures can be turned into boolean existence variables by the search module and made available on a component basis to the rest of the system.

Geometrical construction

There are two common types of measurement required in generic inspection: measurement of the dimensional properties of components and measurement of the material quality or finish. The dimensional properties are generally compared to design tolerances. The material quality is assessed over critical regions of the assembly.

Our inspection description is based around geometries because tolerances are generally expressed in terms of the dimensions of the geometries used in the design whilst areas of the assembly which require surface assessment can also be described by geometric constructions.

Figure 2a provides an idealized example to demonstrate the points to be made in the following sections. The figure shows an assembly which is a sheet with a hole cut from it. Consider the task of measuring the angle of the top left corner. When an inspection is performed on the object the image search technique is used to locate the landmark points on the sheet's rectangular boundary and the edge of the circular hole. A close up view of the object edge (*Figure 2b*) shows that it has undulations and that the object search has performed accurately. The angle measurement depends upon considering the two edges as straight lines as in the

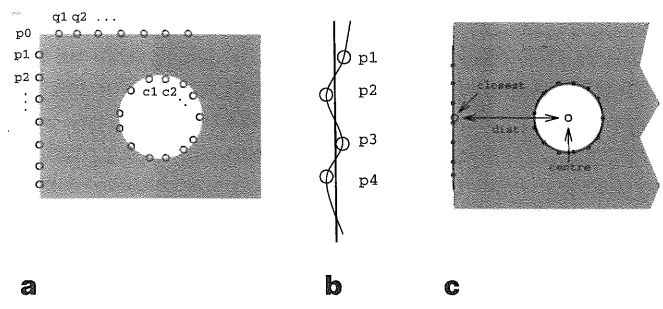


Figure 2 (a) Example shape for inspection; (b) close-up of the landmark points on the edge; (c) example measurement using the `closest` and `distance` operations

original design. The inspection description therefore reconstructs a straight line from the points which have been found.

In general, the geometric task description consists of describing the construction of geometries from other geometries using operations. In the example above, we have constructed lines `l1` and `l2` from the edge points:

```
l1 : line ([p1, p2, p3, p4])
l2 : line ([q1, q2, q3, q4])
```

The angle can then be calculated from the lines:

```
a : angle (l1, l2)
```

In the following sections, we discuss what types are required to perform a range of inspections and what operations should be available to act on each type. We also show how descriptions can be written to handle intentional variations in assemblies and how task descriptions are executed to obtain results.

Types

Only a small range of geometries and other types are required within the system to support a wide set of inspection tasks. The current system includes points, lines, circles, regions, booleans, reals and groups. It is accepted that this list may not be complete, but the important point is that only a few more geometric types are required to tackle the majority of cases. A previous study[5] of 17 industrial production-line inspection problems covering a wide range of products identified the requirements for only three extra geometric types: boundary (representing general polygonal outlines), ellipse and hyperbola.

Of the types in the system, groups, regions and booleans deserve individual explanation. Groups provide a way of collecting together sets of possibly mixed type values and treating them as a single entity. A region is a geometry defined by a group of boundary points and may represent an irregular connected area of the image. Boolean types are used in the logical control of *if then else* constructs in the description.

Operations

It is object search which locates the landmark points on objects in the image. Operations provide the mechanism for constructing new geometries based on these points. Operations are the basis for the user description of the whole inspection task.

We have already seen (*Figure 2b*) operations used to describe the construction of a line from a group of points and the subsequent calculation of the corner angle. The operations provided by the system are for the most part restricted to those which act on or produce geometric types; the number required to provide satisfactory functionality is limited. It is expected that extra functionality such as arithmetic capability will be provided by links to other more appropriate tools.

There are 17 user operations provided in the current system: `line, point, circle, region, [], if, closest, distance, derive, min, max, sd, mean, less, greater, between` and `greybox`. From this list of functions two of the operations which have been identified as providing real descriptive power are `closest` and `distance`. These are both overloaded functions and can take the following forms:

- `closest (<any geometry>,point)` returns the position on the geometry which is closest to the point.
- `closest (<any geometry>,[points])` returns a group of positions, one for each point in the group `[points]`.
- `closest (<any geometry>,<any geometry>)` returns the point on the first geometry which is closest to the second geometry

Similar definitions hold for the distance operation:

```
distance (point,point)
distance (point,[group])
distance ([group],[group])
```

As an example of the operation of distance and closest we will extend the example of *Figure 2* to measure the distance between the centre of the cut out hole and the left edge of the assembly (see *Figure 2c*):

```
hole    : circle([c1,c2,...])
centre  : hole.centre
closest: closest(edge,centre)
dist    : distance(closest,centre)
```

Execution

The geometrical descriptions in the inspection system can be thought of as forming an acyclic dependency graph. Edges in the graph correspond to operations and vertices to the values in the task. Evaluation is lazy and prompted by a request for the value of some entity in the task. In the simple example above it might be a request for the value of `dist`. Execution recursively evaluates all the dependencies until it reaches the landmark points. The system then makes a request to the object search module to evaluate the position of the required landmark points. The object search selects the models relevant to the required points and initiates the search. When the search is complete the search module returns either the point locations or an indication that the points cannot be found. After the landmark points have been determined the recursive dependency calculation unwinds. The values of each of the dependent nodes are calculated until finally the requested value can be returned.

To provide the user with a fast prototyping ability the system is interpretive; changes to inspection task description are effected immediately. Out-of-date management is built into the execution strategy to avoid unnecessary computational effort when changes occur.

Variants

There are situations in inspection where the components to be inspected may not all be identical but may be any one of a set of variants. In our example of brake

inspection, the production line contains a mixture of right- and left-hand assemblies. There is no prior information as to which variant is currently being inspected.

The notion of existence has already been discussed in object search. The existence of certain objects can be used to determine which variant is being inspected. Returning to the example of *Figure 2* we could assume that a variant condition was that an elliptical hole was cut in the sheet instead of a circular one. Thus we might define:

```
c-hole : circle([c1,c2,...])
e-hole : ellipse([e1, e2,....])
```

We require a method for dealing with variants without having to write a separate task description for each possible case. In the example above, once we have fitted the appropriate geometry to the hole the rest of the task is identical.

The `if` construct provides a facility for the selection of one of two objects on the basis of a binary value:

```
if (<binary value>,<object(true)>,
<object(false)>)
```

We can use the if construct in our example to select the appropriate object for the rest of the task to work on:

```
hole   : if(exists(c-hold),c-hole,
         e-hole)
centre : hole.centre
```

When this is executed the point `centre` will receive the position of the centre of whichever type of hole is found in the currently inspected sheet.

Surface property measurement

In the introduction we identified two major aspects to generic industrial inspection. The first was dimensional measurement and the second was surface quality measurement. As with dimensional measurement the objective in surface property measurement is to allow flexibility, adopting a methodology which is user-programmable and user-understandable. We achieve this using the greybox operation.

The greybox surface measurement operation (as opposed to blackbox where the user has no control over the measured properties) is specified like any other operation in the inspection task:

```
greybox(<geometry>,[<properties>])
```

The property list contains a list of properties which the user thinks are relevant to the inspection of the area. The words are in a language familiar to the user, e.g. roughness, streak, blemish, etc. For example, the user may define:

```
surface-vals : greybox(pad-region,
[blemish])
```

Internally, the words are associated with particular texture operators, e.g. Laws texture energy features[6].

Every geometric type defined in the system is expected to be able to supply a set of points, at a specified resolution, describing the area in the image which the geometry covers. This set of points is used as the region of application for the selected texture operators. The greybox operator returns a group of numbers which represent the outputs of the texture operators supplied in the list. We desire as much autonomy as possible from the greybox operation. There is, therefore, no definition of the contents of the group; all that is known is that it is a set of measures related to the properties specified. The group of numbers are fed directly to a classifier for analysis.

Classifiers

We need classifiers not only for surface measurement but also for generating decisions from a geometrical measurement. Often these decisions will be determined by allowed tolerances, but when such tolerances are not available results may be specified by training using examples of both pass and fail classes. For example, we may wish to classify on the basis of the angle between the edges in *Figure 2*:

```
pass : classify(angle-size,a)
```

where `a` is the observable used to produce a boolean result named `pass`. Similarly for inspection of surface quality we have:

```
pass : classify(pad-surface,
surface-vals)
```

When this description is executed the `pad-surface` classifier will classify on the basis of the surface properties held by the `surface-vals` observable. In this case, `pass` will become a boolean value indicating the acceptability of the surface.

In our experimental system we have retained classification as a distinct operation so that it can be used to classify both geometric values and the output of the surface property measurement. Since the greybox output is always applied directly to the classifier it is probable that a future system would collapse the surface property measurement and its associated classification into one operation, e.g.:

```
surf_prop(<geometry>,[<properties>])
```

Classification of geometric values would be retained as a separate operation.

The design of an appropriate general purpose classifier, while within the spirit of the approach we have taken, has not featured in our design to date. We have implemented a straightforward linear classifier to demonstrate how the concept of classification could be used within task description. The question of classifier design and training for a system such as ours is a substantial one. Ongoing work[7] in the area of classification in the face of a low number of training examples is especially relevant to industrial inspection.

EXAMPLE APPLICATIONS

We have used the inspection of brake assemblies as an example of a complex inspection task for which we have direct experience of a hand-crafted solution[2]. This exemplar highlights many of the common requirements of an inspection: checking to see if components are present, measuring the positions of components and determining whether surfaces are finished properly. We have selected three subtasks to demonstrate different aspects of the problem.

Brake lining thickness

One of the requirements of the inspection task is a measurement of the thickness, circularity and co-circularity of the brake shoe surfaces. *Figure 3* shows the geometric task which was entered at the command line to perform the operation. At any time during the prototyping of the inspection the system is able to display the geometric constructions and print task values. *Figure 1* was obtained by entering the following statement at the command line:

```
Display c_in, c_out, out_best, in_ends;
```

The system updates any values automatically before displaying or printing them. If a new image is loaded and the statement `Print results` is issued the system will search the image for the new positions of the objects and recalculate all the required values before returning with a result.

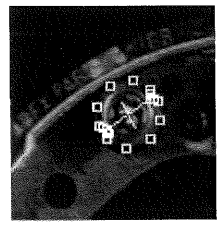
Retaining pin location

The second example from the exemplar is a check to see whether the shoe retaining pin is positioned correctly in the cap. The description and an example of the display are shown in *Figure 4*. Geometric descriptions tend to be compact and apart from training the model the text shows all the statements which are required to perform the angle measurement. If either the pinhead or cap cannot be found in the image, then the system returns the angle as undefined.

The model definition statements for this inspection instruct the system that two models are to be used. The statements specify that the first model is used to locate the general position of the cap within the shoe assembly. They also specify that, using this position as a constraint, the second model composed of a cap and pinhead alone should rotate freely in order to obtain a final accurate position.

```
# p_in and p_out are the points
# on the lining edge

c_out :  circle(p_out);
c_in :  circle(p_in);

# calculate the distances from each
# of the edge points to the circles

out_best :  closest(c_out,p_out);
out_dist :  distance(p_out,out_best);

in_best :  closest(c_in,p_in);
in_dist :  distance(p_in,in_best);

# get the mean value of the distances

out_dev :  mean(out_dist);

in_dev :  mean(in_dist);

# calculate the points on the inside
# circle closest to the outside points

in_ends :  closest(c_in,
[out_best.0, out_best.8]);

# calculate the difference between the
# separation of the circles at the two
# ends of the lining.

end_dist :  distance(in_ends,out_ends);
ends_dev :  sd(end_dist);

# summarise the results.

results :  [out_dev,in_dev,ends_dev];
```

Figure 3 Key statements required to measure circularity and co-circularity of the brake shoe lining



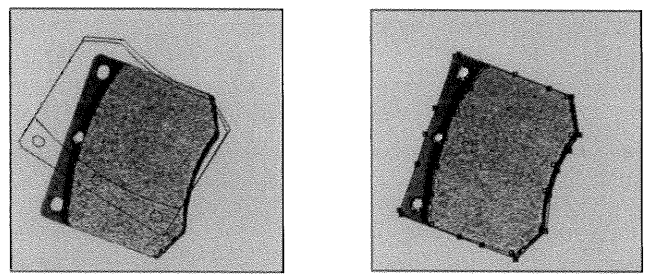
```
# define the centre line of the pinhead and the
slot
# cap and pinhead are aliases for the associated
# groups of boundary landmark points.

cap_l :  line(cap.0,cap.8);
pinhead_l :  line(pinhead.3,pinhead.9);

# measure the angle

angle :  angle(cap_l,pinhead_l);
```

Figure 4 Determination of the location of the pin head in a blind spot. The image shows the two lines and the cap landmark points as found by the object search

Brake pad oil staining

To give an example of surface inspection we demonstrate the construction of a new task on a similar theme, using disk brake pads. *Figure 5* shows the object, a front brake pad from a caliper brake. A number of these pads were prepared with oil stains and the following short task was created to inspect the objects. The task statements with comments are shown in *Figure 6*. Notice that the rotation attribute in the model definition line tells the system that it should search for the assembly at any angle.

Setting up this new inspection, including preparation of the training examples, took only a few hours and



a **b**

Figure 5 (a) Position of the model before search begins; (b) final result

```
# define the parts which are to be
# searched for in the image.

Models pads [[rotate=on]("Outside",
"Inside", "Hole 1","Hole 2","Hole 3")];

# define a classifier to use to
# classify the surface.

Classifiers (oil);

# define the lining to be a region
# covering the area of interest
# ie.  the area bounded by the
# "Inside" landmark points.

lining :  region(pads."Inside");

# measure the surface quaility using
# a grey box operation

values :  greybox(lining,[blotch]);

# finally classification is performed
# leaving oil_stained with the value
# true if the pad surface is stained.

oil_stained :  classify(oil, values)
```

Figure 6 Task required to determine if the brake pads shown in *Figure 5* are oil stained

most importantly involved no addition or change to the inspection system software whatsoever.

DISCUSSION AND CONCLUSION

Not surprisingly, generic tools are composed of individual parts which have strongly generic capabilities. The applicability of the ASM technique to a wide range of image search problems is critical in this respect and underpins the system performance. If subsequent improvements are made to individual aspects such as texture measurement and classification, the modular structure of the system will allow them to be easily incorporated.

In summary, by employing suitable generic components and a modular construction we have been able to construct a visual inspection system which is user configurable but capable of tackling a wide range of inspection problems. Our main objective was to demonstrate that a generic user programmable inspection system is feasible. We believe we have achieved this, demonstrating the functionality of the system in complex inspection tasks.

REFERENCES

1 Bryson, N, Dixon, R N, Hunter, J J and Taylor, C J 'Contextual classification of cracks', *Br. Machine Vision Conf.*, J Illingworth (ed.), BMVA Press (1993) pp 409–418
2 Woods, P W, Taylor, C J, Cooper, D H and Dixon, R D 'The use of geometric and grey-level models for industrial inspection', *Patt. Recogn. Lett.*, Vol 5 (1987) pp 11–17
3 Cootes, T F, Taylor, C J, Cooper, D H and Graham, J 'Active shape models – their training and application', *CVGIP: Image Understanding* (in press)
4 Cootes, T F, Taylor, C J and Lantis, A 'Active shape models: Evaluation of a multi-resolution approach to improving image search', *Proc. BMVC*, York, UK (1994)
5 Jackson, C B *Application generator development*, Automated Visual Systems Ltd, University of Manchester (internal report)
6 Laws, K I 'Rapid texture identification', *SPIE Image Processing for Missile Guidance*, Vol 238 (1980) p 376
7 Frank, I E and Friedman, J H 'Classification: oldtimers and newcomers', *J. Chemometrics*, Vol 3 (1989) pp 463–475

7. **Automation of routine clinical chromosome analysis I. Karyotyping by machine.** J. Graham, *Analyt. Quant. Cytol. Histol., 9: 383 - 390, 1987.*

Reprinted from ANALYTICAL AND QUANTITATIVE CYTOLOGY AND HISTOLOGY
Vol. 9 No. 5, October 1987

Automation of Routine Clinical Chromosome Analysis

I. Karyotyping by Machine

James Graham, Ph.D.

An automated karyotyping sytem suitable for widespread use in clinical laboratories is described. The software is implemented on a general-purpose, commercially available image analyzer (Magiscan 2) using TV input from a conventional research microscope with minimal modification. The analysis is automatic, but operator interaction is used to resolve difficulties. Extensive experience with a routine clinical workload shows that the system is robust and easy to use and that its use results in a substantially increased laboratory throughput.

Chromosome analysis is used in a number of fields for the detection of genetic damage or abnormality. Examples include monitoring of exposure to ionizing radiation or chemical mutagens, monitoring of cancer treatment and genetic counseling.

This paper describes automatic metaphase analysis software running on commercially available image analysis hardware. Together with the metaphase finder described in an accompanying paper,[5] they form a complete chromosome analysis system. This system is intended to be used particularly in the field of clinical karyotyping, where the chromosome constitution (karyotype) of an individual (usually unborn) is assessed for the purposes of genetic counseling or for the diagnosis of a phenotypically observed syndrome. The karyotyping of fetal cells obtained by amniocentesis is usually done only in the case of "high-risk" pregnancies, e.g., if the mother is over 35 years old or there is a history of genetic abnormalities. The demand for clinical amniocentesis is increasingly rapidly[15] due to a number of factors, such as increasing public awareness of the ability to diagnose abnormalities and the identification of new "risk groups," leading to the suggestion of wider population screening.[2,18] This increasing work load can be met either by an increase in the number of skilled (and expensive) laboratory staff members or by the introduction of some form of automated assistance to increase the productivity of the existing staff.

The typical analysis procedure can be outlined briefly as follows. The microscope slide is first scanned at low magnification to identify metaphase cells suitable for analysis. Selected cells are examined at high magnification, and the chromosomes are counted (the number of chromosomes in the cell is a very important diagnostic feature). A few cells in each sample are examined carefully for abnormalities in the morphology or banding pattern of the chromosomes (Figure 1). Perhaps one cell will be photographed, with the chromosomes then cut out of the print and classified into 24 groups; they can be arranged in a tabular array called a "karyogram." Each group in the karyogram contains two homologous chromosomes (with the exception of the male sex

From the Wolfson Image Analysis Unit, Department of Medical Biophysics, University of Manchester, Manchester, England, U.K.

Dr. Graham is Research Fellow in Medical Biophysics.

Address reprint requests to: James Graham, Ph.D., Wolfson Image Analysis Unit, Department of Medical Biophysics, University of Manchester, Stopford Building, Oxford Road, Manchester M13 9PT, England, U.K.

This study was supported in part by a grant from the Wolfson Foundation and in part by sponsorship from Joyce-Loebl Ltd.

Received for publication April 8, 1986.

Accepted for publication January 9, 1987.

0884-6812/87/0905-0383/$02.00/0 © The International Academy of Cytology

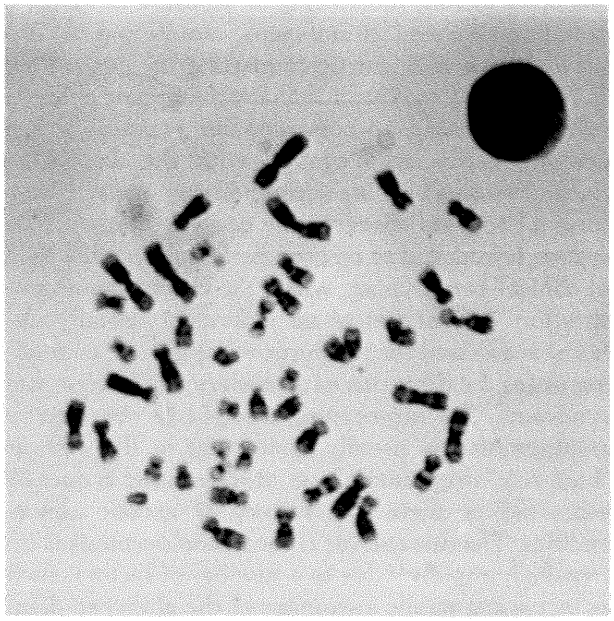


Figure 1
G-banded metaphase cell displayed on the TV monitor.

chromosomes), and abnormalities of structure can be identified and documented easily in this format.

Automation can be introduced into this process, not only in finding metaphase cells,[5] but in counting, measuring and classifying chromosomes and producing a karyogram.

System Design

The automated karyotyping system described here has been implemented on the Magiscan 2 instrument produced by Joyce-Loebl, Gateshead, England, U.K. The instrument is completely software-based; all image acquisition and manipulation processes are programmable, providing the flexibility to tackle the substantial image analysis problems involved. Its architecture has been described by Taylor et al.[21] The CPU, a microprogrammable processor with a cycle time of 150 nanoseconds, has access to a microprogram memory of 4K words and a 128 K-byte macromemory holding high-level programs and data. The 1 Mbyte of image memory is organized as eight bit-planes of 1K×1K pixels. Under program control, images can be defined to occupy any volume of this space, from a 256×256-pixel binary image to a 1K×1K image with 256 gray levels. For this application, TV camera input is used, so that gray-level images are stored at 512×512-pixel resolution with 64 gray levels.

Operator interaction with the machine is possible using a single-pixel-resolution lightpen and a conventional keyboard. The existence of a compact set of software primitives for image analysis[6] enhances both runtime efficiency and program development.

The remainder of the system consists of a conventional research microscope (Orthoplan; Leitz, Wetzlar, West Germany) equipped with a motorized stage capable of holding up to eight slides (Märtzhaüser, Germany). The X-Y and focus movements of the stage together with the secondary magnification (zoom) and lamp brightness are under control of the Magiscan via a microscope control unit supplied by the manufacturer. This unit also controls a Honeywell video hard copy unit.

Chromosome Analysis

This section describes in detail the process of automatic analysis. The description refers to G-banded chromosomes, which are the most commonly occurring type of clinical sample. A similar analysis program exists for unbanded chromosomes, and certain parts of the analysis are more straightforward in this case. Points at which a different approach is adopted for unbanded chromosomes are indicated in the text.

Counting

All metaphases examined must have their chromosomes counted. For this purpose, it is only necessary to identify where chromosomes are, and speed is important. For this reason, the image is segmented using a globally determined threshold. A histogram of gray levels in a central 256×256-pixel square is bimodal or highly skewed, and a threshold value can easily be obtained from it that is biased towards underdetection to avoid joining closely adjacent chromosomes.

Machine-counted chromosomes are marked with a dot so that the operator can quickly identify miscounts, such as undetected chromosomes, composite chromosomes and nonchromosomal debris. Interactive correction of the count can be made by simply indicating the miscounted object with the lightpen. If it is counted already, it is removed; if not, it is included, and the machine count is updated accordingly. Thus, the operator only has to identify erroneous counts as displayed rather than count all chromosomes in the image.

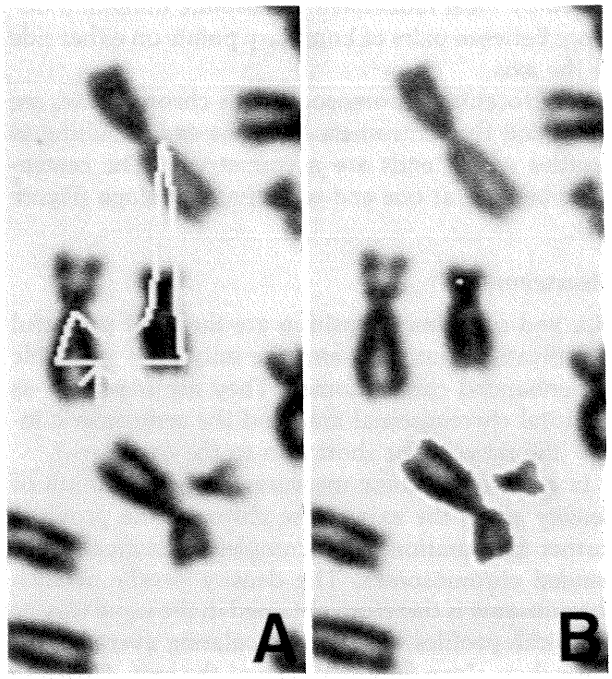


Figure 2
(A) The gray-level histogram obtained in the locality of the pair of touching unbanded chromosomes at the bottom of the picture. Dark gray levels are to the left of the histogram. The arrows indicate two thresholds: a "best" threshold in the middle of the intermodal minimum and an "underdetecting" threshold. (B) Using combined thresholds, the chromosomes can be separated while retaining well-placed boundaries.

Fine Segmentation

For most cells, the analysis stops at the count, but for a few cells—those whose chromosomes are well formed and reasonably well separated—a complete analysis will be performed, resulting in a classification of the chromosomes. The first stage in this process is to obtain a "best" segmentation of the image. The segmentation has two aims: (1) to separate the chromosomes from the background and from each other and (2) to provide well-defined chromosomal boundaries.

In the case of both banded and unbanded chromosomes, thresholding based on local gray-level segmentation is used. Two thresholds are selected from the histogram: (1) a "best" threshold, which selects the gray level at which the overall boundary should be placed, and (2) an "underdetecting" threshold to split the composite. In the case of unbanded chromosomes, these thresholds are easily found (Figure 2A) and the underdetected objects expanded back to the original boundary, producing separate objects with well-placed boundaries (Figure 2B).

The histograms of banded chromosomes, particularly composite ones, are usually more complex, and selection of two thresholds is much more error-prone (Figure 3A). This, together with the gray-level variability within the chromosomes, means that the underdetection and expansion algorithm is often unsuccessful. The best threshold has to be selected rather more conservatively (Figure 3A) since banded chromosomes are frequently lightly stained at the edges. The result is that composite chromosomes occur (Figure 3B).

Axis Fitting and Centromere Finding

In order to make measurements on the chromosome so that classification features can be extracted, it is necessary to find the medial axis and the centromere (Figure 4). The centromere is a characteristic con-

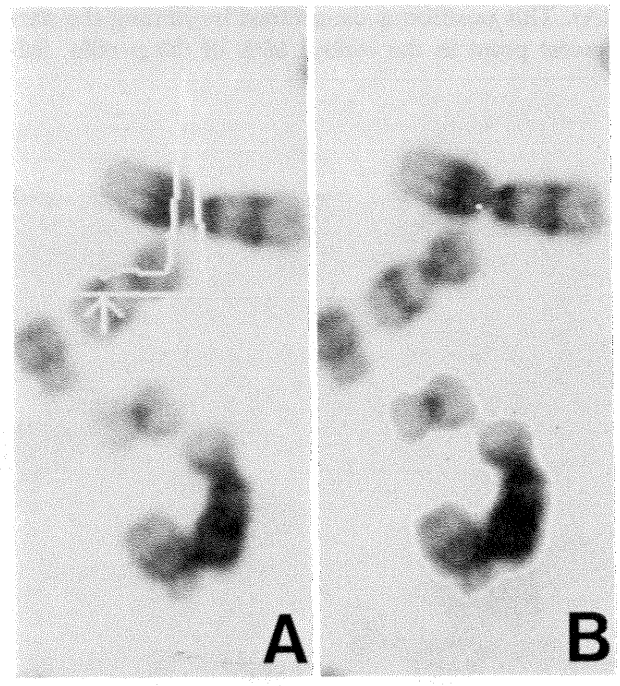


Figure 3
(A) The local gray-level histogram for a pair of banded chromosomes. The distribution is not bimodal in this case, and the "best" threshold is placed at the foot of the background peak. There is no reliable information on which to place an "underdetecting" threshold. (B) Banded chromosomes are not easily separated, and composite chromosomes result.

striction that can occur in a variety of positions along the chromosome length; its position is a very powerful feature in classifying chromosomes.

Most chromosomes are long and thin; they may be either straight or bent. For straight chromosomes, a least-squares straight line fit to the boundary coordinates (in which the minimand is the sum of the squared pythagorean distances of each boundary point to the fitted line[4]) provides a good axis. "Half-width" profiles of the chromosomes are calculated by measuring, for each axis point, the distance to the boundary on either side along the normal to the axis. If the half-width profiles are highly dissimilar, the chromosome is asymmetric and assumed to be bent. A bent axis is generated by fitting a cubic curve to a set of "back bone" points. These points are those lying halfway between opposing boundary points on the normals to the straight axis. A new pair of half-width profiles, generated for the curved axis using local normals at each axis point, can be used to determine the centromere position. Two cases can occur.

Metacentric or submetracentric chromosomes may be large or small; the centromere is in the body of the chromosome, resulting in a dip in the profile (Figure 4A). This position is determined by finding the narrowest point in the central 60% of the profile, followed by local refinement to absolute minimum distance between pairs of boundary points on either side of the axis.

In acrocentric chromosomes, the chromosomes are small and the centromere is at one end, resulting in profiles whose ends are asymmetrical. The centromere belongs at one end with the least slope (Figure 4B).

Measurements

Size and centromere position are the most powerful classification features—and the only ones available for unbanded chromosomes. They are measured as the total chromosomal area and the centromeric index (the ratio of the short arm to the total area).

In addition to these measures, the distribution of density along the axis of the chromosome provides further information for a complete classification of banded chromosomes. The density profile of each chromosome is therefore obtained in the same way as the width profiles, this time calculating average density values along local normals of the axis. Groen et al[9] pointed out that sampling a rectangular grid along lines skew to the cartesian axis produces errors unless the sample points are interpolated between the grid points. In consideration of the method used to extract

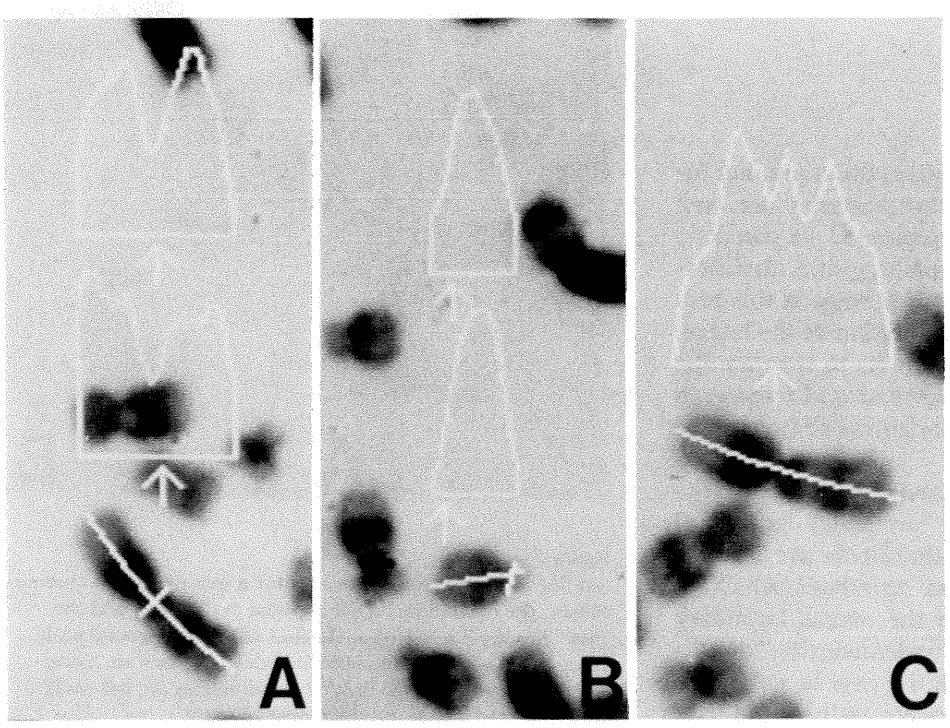


Figure 4
(A) The (curved) axis and "half-width" profiles of a metacentric chromosome. The centromere position is indicated and corresponds to a position where significant minima in the two profiles coincide (arrows). (B) The axis and half-width profiles of an acrocentric chromosome. There are no minima in the profiles, and the centromere is placed at the end of the chromosome where both profiles have least slope. (C) The density profile along a chromosome axis. The arrow marks the centromere position on the profile. The features used for classification represent the gross properties of this profile.

classification features from the profile (next section), such careful measurement is unnecessary, and samples are taken from grid points closest to the sampling line. A typical density profile is shown in Figure 4C.

Classification

Unbanded chromosomes are classified into ten groups approximating to the "Denver" classification using a two-dimensional classifier based on size and centromeric index. Class probabilities are derived from a training set of 100 cells, and assignment is made using the constraint that the classes have fixed sizes. The assignment procedure used always assigns the most likely chromosome to a given group (rather than assigning each chromosome to its most likely group). This does not guarantee a globally optimal classification, but is computationally straightforward and, in practice, gives results that are not much worse than the theoretical best classification, with a misclassification rate of about 5%.[22]

Several authors have tackled the problem of extracting useable features from the density profile for classifying banded chromosomes.[7,8,14,23] The method of Granum[8] has been used by Lundsteen et al[12] of Rigshospital, Copenhagen, to classify a set of carefully selected and measured chromosomes, giving classification error rates of about 2%. It derives features for a Bayesian classifier by multiplying the density profile by a set of weight functions, some of which depend on the centromere position. This has the advantage that it responds to gross features of the profile and is largely insensitive to the frequent small variations in the profile shape between chromosomes of the same group.

The Copenhagen implementation of Granum's classifier has been incorporated into our system for classification of the density profiles described in the previous section. Classification results on more than 2,500 metaphases from amniotic fluid samples in a routine clinical workload show an average misclasification rate of about 8%.[10]

Karyogram

To facilitate clinical evaluation, the chromosomes as classified are displayed in karyogram format. The chromosomes are translated to the appropriate locations in the karyogram image and rotated to have their major axis upright, with their centromeres in line (Figure 5). There is no restriction on the angle of rotation so that, in general, the gray-level value at a pixel in the karyogram has been obtained from a nonintegral point in the original metaphase image. These gray values are calculated by bilinear interpolation between the four closest integral points.

A hardcopy of the karyogram for patient documentation can be obtained instantly via the gray-scale printer.

Operator Interaction

Difficult points in the analysis are overcome by recourse to operator interaction. Chromosome counting is an example of a necessarily interactive procedure in which the operator can use the lightpen to amend the image in a straightforward manner.

There are a number of difficult analysis situations or errors that may occur. The most common are composite chromosomes and misplaced centromeres. Centromere correction can be accomplished by a single stroke of the lightpen, as can the resolution of composites arising from the incorrect segmentation of touching chromosomes. Composites of overlapping chromosomes may require a few lightpen "cuts." Other interactions, for example, to reject nonchromosomal objects or to correct the axis of a badly bent chromosome, are required less frequently but are equally easy, requiring only the lightpen to indicate an object or draw a line in an obvious place.

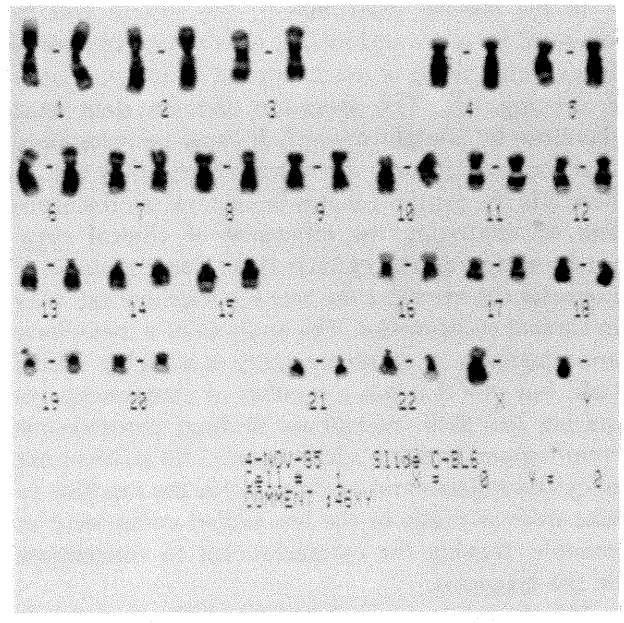


Figure 5
Machine-generated karyogram for inspection by the cytotechnician. Patient details are entered and form part of the hard-copy documentation.

Even in the absence of other errors, classification errors occur (on average, about four per metaphase). These can also be corrected interactively by using the lightpen to move chromosomes around the karyogram image. As the chromosomes are moved to their correct locations, their classifications are reassigned.

Discussion

Machine-aided karyotyping provides a possible solution to the throughput bottleneck being experienced by many clinical laboratories.[15,16] The system described here has been designed with the intention of being used widely in the routine clinical environment. This imposes important constraints on the implementation: constraints of cost, availability and robustness. The system must be cheap enough to be purchased as a piece of capital equipment on a typical health service budget, and if it is to have any effect on clinical workloads, it must be widely available. For this reason, the system was implemented as a software package on a general-purpose, commercially available image analyzer with standard peripherals. (The only modification to the microscope has been the addition of stepping motors to the focus and secondary magnification controls.)

A corollary of widespread use is the ability to handle a wide variety of sample qualities, which requires substantial robustness and flexibility of the software.

In the clinical environment, the system will be operated by a cytotechnician, who cannot in general be expected to have much interest in (or sympathy for) computers. The operation of the system must therefore be straightforward. It must be recognized that these constraints are not arbitrary; they are inherent in the problem that is being tackled: the problem of improving the efficiency of clinical cytogenetics laboratories. Given these constraints, it is important to consider the role the machine can play in clinical cytogenetics. The analysis of a metaphase and diagnosis of an abnormality is a highly skilled task, but one that has a number of components requiring less skill: metaphase finding, chromosome counting and forming a karyogram. An efficient use of cytotechnician and machine is for the machine to take over as much of the less-skilled component as possible, leaving the cytotechnician to concentrate on the diagnosis.

The images to be analyzed are difficult, and it is necessary for the cytotechnician to become involved to some extent in the analysis to resolve difficult situations. There is no harm in this—indeed, it makes good use of the operator's highly developed image analysis skills—provided that the interaction is easy and natural for an operator who has a deep understanding of the image, though not of the machine.

In short, the system is a tool that the cytotechnician can use in performing the analysis. It should be stressed that the machine analysis is automatic. The operator interaction is only required to correct machine errors. The usefulness of the tool must be judged on the following criteria: how easy is it to use and by how much does it improve efficiency?

This system has been under clinical trial at a busy clinical genetics laboratory at Rigshospital, Copenhagen, since 1983 and now contributes to the routine analysis workload in that laboratory. The mode of operation of the system is to run the automatic metaphase finding[5] overnight on eight slides and to analyze ten cells from each slide the following day. Four of these ten cells have karyograms made. The time required to produce a machine-aided karyogram of an individual cell depends on the quality of the preparation and is highly variable between cells; a typical figure is 4.5 minutes.[17] (An earlier study showed that the average time required to analyze unbanded cells was 2.5 minutes.[13]) A more meaningful figure is the total time required to analyze a complete slide, which in this study was found to be 37.5 minutes. Four technicians of varying degrees of familiarity with the system acted as operators during the trial; the throughput of these technicians was approximately double their expected manual rate.

The eight-slide capacity of the motorized microscope stage limits the number of analyses to eight per day in practice. Given the observed analysis time, there is plenty of time to perform this number of analyses in one day—even allowing a 50% dead time between samples—using the equivalent of one operator. Dividing the time between (say) four technicians is obviously a sensible practice. The use of a larger-capacity stage and a less-generous dead time would allow ten or more analyses per day. However, eight analyses per day in a 250-day working year means 2,000 samples analyzed per year; this is about half the current workload of the Rigshospital laboratory with a staff of nine full-time technicians.

Automatic analysis is forming an increasingly important part of the routine workload at Rigshospital. Almost one-quarter of all prenatal samples in 1984 and over one-third of the samples examined in the first nine months of 1985 were analyzed with this system. For this purpose, two work stations are used, each with a computer-controlled microscope. In addition, a third machine, without an automated mi-

croscope, is used to help in the analysis of the remaining manually analyzed samples, where one karyogram is produced to document the analysis. The Magiscan is used for this in preference to the conventional "photograph and scissors" method. Between January and September 1985, 1,729 such machine-aided karyograms were produced.[11]

The single most significant problem with the system at the moment is the number of interactions needed to produce one karyogram. For samples of routine quality, this number is 46 on average.[11] The main requirement for interaction is in the separation of touching chromosomes. The method currently used to try to separate such composites during the automatic analayis is a binary opening operation on the detected objects; this is successful in only a very few cases. Recent experiments show that a method similar to that described for separating unbanded chromosomes may dramatically reduce the number of necessary interactions. This method will be described elsewhere: it does not yet form part of the clinically tested system.

The automated karyotyping system described here is being shown to contribute significantly to the work of a routine clinical laboratory. Of course, work on applying image analysis techniques to chromosome analysis has been in progress for some 20 years now. These developments have been reviewed by Piper et al,[19] who described many useful algorithms and approaches to the problem; these authors also emphasized that the analysis of these images is difficult and there has been little progress towards clinical implementations. The present state of the art has been well described by Rutovitz.[20] One previous well-founded attempt has been made to create an automatic chromosome system suitable for use in clinical cytogenetics laboratories; that was the system put together by Castleman and his colleagues[1] at the Jet Propulsion Laboratory (JPL) in 1975. The requirements for robustness, speed and low cost were beyond the technology of the time, making the system slow and difficult to use. It was never taken past an evaluation phase.[16] The system described here represents, in a sense, the implementation of what the JPL system was striving for. In addition to the extensive clinical trials in Copenhagen, independent trials have been taking place at the Northwestern University Medical School in Chicago[16] and at Guy's Hospital in London.[3]

There are applications of chromosome analysis other than clinical karyotyping that could benefit from machine assistance. In some, such as aberration scoring, interactive analysis of the type described here may not be appropriate; in others, such as cancer cytogenetics, operator interaction may be highly useful. Clinical karyotyping is an application in which interactive automation provides tangible benefits in tackling a real clinical problem while providing valuable experience of the challenges involved in implementing such systems in other fields.

Acknowledgments

I am indebted to colleagues at Rigshospitalet, Copenhagen, particularly Tommy Gerdes for implementation of the banded classifier and Claes Lundsteen, Anne-Marie Lind and Jan Maahr for forceful presentation of the user's point of view and expert assistance in collecting classification data. Many discussions with colleagues here in Manchester, particularly Chris Taylor, have contributed greatly to the system development. Thanks are due to Joyce-Loebl Ltd., for their faith in this system as a saleable product, and their assistance in bringing it to a marketable state.

References

1. Castleman KR, Melnyck JH, Frieden HJ, Persinger GR, Wall RJ: A minicomputer based karyotyping system. *In* Automation of Cytogenetics. Edited by ML Mendelsohn. Asilomar Workshop Conference 751158, 1975, pp 46–49
2. Cuckle HS, Wald NJ, Lindenbaum RH: Maternal serum alpha-fetoprotein measurement: A screening test for Downs syndrome. Lancet 1:926-929, 1984
3. Daker M: The detection of chromosome abnormalities using Magiscan 2. Report to the European Chromosome Analysis Workshop, Leiden, The Netherlands, 1985
4. Dudani SA, Luk AL: Locating straight line edge segments on outdoors scenes. Pattern Recogn 10:145-157, 1978
5. Graham J, Pycock D: Automation of routine clinical chromosome analysis: II. Metaphase finding. Analyt Quant Cytol Histol 9:391-397, 1987
6. Graham J, Taylor CJ, Cooper DH, Dixon RN: A compact set of image processing primitives and their role in a successful application program. Pattern Recogn Letters 4:325-333, 1986
7. Granlund GH: Identification of human chromosomes by using integrated density profiles. IEEE Trans Biomed Engl BME-23:182-192, 1976
8. Granum E: Application of statistical and syntactic methods of analysis and classification to chromosome data. *In* Pattern Recognition Theory and Practice. Edited by J Kittler, KS Fu, LF Pau. Dordrecht, Holland, D Reidel, 1982, pp 337–398
9. Groen FCA, Verbeeck PW, Van Zee GA, Oosterlink A: Some aspects of the computation of banding profiles of chromosomes. *In* Proc 3rd International Joint Conference on Pattern Recognition, Coranado, 1976, pp 547–550
10. Lundsteen C, Gerdes T, Maahr J: Automatic classification of

chromosomes as part of a routine system for clinical analysis. Cytometry 7:1-7, 1986

11. Lundsteen C, Gerdes T, Maahr J: Semiautomated prenatal chromosome analysis. Report to the European Chromosome Analysis Workshop, Leiden, the Netherlands, 1985
12. Lundsteen C, Gerdes T, Philip K: The Rigshospitalet Chromosome Analysis Program System: RACAPS. *In* Proc IVth European Chromosome Analysis Workshop, Edinburgh, 1981, pp 4.7, 1–6
13. Lundsteen C, Gerdes T, Philip J, Graham J, Pycock D: An interactive system for chromosome analysis: Tests of clinical performance. *In* Proc IIIrd Scandinavian Conference on Image Analysis, Copenhagen, 1983, pp 392–397
14. Lundsteen C, Granum E: Description of chromosome banding patterns by band transition sequences. Clin Genet 15:418-429, 1979
15. Machol L, Queenan JT, Morris J, Wallach EE, Prescott G, Hsu L, Schulman JD: Can we meet the increasing demand for genetic amniocentesis? Contemp Ob/Gyn 16:77-84, 1980
16. Martin AO: My life with two image analysis systems. Report to the European Chromosome Analysis Workshop, Leiden, The Netherlands, 1985
17. Philip J, Lundsteen C: Semi-automated chromosome analysis: A clinical test. Clin Genet 27:140-146, 1985
18. Philip J, Tabor A, Bang J, Madsen M: Fetal chromosome analysis: Screening for chromosome disease. Prenatal Diagn 3:209-218, 1983
19. Piper J, Granum E, Rutovitz D, Ruttledge H: Automation of chromosome analysis. Signal Processing 2:203-221, 1980
20. Rutovitz D: Automated chromosome analysis. Pathologica (suppl) 75:210-242, 1983
21. Taylor CJ, Dixon RN, Gregory PJ, Graham J: An architecture for integrating symbolic and numerical image processing. *In* Intermediate Level Image Processing. Edited by MJB Duff. London, Academic Press, 1986, pp 19–34
22. Tso MKS, Graham J: The transportation algorithm as an aid to chromosome classification. Pattern Recogn Letters 1:489-496, 1982
23. Vanderheyt L, Oosterlink A, Van Daele J, Van den Berhe H: Design of a graph-representation and Fuzzy-classifier for human chromosomes. Pattern Recogn 12:201-210, 1980

8. **Automation of routine clinical chromosome analysis II: Metaphase finding.** J. Graham and D. Pycock, *Analyt. Quant. Cytol. Histol., 9: 391 - 397, 1987.*

Reprinted from ANALYTICAL AND QUANTITATIVE CYTOLOGY AND HISTOLOGY
Vol. 9 No. 5, October 1987

Automation of Routine Clinical Chromosome Analysis
II. Metaphase Finding

James Graham, Ph.D.
David Pycock, B.Sc.

Metaphase finding is an essential activity in chromosome analysis, and there is much to gain from its automation. This paper describes software for automatic metaphase finding developed for use as part of a routine clinical chromosome analysis system, principally for samples from blood and amniotic fluid. Since the metaphase finding and analysis programs were intended to be used widely in clinical laboratories, cost and portability were important design features. The metaphase finder has been implemented on a moderately priced, general-purpose image analyzer (Magiscan 2), which controls a standard research microscope with motorized stage and focus. Metaphases are detected using fast gray-level processing on whole fields of view, followed by binary processing to produce a figure of merit for each detected object. Clinical experience has shown that this ability to rank detected objects on the basis of their suitability for analysis is a critical feature in determining the usefulness of an automatic metaphase finder.

A fundamental stage in any task involving visual chromosome analysis is the location of dividing cells. Most applications involve the use of cells at metaphase, when the chromosomes are most condensed, and so this operation is usually described as metaphase finding, although cells at prometaphase or prophase may be required. Dividing cells are fairly rare occurrences in most types of samples used for chromosome analysis, and only a small proportion of those located are suitable for analysis; thus, a large part of the time a cytogeneticist spends in analyzing a sample is devoted to finding appropriate cells. This may be as much as 25% to 50% of the time for samples from amniotic fluid and even more in such cases as bone marrow preparations, in which an entire microscope slide may yield only one or two metaphases. Since cytogeneticists are highly trained and expensive, it makes good operative and economic sense to relieve them of this drudgery. Much work has been devoted to using image analysis as the basis of automatic metaphase finders.[5,6,12]

The accompanying paper[3] describes a metaphase analyzer intended primarily for clinical use, i.e., with the analysis of samples from blood or amniotic fluid as its principal aim. An automatic metaphase finder is an essential component of a system for automatic metaphase analysis. The metaphase analysis software was implemented on a moderately priced, general-purpose image analyzer because cost is a limiting factor in the implementation of real solutions to the problem of routine analysis. This paper describes a metaphase finder implemented on the same equipment. The metaphase finder program is separate from the analysis software and can stand alone, providing metaphases for analyses that are not automated, such as screening for fragile X chromosomes.

Figure 1 shows a field from a slide of amniotic fluid culture as seen on a TV monitor. The metaphase cell is to be identified and distinguished from the nondividing cells and debris. Most fields on a slide would not contain any metaphases, but may contain nondividing cells and possibly some debris. For the purposes of a finding algorithm, metaphases are clusters of 20 to 40 small dark objects, spread over an area

From the Wolfson Image Analysis Unit, Department of Medical Biophysics, University of Manchester, Manchester, England, U.K.

Dr. Graham is Research Fellow in Medical Biophysics.

Mr. Pycock is Research Fellow in Medical Biophysics.

Address reprint requests to: James Graham, Ph.D., Wolfson Image Analysis Unit, Department of Medical Biophysics, University of Manchester, Stopford Building, Oxford Road, Manchester M13 9PT, England, U.K.

This study was supported in part by a grant from the Wolfson Foundation and in part by sponsorship from Joyce-Loebl Ltd.

Received for publication April 8, 1986.

Accepted for publication January 9, 1987.

0884-6812/87/0905-0391/$02.00/0 © The International Academy of Cytology

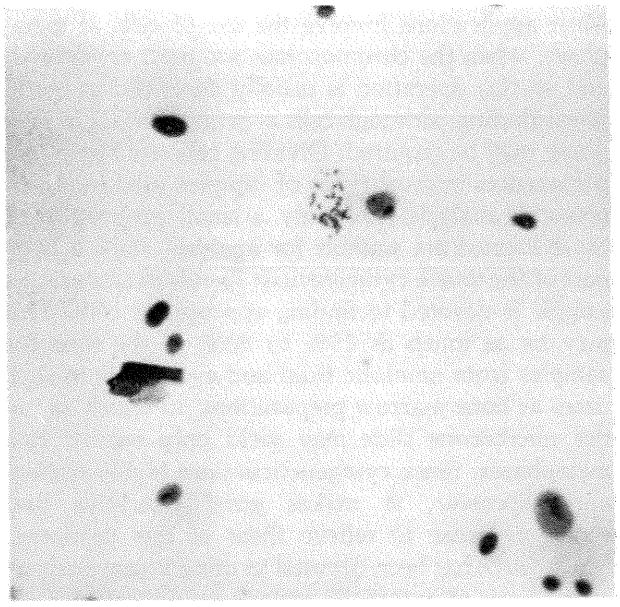


Figure 1
A typical field of view from an amniotic fluid slide.

whose extent is roughly predictable. The metaphase finder's job is to identify such regions quickly and reliably and to make some measure of the "quality" of the metaphase, i.e., its suitability for analysis.

Implementation

Hardware

Like the metaphase analysis program described in the accompanying paper,[3] the metaphase finder is implemented on a Magiscan 2 image analyzer produced by Joyce-Loebl, Gateshead, England, U.K. The architecture of this instrument was described by Taylor et al.[11] From the point of view of metaphase finding, the salient features of the instrument are: (1) the processor is fast and the instruction set is optimized for fetching and manipulating image data over small neighborhoods, thus allowing efficient local gray-level operators to be applied; (2) data structures and low-level (microcode) routines exist that allow efficient operation over arbitrarily shaped lines and regions[4]; and (3) all operations are performed in software: there are no special hardware modules. A library of general-purpose image analysis routines exists in microcode and is accessible from high-level programs.

Images are obtained via a TV camera from a standard research microscope (Leitz Othoplan) fitted with a stepping stage (Märtzhaüser) having a capacity of eight 2.5-cm×7.5-cm slides. The microscope is fitted with motorized drives for the focus and secondary magnification (zoom) controls; the latter is not required for metaphase finding, however. The X-Y movement of the stage has a step size of about 5 μm and the focus is about 0.1 μm. All the motor drives and the microscope lamp brightness are under control of the Magiscan by way of its peripheral microscope controller.

Defining Regions of Interest

The speed penalty implied by not being able to have some critical part of the operation performed in dedicated hardware is to some extent overcome by the ability to define regions of interest and to concentrate the discrimination effort on these. By the nature of the problem, regions of real interest constitute a very small proportion of the total area searched. A useful cue in defining regions of interest, and one that is exploited efficiently by fast local gray-level processing, is the local gray-level variability. Metaphases are regions of stippled texture, 50 μm to 100 μm in diameter. At the magnification used (160×), the pixel size is about 1 μm, so measuring gray-level variation within a 32×32-pixel box should be capable of distinguishing between potential metaphases and background or large cells.

The texture measure used is illustrated in Figure 2 and can be written

$$D_i(i,j)+D_j(i,j)$$

where $D_i(i,j)$ and $D_j(i,j)$ are second derivative operators in the X and Y directions, respectively, and are defined as follows:

$D_i(i,j)=2g(i,j)-g(i-2,j)-g(i+2,j)$
if $g(i,)>g(i-2,j)$ and $g(i,j)>g(i+2j)$;
$D_i(i,j)=0$ otherwise.
$D_j(i,j)$ is similarly defined with respect to the Y axis;
$g(i,j)$ is the image value at pixel(i,j) in the box.

If the integrated texture over the 32×32-pixel box is above threshold, the entire box is masked in a binary mask image. The box origin is moved forward by one pixel, and the texture from the new line thus generated is substituted for that of the old line left behind; thus, the box and its associated texture rolls along the image with masks being set up where the texture measure is high. This results in irregularly shaped regions of interest being defined (Figure 3A).

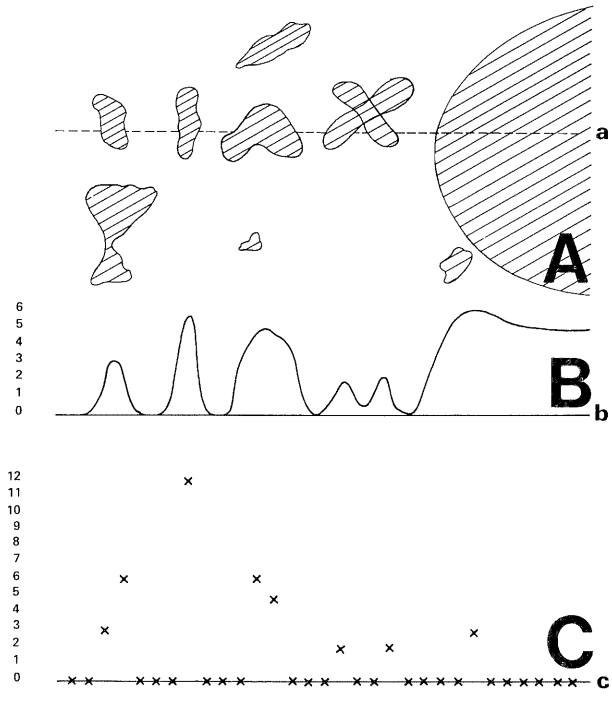


Figure 2
The texture operator. (A) A collection of objects that may be required to be differentiated and a scan line along which the X component of the operator is to be applied. (B) The gray-level density distribution along the line (gray values in arbitrary units). (C) The result of applying the texture operator in one dimension along the line (see the text). Notice that high responses come from isolated small objects, and a large step at the edge of an extended object gives only a small response. Similar responses are obtained in the Y direction, and the operator value is the summed response over a small square neighborhood.

Analysis of Regions of Interest

Attention can now be focused on a slightly expanded (binarily dilated) version of the region of interest. The simplification arising from gray-level thresholding is of use here, but problems of staining and background variability can be overcome by selecting the threshold on the basis of a local histogram of gray levels. A collection of "objects" in the binary image is obtained (Figure 3B). The following binary processes allow some measure of assessment of the quality of the candidate region. The objects are first counted: some 20 to 40 objects might be expected of a metaphase, depending on its quality and how good a threshold was obtained. The mean size of the objects can be determined by binary opening (an erosion followed by a dilation, which tends to remove objects smaller than the erode/dilate mask) (Figure 3C). A diameter of two or three pixels might be expected for chromosomes at a 160× magnification. Binary closing (a dilation followed by an erosion, which tends to fill gaps smaller than the mask size) is used to measure object separation. Five or six pixels separation might be typical of a metaphase (Figure 3D). The area of the closed object, i.e., of the metaphase, and the total detected area are also useful measures for assessing metaphase quality. For each of these parameters, a range of values will be appropriate for metaphases, with other values being indicative of nonmetaphases. A trapezoidal merit function (Figure 4) is used for each parameter, giving full weight to values corresponding to good metaphases, tailing off to zero for values corresponding to poor metaphases or nonmetaphase objects. The break

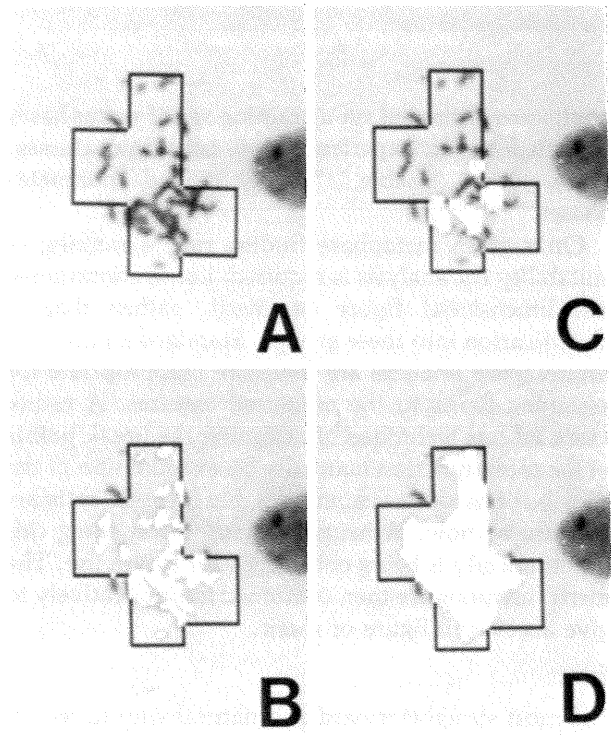


Figure 3
Analysis of the regions of interest. (A) An irregularly shaped region of interest defined by a pass of the texture operator. (B) Segmentation within this region using a threshold based on a local gray-value histogram. (C) Binary-opened version of Figure 3B. Most isolated objects are removed, but clumpy areas remain. This allows a crude size distribution to be calculated. (D) Binary-closed version of Figure 3B. This allows object separation and overall metaphase size to be assessed.

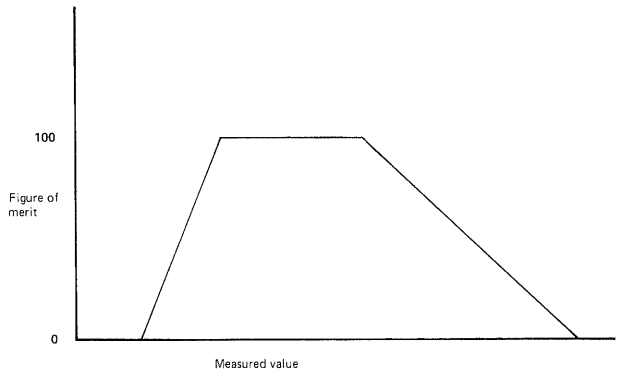


Figure 4
A trapezoidal function used in obtaining a figure of merit from measurements such as those in Figure 3. A range of values produces a maximum figure of merit of 100. The function tails off linearly to zero on either side of these optimum values. The junction positions between the flat and sloping sections are determined from training data.

points are estimated on a training set of metaphases classified by an experienced operator into classes: "very good," "usable," "unusable" and "nonmetaphases."

On a given metaphase finding run, a ranking of suitability for analysis is required, i.e., a continuous, one-dimensional figure of merit, rather than a classification into these groups. Standard methods of multivariate analysis are therefore inappropriate for assigning limits to the measured features. A rather more *ad hoc* technique of assigning the break points in the merit function manually by examination of the distributions of measurements has proved satisfactory up to now. A heuristic method for doing this automatically is being considered for future use. The merit functions are then combined multiplicatively to give an overall figure of merit.

Focus

The most straightforward and natural way to search a slide is to select a rectangular area and step through it in a raster pattern. The X and Y coordinates of the microscope stage for each field are easily calculated. The Z-position (focus) varies in an unpredictable way over this area. It is crucial for the efficient operation of the metaphase finder that every field examined should be in focus. An early approach to this problem was to have the operator focus the microscope at the four corners of the scan area and to determine the Z coordinate of each raster field by bilinear interpolation. For small search areas, this method works quite well, but in order to obtain sufficient cells for analysis, larger areas are required and autofocus must be implemented.

The metaphase finder uses a standard light microscope with minimal alteration: there is no special autofocus hardware, which would allow continuous focus monitoring. Instead, autofocusing algorithms are applied at selected fields. The optimum use is made of the information obtained from these fields if the autofocus algorithm is applied before scanning to a sparse grid of fields covering the search area, with the focus at interstitial fields being determined by bilinear interpolation as in the "four corner" method described above. The algorithm for determining best focus at a particular field is supplied by Joyce-Loebl as part of the operating software and will not be described here.

Practical Operation

Overall control of the scanning resides in a driving file on disk, which contains information about the area of slide to be scanned, appropriate parameters for figure-of-merit calculation, number of metaphases to be found before terminating the scan, etc. Typically, one such file exists for every sample type likely to be examined (amniotic fluid, blood, bone marrow, etc.).

The operator's tasks in initiating the scan are (1) to define a scanning origin, (2) to approve the focus at the four corners of the scan area for each slide and (3) to enter a slide identifier under which results are to be filed. The remainder of the operation—generation of a focus grid and scanning metaphases—is automatic and proceeds without intervention, typically overnight.

Discussion

The metaphase finder described here is designed to be part of a chromosome analysis system for routine clinical use. It is argued in the accompanying paper[3] that clinical use constrains the implementation to involve only readily available, relatively inexpensive hardware. This argument applies equally to a metaphase finder, whether it is part of an automatic analysis system or a stand-alone tool. We have therefore avoided solutions involving hardware specifically designed for metaphase finding. The principal penalty arising from basing this implementation on a general-purpose, TV-based image analyzer is loss of speed. The strategy for analyzing each field—rapidly

Table I *Typical Performance Figures for Metaphase Finders**

	Blood	Amniotic fluid
Scanning speed (/sq cm)	13′7″	13′48″
False positives (%)	3	30
False negatives (%)	40	29
Ranking†	7	—

*Adapted from Bresser.[1]
†Number of relocations necessary to obtain five cells of sufficient quality to produce a karyogram. The ranking for amniotic fluid appears to be about 10 from other studies.[7,9]

defining small regions of interest and applying detailed analytical methods only to them—goes some way towards reducing this time penalty.

Clinical Performance

The main clinical workload is in blood and amniotic fluid preparations (with a small but increasing proportion of chorionic villi samples). These preparations provide a fairly rich source of metaphases (compared, for example, to bone marrow samples), of which only a few are required for analysis. This affects the performance requirements for the metaphase finder. We have some considerable experience in operating on routine samples from a clinical trial in the chromosome laboratory at Rigshospitalet, Copenhagen.[9]

Performance figures are shown in Table I. These are summarized from results obtained in an independent trial of metaphase finders[1] and are based on 0.5-cm × 0.5-cm areas of four blood and two amniotic fluid samples. The fact that so few measurements were made reflects the extreme tedium and difficulty involved in making them. They are, however, in general agreement with our previous measurements of the same attributes.[8] More important than the actual values, or their accuracy, is what they mean for practical operation.

The speed of search is such that a 2-cm × 1-cm area of the slide can be scanned in about half an hour. This area is normally adequate to provide a sufficient queue of metaphases for analysis from amniotic fluid samples. It should be noted that, since metaphase finding takes place unsupervised overnight on a number of slides, speed is not a matter of great importance. There is no time cost imposed on the day-to-day running of the laboratory. Assuming an eight-hour working day and taking eight slides as a typical analysis load, there is time to scan a 2-cm × 4-cm area of each slide overnight. This may be necessary for chorionic villi or bone marrow preparations; our experience with these is limited.

The false-positive and false-negative figures reflect the frequency of occurrence of good metaphases in the two types of sample. Blood slides are heavily populated with good metaphases, and parameters can be adjusted so that only the best metaphases are retained, keeping false positives low. In the case of amniotic fluid samples, metaphases, particularly good quality metaphases, are much rarer, and the balance between false positives and false negatives is more delicate. The figures indicate a rough equivalence of the two error rates, but say nothing about the quality of the metaphases involved, i.e., how many of the false negatives were metaphases useless for analysis.

Metaphase Quality

The question of metaphase quality is a very difficult one to address in such a test, but is of central importance. Experience of operating this metaphase finder as part of a routine analysis regime[9] shows that ability to rank metaphases according to their quality and to provide the operator with good quality cells for analysis is the most important factor in achieving efficient throughput. Metaphases missed totally do not matter, provided that not many of them are analyzable, and nonmetaphases in the metaphase list are only of consequence if they are ranked as good metaphases and presented to the operator for analysis. The ranking figure in Table I is an attempt to quantify this property of the metaphase finder and is the number of relocations necessary to obtain five cells of a sufficient quality to produce a karyogram. This figure was not calculated for amniotic fluid samples in the study; however, figures for amniotic fluid samples are available from routine analysis in Copenhagen. A full analysis of a slide consists of ten counts and four karyograms from ten cells. Clinical experience has been that, on average, 39 potential metaphases have to be relocated from the list to achieve this. In order to reduce this figure further, the metaphase finder has been running in Copenhagen at higher magnification (256× rather than 160×), by using a higher secondary magnification and altered parameters for figure-of-merit calculation, giving an average of about 20 potential metaphases relocated to complete the analysis.[7,9] That is, on average, every other cell found is analyzable. The increase in scanning time (which is approximately quadrupled) is well compensated for, since overnight scanning imposes no penalty on throughput.

It is worth digressing for a moment to point out why ranking efficiency is a more critical requirement

of a metaphase finder than either speed or overall error rates. The effect on clinical throughput is determined by the amount of time a cytotechnician spends operating the system. The speed of unsupervised scanning is of little importance. Error rates are only of significance insofar as they affect the number of analyzable-quality metaphases presented to the operator. The number of objects that require relocation to complete an analysis is a sensitive measure of the efficiency of the metaphase finder's contribution to the analysis. A cytotechnician operating the metaphase analyzer[3] performs much more efficiently if presented with a succession of high-quality metaphases than if a number of unsuitable candidates must be viewed between analyzable cells. This is the case despite the fact that rejecting an unsuitable cell requires only a few seconds' attention. There seem to be two reasons for the high correlation of operator efficiency and ranking efficiency. One is psychological: the operator simply gets bored with seeing unanalyzable metaphases. The other is practical: the operator will try to make the analysis from the first ten usable cells that can be found. If these are not the best cells in the queue, due to poor ranking, the analysis will necessarily take longer.

Comparison with Other Metaphase Finders

Previous implementations of metaphase finders[5,6,12] have emphasized the importance of speed and low error rates (particularly false negatives). They have operated by applying gray-level thresholding to each field and identifying clusters of above-threshold objects. Depending on the hardware, this is accomplished either by identifying the frequency of threshold crossings along a line and associating results from successive lines[5] or applying binary open-and-close operations to the whole segmented image.[12] We have found global thresholding to be insufficiently robust for dealing with samples from routine stock, due to staining variability within and between samples. This is particularly true in the clinical samples since G-banded staining is almost always used. The method of using texture cues was adopted to make detection more robust.

Published data on metaphase finder performance are difficult to find. The only systematic attempt to make a comparison of metaphase finder performance to date has been that by Bresser,[1] from which Table I was extracted. This metaphase finder was compared with two other TV-based systems. Results in terms of scanning time and absolute error rates were broadly similar over blood and amniotic fluid samples for all three. The study did not address the difficult topics of robustness and quality assessment. No comparison was made with a metaphase finder using special line-scanning hardware, although it is obvious that this strategy would result in much faster scanning.

For example, a metaphase finder using hardware highly optimized for continuous scanning[10] has been tested by Finnon et al.[2] Their assessment was aimed at measuring its suitability for aberration scoring on unbanded blood cells. Scanning times were very low (about two minutes per slide), and detection efficiency was high. Some quality assessment was applied: scorable metaphases tended to be ranked higher than nonscorable ones or debris. It is difficult to extrapolate from these results to the requirements of a clinical karyotyping regime in which only a small number of very-well-formed metaphases is required, mainly from amniotic fluid samples. At the scanning speeds reported, the entire annual workload at Rigshospitalet could be scanned in a few days. Such high performance is probably unnecessary for clinical purposes.

The metaphase finder described in this paper now contributes substantially to the routine workload of the clinical chromosome laboratory at Rigshospitalet. In the first nine months of 1985, it was used together with automated analysis for 854 prenatal samples: slightly more than one-third of the prenatal workload in that period.[7] It has also been used to find cells for other nonautomated analyses, such as screening for fragile X chromosomes. We have no experience of applying this metaphase finder in other fields, such as cancer cytogenetics or aberration scoring. The ability to tune the parameters using the driving file should allow it to be used in other fields unless very fast scanning is seen to be an essential property.

Acknowledgments

We are indebted to the staff of the Chromosome Laboratory, Rigshospitalet, Copenhagen, particularly Claes Lundsteen, Tommy Gerdes, Jan Maahr and Anne-Marie Lind, for many discussions and for patience during the installation of the system. Many colleagues here in Manchester, particularly Chris Taylor, have contributed much in the form of ideas and encouragement. Joyce-Loebl Ltd., in addition to showing commercial interest and testing the software to destruction, provided most of the microscope control and autofocus software.

References

1. Bresser M: Automated metaphase finding. Report on European Working Group meeting on Automated Chromosome Analysis, Copenhagen, Denmark, 1984, pp 9–11
2. Finnon P, Lloyd DC, Edwards AA: An assessment of the metaphase finding capability of the Cytoscan 110. Mutation Res 164:101-108, 1986
3. Graham J: Automation of routine clinical chromosome analysis: I. Karyotyping by machine. Analyt Quant Cytol Histol 9: 383-390, 1987
4. Graham J, Taylor CJ, Cooper DH, Dixon RN: A compact set of image processing primitives and their role in a successful application program. Pattern Recogn Letters 4:325-333, 1986
5. Green DK, Bayley R, Rutovitz D: A cytogeneticist's microscope. Microsc Acta 29:237-245, 1977
6. Johnson ET, Goforth LJ: Metaphase spread detection and focus using closed circuit television. J Histochem Cytochem 22: 536-543, 1974
7. Lundsteen C, Gerdes T, Maahr J: Semiautomated prenatal chromosome analysis. Report to the European Chromosome Analysis Workshop, Leiden, the Netherlands, 1985
8. Lundsteen C, Gerdes T, Philip J, Graham J, Pycock D: An interactive system for chromosome analysis: Tests of clinical performance. *In* Proc III Scandinavian Conference on Image Analysis, Copenhagen, 1983, pp 392–397
9. Philip J, Lundsteen C: Semiautomated chromosome analysis: A clinical test. Clin Cytogenet 27:140-146, 1985
10. Shippey GA, Bayley RJH, Farrow ASJ, Rutovitz DR, Tucker JH: A fast interval processor. Pattern Recogn 14:345-356, 1981
11. Taylor CJ, Dixon RN, Gregory PJ, Graham J: An architecture for integrating symbolic and numerical image processing. *In* Intermediate Level Image Processing. Edited by MJB Duff. London, Academic Press, 1986, pp 19–34
12. Vrolijk J, ten Brinke H, Ploem S, Pearson PL: Video techniques applied to chromosome analysis. Microsc Acta [suppl] 4:108-115, 1980

9. **The transportation algorithm as an aid to chromosome classification.**
MKS Tso and J. Graham, *Patt. Recog. Lett., 1: 489 - 496, 1983.*
doi:10.1016/0167-8655(83)90091-0

Pattern Recognition Letters 1 (1983) 489–496
North-Holland

July 1983

The transportation algorithm as an aid to chromosome classification

M.K.S. TSO
Department of Mathematics, UMIST, P.O. Box 88, Manchester, M60 9PT, U.K.

J. GRAHAM
Department of Medical Biophysics, Stopford Building, The University of Manchester, Oxford Road, Manchester, M13 9PT, U.K.

Received 7 April 1983

Abstract: An algorithm is presented which obtains a constrained maximum likelihood classification of homogeneously stained chromosomes. Significantly improved results over both a context-free and a plausible context-driven classification are obtained. Extension to banded chromosomes and abnormal cells are discussed.

Key words: Classification, maximum likelihood, automated chromosome analysis, transportation algorithm, linear programming.

1. Introduction

Much of the effort in the analysis of chromosomes is directed towards producing a karyogram in which the 46 chromosomes of a (normal) human cell are displayed in their correct groups. A trained human operator can achieve this classification with an extremely low error rate (Lundsteen et al. (1976)). However it is a painstaking task and there is still considerable interest in developing an automated system having a comparable level of accuracy (see Piper et al. (1980)).

Rutovitz (1977) and Piper et al. (1980) have observed that the task of identifying the correct grouping of chromosomes is a 'context-conditioned' operation when performed by a human operator. That is, all the chromosomes in a cell are taken into account by the operator in making individual assignments. In particular the operator knows at the outset how many chromosomes in total should be assigned to each group. Previous attempts at automatic classification have tended therefore to include a final rearrangement algorithm that takes a 'context-free' allocation and iterates to one satisfying the group total constraint. Rutovitz (1977) has described such an algorithm that operates through a sequence of 'cascade' transfers between groups but no analysis of the optimality of this algorithm is given.

We propose in this paper a new formulation of the allocation problem which adopts a maximum-likelihood approach, but which allows an optimal allocation to be determined by an algorithm for solving the 'transportation' problem of linear programming also known as the 'Hitchcock' problem (Hitchcock (1941)). We give the results of a study involving 110 homogeneously stained human metaphase cells showing that significantly improved classifications can be expected using this algorithm.

2. Mathematical formulation

We have divided the chromosomes of a normal cell into ten groups, approximating the 'Denver'

0167-8655/83/$3.00 © 1983, Elsevier Science Publishers B.V. (North-Holland)

classification (Denver conference (1960)). These classes and the number of chromosomes they contain are shown in Table 1. The ambiguity in the totals for groups 5 and 10 arises from the fact that the sex chromosomes, the X and the Y, respectively, belong in these groups. We have given the correct totals for a male (XY) cell in parentheses.

We shall assume for the moment that the sex of the cell is known and that a complete set of chromosomes is present. Let I and J denote the index sets $I=\{1,\ldots,10\}$ and $J=\{1,\ldots,46\}$. The chromosomes to be classified are arbitrarily ordered and p measurements are made on each by an image analysis system. The resulting set of 46 p-vectors ξ_j, $j\in J$, belong to a p-dimensional *feature* space. We shall assume that chromosomes from the i-th group have a probability density function $f_i(\xi)$ defined over the feature space. The maximum-likelihood discriminant rule (see e.g. Mardia (1979), p. 300) allocates the j-th chromosome to the group k satisfying

$$L_K(\xi_j)=\max_i\{L_i(\xi_j)\}$$

where $L_i(\xi)$ is the likelihood function ($L_i(\xi)\propto f_i(\xi)$). However, this results in a context *independent* classification of individual chromosomes. We require a context dependent classification that provides the correct group totals. We may formulate this latter problem as the constrained optimisation

$$\text{Maximise } \log L=\sum_{\substack{i\in I\\ j\in J}}\log L_i(\xi_j)x_{ij} \qquad (1)$$

where $x_{ij}=0$ or 1 and satisfy the constraints

$$\sum_{j\in J}x_{ij}=n_i,\quad i\in I, \qquad (2)$$

and

$$\sum_{i\in I}x_{ij}=1,\quad j\in J. \qquad (3)$$

$X=(x_{ij})$ is an 'allocation matrix' of indicator variables such that $x_{ij}=1$ if chromosome j is allocated to group i. The constraints (3) specify that X is a valid allocation matrix with a single '1' in each column, while the constraints (2) specify the total number of chromosomes allocated to each group. The maximand in (1) is the log of the joint likelihood function of the total allocation assuming independence of the distributions $f_i(\xi)$, $i\in I$.

Table 1
The ten classification groups, their expected populations and their relation to the Denver classification

Denver Notation	Class (i)	n_i
1	1	2
2	2	2
3	3	2
4–5	4	4
6–12, X	5	16(15)
13–15	6	6
16	7	2
17–18	8	4
19–20	9	4
20–21, Y	10	4(5)

It is not difficult to see that this problem in 0–1 variables can be solved by the transportation algorithm. The transportation analogy arises as follows. Regard the assignment of a chromosome j to a group i as a movement along a route j–i giving rise to an additive cost c_{ij}. If we define

$$c_{ij}=-2\log L_i(\xi_j)$$

then $c_{ij}\geq 0$ and the problem of finding an optimal assignment maximising (1) becomes that of minimising the transportation cost function

$$\sum_{i,j}c_{ij}x_{ij}$$

finishing with a prespecified total number of chromosomes in each group. If we introduce the non-negativity constraints $x_{ij}\geq 0$, $i\in I$, $j\in J$, the 0–1 constraints on x_{ij} become superfluous. (Since the right-hand sides of (2) and (3) are integers, any optimal solution to the transportation problem produced by the algorithm must be integer (Trustrum (1971), p. 36). Hence (3) and the non-negativity conditions force x_{ij} to be 0 or 1, $i\in I$, $j\in J$, in the optimal solution).

This formulation possesses a number of striking advantages, namely the following:

(1) *Optimality:* A solution algorithm is available which is known to terminate at an optimal solution. The solution reached will be globally optimal, thus providing the overall maximum-likelihood allocation subject to the group total constraints.

(2) *Sexual attraction:* If the sex of a cell is not known in advance it may be left unspecified, and determined by the algorithm. In this case we set the group totals at their maximum value viz. $n_5 = 16$, $n_{10} = 5$, and take up the 'slack' in the allocations by creating a 'fictitious' 47th column for the X-matrix and forcing either $X_{5,47} = 1$ or $x_{10,47} = 1$. We ensure that these variables contribute nothing to the cost function by defining their associated cost coefficients to be zero. Forcing the slack allocation into one of groups 5 and 10 is achieved by assigning a large negative value to the coefficients of $x_{i,47}$ $(i \neq 5, 10)$ in (1). The algorithm will then be attracted to the 'most likely' choice of sex.

(3) *Missing chromosomes:* Whilst in the study presented in Section 3 we have used cells with the correct chromosome complement, this is not a requirement of the algorithm. If one or more chromosomes are missing, the deficient groups may be readily identified. The final column(s) of the X-matrix are regarded as fictitious 'slack' columns in much the same spirit as (2) above. The corresponding cost coefficients in the objective function are set to zero and the final allocation of these columns enables the deficient groups to be identified.

(4) *Extra chromosomes:* Extra chromosomes are assigned to a 'surplus' 11th group by the algorithm, the complement of this group being determined by counting. Chromosomes assigned to this group may subsequently be identified by context independent assignment to one of the 10 groups using a maximum likelihood rule.

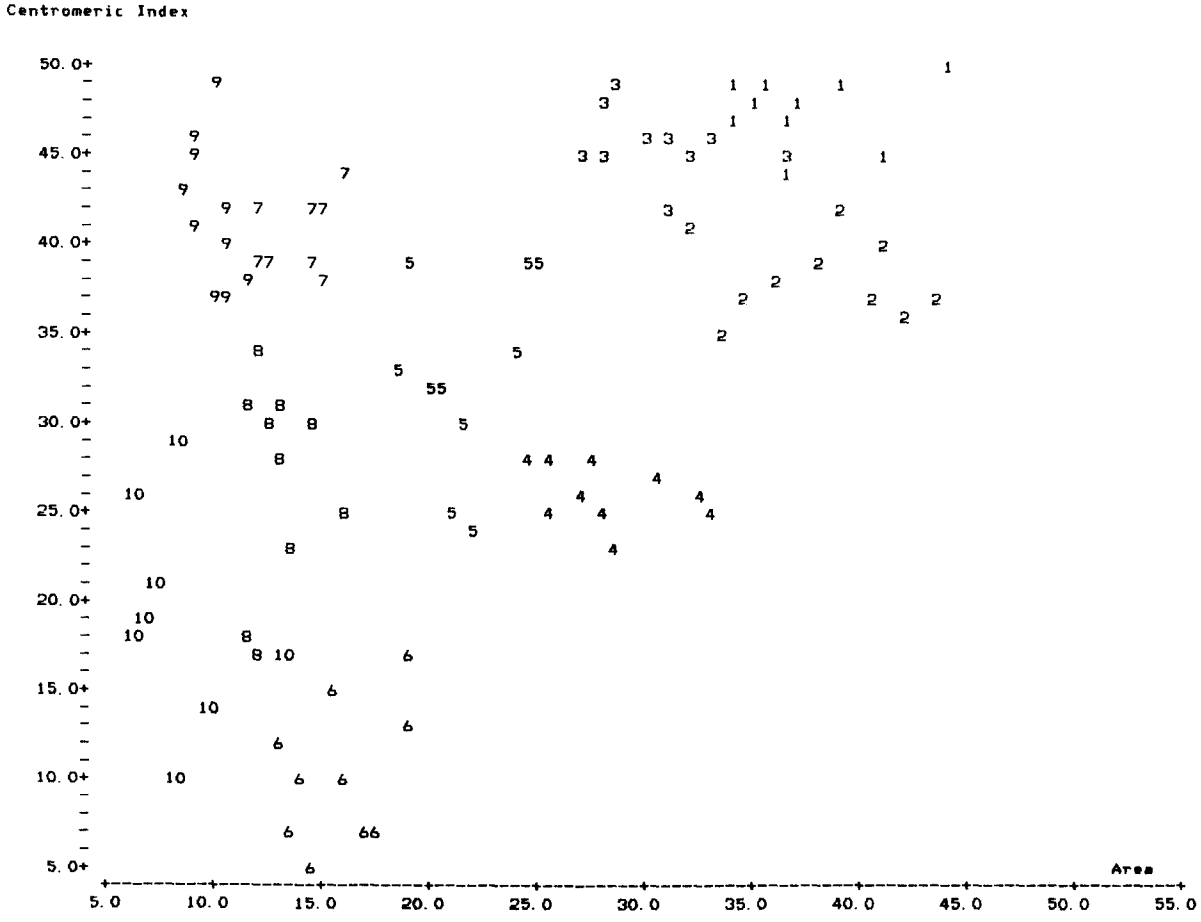


Fig. 1. Scatter of measurements from the 'training' set in feature space showing the true classification.

3. Results of analysis on 110 cells

One hundred and ten homogeneously stained metaphase cells from peripheral blood were selected from a routine clinical population. Roughly equal numbers of male and female cells were used in varying states of contraction and stain density and each containing the full complement of 46 chromosomes. Chromosome lengths, areas and the corresponding centromeric indices were measured using a television based image analyser (Magiscan-2, Joyce Loebl), running specially designed chromosome analysis software (JG, in preparation). Length and area measurements are highly correlated, and for this study, the classification features used were area and area centromeric index only.

Each cell was independently classified by an experienced human operator. It was decided to fit a 4-parameter bivariate normal distribution to the measurements from each chromosome group. (The errors in the determination of area and of area centromeric index were assumed to be uncorrelated.) Chromosomes from 20 cells were used to fit the distributions. These were divided into the ten groups and gave rise to the parameter estimates given in Table 2. Each determination was based on a minimum sample size of 40 chromosomes.

Let μ_i be the mean vector of measurements on area and centromeric index for chromosomes of the i-th group. Let Δ_i be the covariance matrix of these measurements (which we shall assume to be diagonal). The assumption of a normal distribution gives rise to the cost penalty

$$c_{ij} = \log|\Delta_i| + (\xi_j - \mu_i)^{\mathrm{T}} \Delta_i^{-1} (\xi_j - \mu_i) \qquad (4)$$

for assigning ξ_j to group i.

A total of 90 cells were classified by solving (1)–(3) using the transportation algorithm with cost matrix defined by (4). The results are shown in Table 3. The first three columns of this table represent respectively the 'one at a time' context-independent classification, the results obtained using a 'benchmark' allocation procedure and those obtained using the transportation algorithm.

The benchmark allocation procedure was to assign chromosomes to groups sequentially starting with the highest likelihood assignment, but not allocating chromosomes to full groups. Trying to assign a chromosome to a full group implies that chromosomes already assigned to that group have a higher likelihood of belonging there. The rejected chromosome must later be assigned to another (unfilled) group with lower likelihood. Thus a high likelihood solution is obtained, but global optimality is not guaranteed.

Use of the transportation algorithm resulted in an overall misclassification rate of 3.3% which is a clear improvement on the 5.2% achieved by the benchmark algorithm. The 'one at a time' approach achieved a rather higher misclassification rate of 6.4%. On a cell by cell basis the transportation algorithm was able to classify almost 50% of cells without error.

Table 2
Group parameters based on a 'training' set of 20 cells. A = Area normalised to mean 20, CI = Centromeric index by area

	A		CI	
Group	mean	s.d.	mean	s.d.
1	35.3	2.5	47.3	1.9
2	34.4	2.5	40.1	3.0
3	28.7	2.2	47.0	2.0
4	26.3	1.7	28.7	3.4
5	21.5	2.3	36.6	5.0
6	16.8	1.4	18.1	5.8
7	15.5	1.6	42.4	4.1
8	14.5	1.2	32.3	5.9
9	12.3	1.1	44.1	4.2
10	10.5	1.6	26.5	7.3

Table 3
Comparison of allocation procedures on an independent set of 90 cells

	Allocation			
	One at a time	Bench-mark	Trans-portation	True
No. of chromosomes misallocated	265	214	135	—
% of total	6.4	5.2	3.3	—
% of cells correctly allocated	10	33	46	—
Mean value of $-2\log L_{\max}$	0	24.7	6.7	12.5

The final row in Table 3 further confirms the improvement in classification accuracy obtained by the algorithm. This row contains the mean values of the objective function, $-2 \log L$, for the final classification determined by each algorithm. This quantity was also calculated for the 'true' classification of each cell determined by human operator. (These values have been shifted by subtracting the column minimum from each column of the cost matrix, so that $-2 \log L = 0$ corresponds to the overall 'one at a time' optimal allocation. Since this assignment is unconstrained, it represents the absolute maximum global likelihood. The difference between the calculated cost function and zero is a measure of how much likelihood we are 'losing' in conforming to the constraints.)

The distribution of the number of misclassifications per cell can be seen in Figure 2. These histograms confirm what is to be expected, that whilst the distribution has a recognisably smooth form using the 'one at a time' approach, it becomes quite distinctive if group total constraints are employed. For example, the possibility of a single misallocated chromosome in a cell only arises because the sex of the cell is initially unknown. In fact the sex was determined correctly by the algorithm in all but four cases and therefore a single misallocation had only a remote chance of occurrence. Similarly it is less usual to observe three misallocations than four. We are currently studying the nature of these distributions.

Further insight into the pattern of misclassifications can be obtained by examining the misclassification matrix shown in Table 4. A similar table was produced by Paton (1969) in early work embodying a maximum-likelihood approach. However the results given there apply to context independent classification on a rather smaller set of carefully selected chromosomes. Table 4 shows for example that the pairs of groups most likely to be confused with each other are groups 4 and 5, and groups 7 and 8 (this latter result is not unexpected as our groups 7 and 8 correspond to a single group in the 'Denver' classification). A more detailed analysis of the 'lack of fit' of our model can be made on the basis of this table but will not be attempted here.

Table 4
Misclassification matrix for 90 cells. Diagonal elements, corresponding to correct allocations, have been omitted for clarity

Assigned Group	True Group									
	1	2	3	4	5	6	7	8	9	10
1	*	2	7	·	·	·	·	·	·	·
2	3	*	·	·	·	·	·	·	·	·
3	6	1	*	·	·	1	·	·	·	·
4	·	·	·	*	16	·	·	·	·	·
5	·	·	1	16	*	2	2	·	·	·
6	·	·	·	·	·	*	·	1	·	3
7	·	·	·	·	·	·	*	16	7	·
8	·	·	·	·	·	1	12	*	7	6
9	·	·	·	·	·	·	9	5	*	3
10	·	·	·	·	·	1	·	4	3	*
Totals	180	180	180	360	1395	540	180	360	360	405

4. Mathematical details

The purpose of our study at this stage was to confirm that formulation as a transportation problem would lead to improved cell classifications. For this purpose an algorithm was developed for the general transportation problem. This used the predecessor index method of Glover and Klingman (1970) to locate the so called θ-circuit through the tableau. The programs were written in FORTRAN and developed on SERC's Prime 750 at UMIST. The initial allocation provided by the 'benchmark' algorithm was used as an initial basic feasible solution for the transportation algorithm, which in fact corresponds to using the matrix minimum method of initialisation mentioned in Trustrum (1971, p. 38). As a transportation problem, the classification problem is degenerate and the ε-method (see e.g. Trustrum (1971, p. 40)) was employed to forestall possible problems arising through the degeneracy.

The assumption that chromosomes from a single group are distributed normally in the feature space is made merely for convenience. One has only to observe that centrometric index is constrained to lie between 0 and 0.5 (prior to scaling) to see the weakness of such an assumption, yet the classifications achieved based on a normal premise were remarkably good in practice. It is quite possible, however, that an improved classification can be obtained using empirical density functions en-

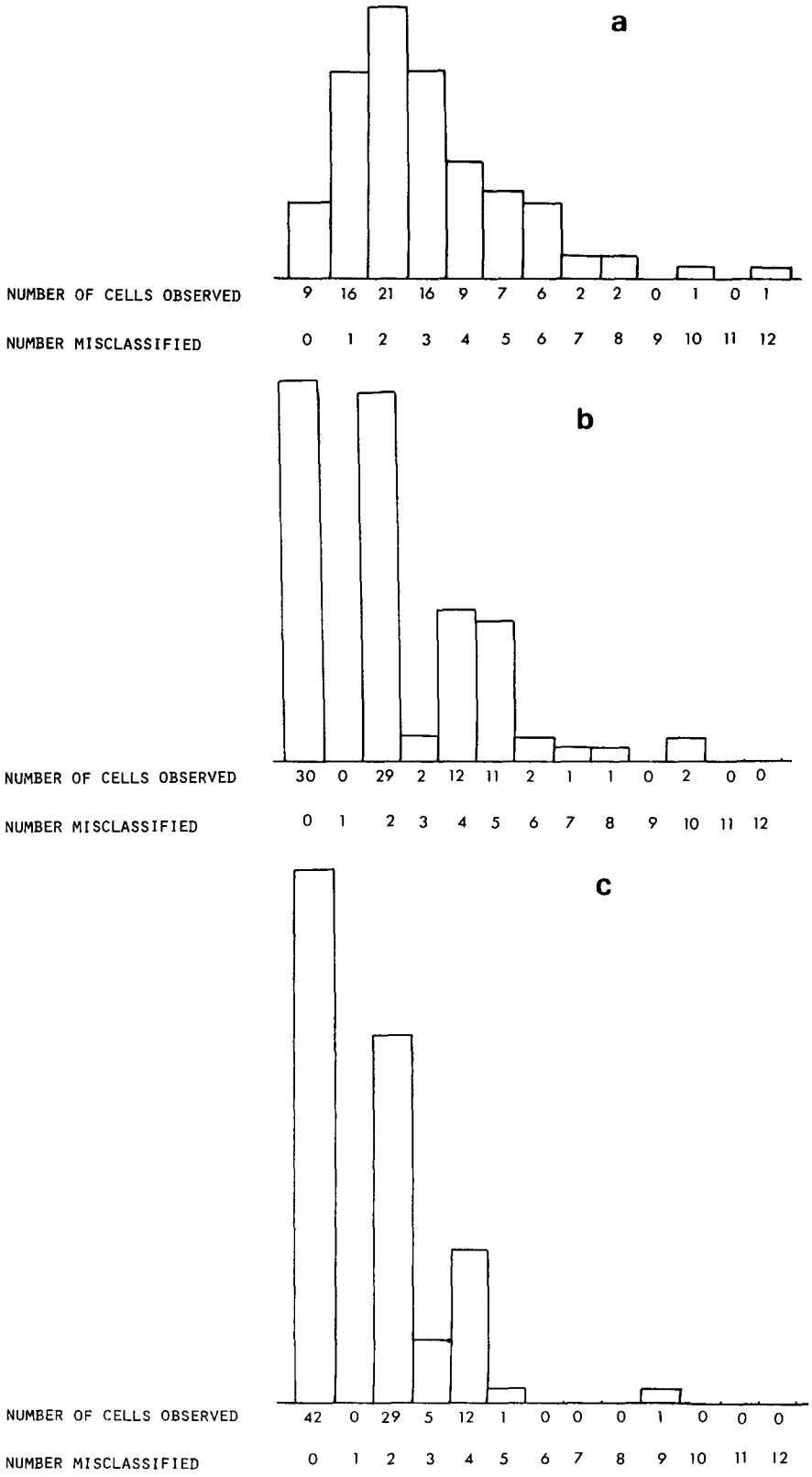


Fig. 2. Histograms showing the number of misclassified chromosomes by (a) context-free classification, (b) 'benchmark' context-driven assignment, (c) transportation algorithm.

coded as a look up table. These may be obtained by a method described by Paton (1969) and attributed to Lejeune and Turpin (1965, p. 61) whereby a Gaussian point spread function is applied to each chromosome of the training set.

5. Discussion

From our results it is apparent that good classifications can be obtained by using the formulation (1)–(3) in conjunction with the cost matrix defined by (4). We now outline possible directions for further work.

(1) It is of interest to know whether a more efficient procedure can be developed that explicitly takes into account the 0–1 nature of our problem. In the extreme case of our formulation when each group contains a single member ($n_i = 1$; $i \in I$) the problem can be solved by a procedure known as the assignment algorithm (see e.g. Spivey and Thrall (1970)). Piper et al. (1980), in a thorough discussion of factors influencing the economic viability of algorithms for automatic chromosome classification, have noted that speed is as important a factor as accuracy of classification in determining overall system cost, since a fast algorithm allows the possibility of reducing the error rate of classifications by karyotyping a number of cells.

(2) The distribution of chromosomes in the feature space, as we have noted in Section 4, is an aspect requiring further study. One question is whether the use of empirical probability density functions can result in improved classifications. Another question is whether better discrimination might be achieved by including more chromosome measurements, thus creating a higher chromosome feature space. Both these questions can be examined through the misclassification error rate and the misclassification matrix. Furthermore the mean difference between the value of $-2 \log L$ for the 'true' classification and the theoretical minimum provides some indication of the degree of information loss in passing from the image to the feature space.

(3) Although this study has been concerned with homogeneously stained chromosomes, we should emphasize that the techniques apply equally to banded chromosomes; these can be classified into 24 groups each with at most two members. In this case we are closer to the conditions required for application of the 'assignment algorithm' mentioned above. Granum (1981) has observed dramatic improvements in the error rate for classifying banded chromosomes in going from a 'context-free' classification to a constrained classification using an algorithm similar to the 'benchmark' assignment described here.

(4) We have given the procedure for identifying the groups relating to missing or extra chromosomes. This is a matter of great practical importance as certain abnormalities are characterised by either a missing or an extra chromosome in a certain group. Practical trials are required to establish the effectiveness of the procedures given in coping with cell abnormalities.

6. Conclusions

Using the method of maximum likelihood, we have formulated the chromosome classification problem as a constrained optimisation. We have shown that the optimal maximum-likelihood allocation can be determined by the transportation algorithm. Using this algorithm we have classified 90 human cells into the Denver classification achieving an overall misclassification rate of 3.3%.

If we regard the optimal allocation as in some sense making the best use of the available information in the feature space of measurements on chromosomes, the differences in log likelihood between the optimal and the 'true' allocations provides a measure of the loss of information in passing from the image to the feature space. The validity of this measure is due to the guarantee of global optimality which the transportation algorithm provides. This measure may be used to investigate the discriminatory power of different parametrisations of the feature space.

We have indicated the need for further mathematical investigation of the distribution within the feature space and indeed the definition of the feature space itself. Further examination of the iterative algorithm is likely to yield valuable gains in efficiency and hence improve the cost effective-

ness of any system for automated chromosome analysis operating in a 'production' environment.

There is every promise that our approach will be directly applicable to the classification of banded chromosomes. On the clinical side, tests will be necessary to examine the ability of the algorithm to cope with abnormal cells.

Acknowledgement

The authors are grateful to Dr. Claes Lundsteen of the Chromosome Laboratory, Rigshospital, Copenhagen, for expert assistance in collecting the data used in this study.

References

Denver Conference (1960). A proposed standard system of nomenclature of human mitotic chromosomes. *Lancet* 1, 1063–1065.

Glover, F. and D. Klingman (1970). Locating stepping-stone paths in distribution problems via the predecessor index method. *Transportation Science* 4, 220–225.

Granum, E. (1981). Application of statistical and synactical methods of analysis and classification to chromosome data. In: J. Kittler, K.S. Fu and L.F. Pau, eds. *Pattern Recognition Theory and Applications.* NATO Advanced Study Institute Series. Reidel, London.

Lejeune, J. and R. Turpin 61965). *Les Chromosomes Humains*, Gauthier Villars, Paris.

Lundsteen, C., A.-M. Lind and E. Granum (1976). Visual classification of banded human chromosomes I. Karyotyping compared with classification of isolated human chromosomes. *Ann. Hum. Gener.* 40, 87–97.

Paton, K. (1969). Automatic chromosome identification by the maximum-likelihood method. *Ann. Hum. Genet.* 33, 177–184.

Piper, J., E. Granum, D. Rutovitz and H. Ruttledge (1980). Automation of chromosome analysis. *Signal Processing* 2, 203–221.

Rutovitz, D. (1977). Chromosome classification and segmentation as exercises in knowing what to expect. In: E.W. Elock and D. Michie, eds. *Machine Intelligence* Vol. 8. Ellis Harwood, London.

Spivey, W.A. and R.M. Thrall (1970). *Linear Optimization.* Holt, Rinehart and Winston, New York.

Trustrum, K. (1971). *Linear Programming.* Routledge and Kegan Paul, London.

Mardia, K.V., J.T. Kent and J.M. Bibby (1979). *Multivariate Analysis.* Academic Press, London.

10. **An efficient transportation algorithm for automatic chromosome karyotyping.** M. Tso, P. Kleinschmidt, I. Mitterreiter and J. Graham, *Patt. Recog. Lett., 12: 117-126, 1991.* doi:10.1016/0167-8655(91)90057-S

Pattern Recognition Letters 12 (1991) 117-126
North-Holland

February 1991

An efficient transportation algorithm for automatic chromosome karyotyping

Michael Tso
Department of Mathematics, UMIST, P.O. Box 88, Manchester M60 1QD, UK

Peter Kleinschmidt and Ilse Mitterreiter
Fakultät für Mathematik und Informatik, Universität Passau, Innstrasse 33, W-8390 Passau, Germany

Jim Graham
Department of Medical Biophysics, The University of Manchester, Oxford Road, Manchester M13 9PT, UK

Received 27 July 1990

Abstract

Tso, M., P. Kleinschmidt, I. Mitterreiter and J. Graham, An efficient transportation algorithm for automatic chromosome karyotyping, Pattern Recognition Letters 12 (1991) 117-126.

We have implemented an algorithm for the special case of the transportation problem with unit demands to assist in the automatic classification (karyotyping) of human chromosomes by image analysis. Use of the algorithm permits prior knowledge of the number of chromosomes of a certain type in a normal human cell to constrain the classification. A study involving the classification of three large datasets is described and a comparison is made of the maximum likelihood and Bayesian approaches.

Keywords. Chromosome classification, karyotyping, transportation algorithm, image processing, context constrained classification, maximum likelihood, Bayesian methods.

1. Introduction

Karyotyping is the process by which chromosomes in a dividing cell, suitably stained are identified and allocated to one of a number of groups. This is an important clinical process, since the identification of abnormalities in chromosomes of particular groups may be diagnostic of certain clinical syndromes.

To form a *karyotype* of the 46 chromosomes in a normal human cell, they are stained to exhibit a series of bands along their length (Figure 1) when viewed under the microscope. This banding pattern, together with the size and shape of the chromosomes, is used to assign them into the 24 groups shown in Figure 1. In classifying the chromosomes in this way, the following information is used:

(1) Each group contains two identical (homologous) chromosomes, with the exception of groups X and Y which contain the sex chromosomes.

0167-8655/91/$03.50 © 1991 — Elsevier Science Publishers B.V. (North-Holland)

Figure 1. A typical 24 hour *Chorionic Villus* preparation karyotyped on the Image Recognition Systems Cytoscan.

(2) Female cells contain a homologous pair of X chromosomes while the male cells contain one X and one Y chromosome.

These constraints on the classification allow doubtful cases to be properly assigned and contribute to high classification accuracy. The measurement and classification of chromosomes can be performed automatically (Granum et al. (1989), Groen et al. (1989), Piper and Granum (1989)) and studies have shown that the error rate of automatic procedures can be substantially improved if these constraints are incorporated (Piper (1988)). An efficient method of incorporating these constraints into automatic procedures is the subject of this paper.

Rutovitz (1977) recognised that the error rate of a classification could be reduced by taking this contextual information into account and proposed a 'cascade' algorithm for satisfying the karyotyping constraints. Slot (1979) sought to develop statistical multiclass classification procedures to incorporate constraints on class totals. Tso and Graham (1983) showed that a linear programming formulation was applicable and used the transportation algorithm to obtain a globally optimal maximum-likelihood classification satisfying the constraints. They noted the requirement for an efficient transportation procedure that exploited the special structure of the problem and presented the results of a study showing improved classification of unbanded chromosomes into the 10 Denver groups. However, recent work in automated chromosome classification has focussed almost entirely on banded chromosome images which allows classification into 24 groups.

The study of algorithms for the transportation problem remains an active area of study in operational research and in the context of optimization on networks (Tso (1986)). In particular, Kleinschmidt and co-workers (1987) have recently proposed an efficient algorithm for the special case of the transportation problem with unit demands, which is the case applicable to karyotyping.

In this paper we present the results of a study on three large datasets of banded chromosome images using the algorithm proposed by Kleinschmidt et

al. (1987). We compare the improvement in classification accuracy gained using the transportation procedure with published results obtained using a suboptimal algorithm RC3 (Piper (1986)) which is a modified version of the 'cascade' algorithm proposed by Rutovitz (1977). In conclusion we discuss the formulation as a transportation problem of an alternative Bayesian approach.

2. Constrained classification — mathematical formulation

For ease of exposition we shall first consider the problem of karyotyping the 46 chromosomes in a female XX cell. Here the cell is known to contain 22 pairs of homologous chromosomes and the pair of sex chromosomes, making 23 pairs in total. We define $c = \{c_1, \ldots, c_{46}\}$ where $c_j = i$ iff chromosome j is assigned to group i to be a *classification vector* for the cell. The karyotyping constraint specifies that a valid classification vector for the 46 chromosomes in the cell should be a permutation of the vector

$$c_0 = (1, 1, 1, 2, 2, 3, 3, \ldots, 23, 23) \tag{1}$$

in which each of the pair group indices occurs twice. (Group 23 is taken to represent the X chromosome). We shall denote by K the class of such classifications satisfying the karyotyping constraints.

A permutation based approach was adopted by Slot (1979) in his study of karyotyping as a multiclass classification problem. An alternative viewpoint proposed by Tso and Graham (1983) is to regard such a classification as being represented by a (23×46) *assignment matrix* $X = (x_{ij})$ of 0's and 1's whose rows correspond to pair groups and whose columns correspond to the chromosomes. The entries of the matrix X define an assignment as follows:

$$x_{ij} = \begin{cases} 1, & \text{if chromosome } j \text{ is assigned} \\ & \text{to class } i, \\ 0, & \text{otherwise,} \end{cases}$$

and they satisfy the constraints

$$\sum_{i=1}^{23} x_{ij} = 1, \quad j = 1, \ldots, 46, \qquad \sum_{j=1}^{46} x_{ij} = 2, \quad i = 1, \ldots, 23, \tag{2}$$

representing respectively the conditions that each chromosome is assigned once only and that precisely two chromosomes are assigned to each group. This formulation is the basis of the linear programming transportation approach described in the next section.

When the sex of the cell to be karyotyped is initially unknown, as is usually the case in practice, the freedom to choose between an XX female cell and a XY male cell on the basis of the data can be incorporated by the following straightforward modifications to the model.

The classification vector c is augmented by a dummy entry c_{47} which is constrained to be either a group 23 (X-chromosome) or a group 24 (Y-chromosome). Allowable classifications are obtained by permutations of a new vector c_0 (modified by a final '24' in position 47) that constrain the last entry to be one of the sex groups. The corresponding changes to the X-matrix are the addition of an extra row corresponding to the Y-chromosome group 24, and an extra 47th column whose assignment is allowed only to groups 23 and 24. This amounts to defining the additional 0–1 variables $s_1 = x_{23,47}$ and $s_2 = x_{24,47}$ satisfying $s_1 + s_2 = 1$ and

$$\sum_{j=1}^{46} x_{23,j} + s_1 = 2, \qquad \sum_{j=1}^{46} x_{24,j} + s_2 = 1. \tag{3}$$

In linear programming terms s_1 and s_2 are known as slack variables. With (3) modifying (2), the set K of allowable classifications is now expanded to include both male and female karyotypes.

Occasionally it may happen that a cell is incomplete so that the number of chromosomes in the cell to be karyotyped is n where $n < 46$. This occurs when two or more chromosomes are overlapped in such a way that they cannot be reliably segmented by the image processing algorithm, or if a chromosome is missing due to an abnormality or a preparation artifact. In such cases only the first n

elements of c represent actual assignments, and the class K should be further restricted to rearrangements of c_0 that are distinct in the first n elements.

3. Maximum-likelihood and the transportation model

Maximum-likelihood (ML) provides a criterion for selecting an optimal classification c from K given a dataset of n ($\leqslant 46$) feature vectors $D = \{\xi_1, \ldots, \xi_n\}$ measured on the chromosomes images of an unknown cell. Suppose that $f_i(\xi)$ is the known probability density function of the feature vector ξ for a chromosome of class i. If ξ is measured on an unclassified chromosome, the likelihood of this chromosome belonging to group i is defined as $L(i \mid \xi) = f_i(\xi)$. Assuming independence of the distributions, the joint likelihood of an arbitrary classification c is given by

$$L(c \mid D) = \prod_{j=1}^{n} L(c_j \mid \xi_j). \tag{4}$$

The ML classification is obtained by maximizing (4) with respect to c. Equivalently, we may determine the assignment matrix X satisfying (2) modified by (3) to minimize

$$\begin{aligned} l &= -\log L(c \mid D) \\ &= \sum_{j=1}^{n} \sum_{i=1}^{24} [-\log L(i \mid \xi_j)] x_{ij} \\ &= \sum_{j=1}^{n} \sum_{i=1}^{24} \gamma_{ij} x_{ij} \qquad \text{say.} \end{aligned} \tag{5}$$

The optimization has been cast into the form of a linear transportation problem with cost matrix $\Gamma = (\gamma_{ij})$, the matrix of log likelihoods. This problem takes its name from the following supply-demand model which has been extensively studied in the context of operational research.

Suppose we have to transport a number of units of a commodity from a given set of m sources to n destinations. We are given a_i, the number of units available at source i, and b_j, the number of units required, or the *demand* at destination j. The cost of transporting a unit quantity from source i to destination j is γ_{ij} and this cost varies linearly as the number of units transported along this route.

The problem of determining a minimum cost shipment plan to satisfy all the demands is known as the transportation problem. Karyotyping can be viewed as a transportation problem in which each source is a chromosome class, and the chromosomes themselves are regarded as destinations having unit demand. If the costs $\{\gamma_{ij}\}$ are defined as in (5) to be log likelihoods, then the solution to the transportation problem is a maximum-likelihood classification.

4. The algorithm

The transportation problem (TP) is a type of linear programming problem for which, in the general case, algorithms are known that provide a globally optimal minimum cost solution. The special case of unit demands however means that more efficient procedures are available than the 'stepping-stone' procedure used in Tso and Graham (1983). The lack of a ready algorithm, on which for example timings could be based, gave rise to doubts about the practicality of the linear programming approach for routine clinical use in an interactive system (see Piper (1986, p. 392)).

Kleinschmidt et al. (1987) proved that the class of transportation problems with unit demands could be solved by an algorithm based on the pivot rule proposed by Balinski (1985). This algorithm was incorporated into a program PCCS (the Passau Chromosome Classification System) which obtained optimal classifications for the three large datasets described in Section 5. Our study, using a Sun workstation, confirmed timings well within acceptable limits for the method to be usefully incorporated in practical karyotyping systems. The algorithm has been shown to solve similar sized problems in less than 1 sec. on a IBM-PC, which represents a negligible overhead. In general, the algorithm is known to have a time complexicity of $O(mn^2)$ for transportation problems with unit demands (Kleinschmidt et al. (1987))—m is the number of supply nodes or *chromosome classes* and n is the number of demand nodes or *chromosomes*. Thus it is known to be efficient for such problems. For karyotyping, we would expect the likelihood matrix to contain a large number of

Table 1
Summary of the datasets used in the study

	Copenhagen	Edinburgh	Philadelphia
No. of cells	180	125	130
No. of chromosomes	8106	5548	5947
Mean no. of chromosomes per cell	45.0	44.4	45.7

zeroes since the features will have been selected to give good discrimination between classes, so restricting the number of possible classifications for each chromosome. Thus the resulting cost matrix will be *sparse* in the sense that it will contain a proportion of effectively infinite entries corresponding to forbidden assignments. For sparse problems with degree of sparsity κ—defined as the number of non-excludable assignments—the time complexity of the computation is known to be proportional to $n(\kappa + m \log m)$ suggesting room for some slight performance improvement if sparsity is taken explicitly into account.

5. A study on three datasets

We have evaluated the benefit of using PCCS by karyotyping cells in three large datasets using likelihood data supplied by the MRC Human Genetics Unit in Edinburgh (Table 1). The datasets are designated by their laboratory of origin (Copenhagen, Edinburgh or Philadelphia), and consist of cells which have each been classified by an experienced cytogeneticist.

Each dataset represents metaphases prepared under significantly different laboratory conditions. The Copenhagen dataset is in many ways a model dataset of carefully selected and measured cells which can provide a benchmark measure of performance for an algorithm operating under ideal conditions. The Edinburgh and Philadelphia datasets are more representative of the quality of cells to be found in a routine chromosome laboratory. The Copenhagen and Edinburgh datasets were imaged respectively by microdensitometer from film and by TV camera. These datasets came from peripheral blood cells. The Philadelphia dataset was collected using the commercially available Cytoscan system which employs a mechanically scanning linear CCD array. The cells in this dataset came from a *Chorionic Villus* preparation in which the appearance of the cells is known to be consistently poorer than for other clinical preparations. Further details of the datasets can be found in Piper and Granum (1989).

The features used in this study were weighted density distributions as proposed by Granum (1982). The cost matrix input to the transportation procedure was a matrix of log likelihoods obtained by fitting a zero-correlation multivariate normal distribution to each class in a 16-feature space. Consistency with the features employed in Piper and Granum (1989) ensured that a valid comparison could be made with previously reported results. Each dataset was divided into independent test and training subsets whose rôles were later reversed so that each cell was used once for training and once for classification. Each dataset was classified by three methods: (1) by context independent maximum likelihood, (2) by the transportation algorithm PCCS, and (3) by the 'cascade' algorithm RC3 (Piper (1986)).

6. Results of study

Table 2 shows a comparison of the error rates for each dataset expressed as a percentage of chromosomes incorrectly classified. Figures are means over the relevant dataset.

All three classifiers confirm the ranking of the datasets in terms of image quality. On the Edinburgh and Philadelphia datasets the improvement in error rate resulting from using either PCCS or the sub-optimal procedure RC3 is about 2%, confirming the earlier results of Piper and Granum (1989). On the Copenhagen dataset PCCS marked-

Table 2
A comparison of PCCS against other procedures—observed error rates expressed as percentages

		Copenhagen	Edinburgh	Philadelphia
Context independent ML Classification		6.5	18.3	22.8
Rearrangement classifier RC3		5.7	16.4	20.6
Transportation procedure PCCS		4.4	15.5	19.9
% Improvement over context independent classification	RC3	0.8 ± 0.2	1.9 ± 0.5	2.2 ± 0.7
	PCCS	2.1 ± 0.5	2.8 ± 0.9	2.9 ± 1.0
95% confidence interval for mean % improvement using PCCS in preference to RC3		(0.8, 1.7)	(0.1, 1.6)	(−0.2, 1.5)

ly outperforms RC3 giving an improvement over context independent classification of 2.1% compared to 0.8%.

Although the improvements we observed may appear to be numerically small, we do in fact have a useful reduction in the error rate. Taking the Copenhagen dataset for example, the recognition rate for context independent ML classification is 93.5%. An improvement of 2.1% in the recognition rate means that some 30% of the errors remaining are removed using PCCS. For the Edinburgh and Philadelphia datasets the corresponding improvements are 15% and 13% respectively. From Table 3 we see that the proportion of Copenhagen cells that are completely correctly classified has risen dramatically from 15.6% to 46.7%. In Section 7 we suggest that the reason for this lies in the choice of loss function implicitly assumed in the maximum likelihood approach.

The confidence intervals and the error bounds shown in Table 2 are all based on large sample normal approximations and a significance level $\alpha = 0.05$.

Table 3
Number of totally correctly classified cells

	ML	RC3	PCCS
Copenhagen	28 (15.6%)	41 (22.8%)	84 (46.7%)
Edinburgh	0	0	1 (0.8%)
Philadelphia	0	2 (1.5%)	5 (3.8%)

In Figure 2 we show histograms of the number of errors made per cell for each procedure. For ease of comparison, both the procedures PCCS and RC3 are displayed against the results of context independent classification in a single histogram. These histograms show the improvement in the number of cells which are totally correctly classified, which is particularly marked in the Copenhagen dataset. Also noticeable is a general shift to the left when the karyotyping constraints are applied, signifying a reduction in the mean error rate.

7. A Bayesian analysis

In this section we show that our maximum-likelihood analysis has a Bayesian interpretation in which the loss function is taken to be either 0 if the karyotype is correct, or 1 otherwise, i.e., equal weight is attached to an erroneous cell karyotype irrespective of the number of individual chromosomes incorrectly classified. Use of this loss function results in a classifier that will maximize the expected number of 100% correctly classified cells in any random sample of cells, thus helping to explain why our maximum-likelihood procedure seems to do so well by this criterion on the datasets used in this study.

We then examine the Bayesian loss function that attaches a weight proportional to the number of errors in the karyotype. This is shown to result in a classifier that minimizes the expected number of errors in the karyotype. We briefly consider implementation of this classifier, and show that, for

this classifier, a transportation procedure is still applicable if the cost matrix is suitably redefined.

In the Bayesian approach we regard the class K of possible classifications for a cell as the discrete set of possible states of nature $\{\tilde{c}\}$ each of which has equal prior probability of occurrence. We first consider the loss function

$$\Delta_0(c,\tilde{c}) = \begin{cases} 0, & \text{if } c = \tilde{c}, \\ 1, & \text{otherwise} \end{cases} \tag{6}$$

associated with a decision to classify a cell as c when the true state is $\tilde{c}$. The number of cells correctly classified out of a random sample of N can be expressed in terms of this loss function as

$$N(\text{correct}) = \sum_{k=1}^{N} [1 - \Delta_0(c_k, \tilde{c}_k)],$$

and by taking expectations we see that the Bayes classifier minimizing the expected loss will also maximize the expected number of correctly classified cells in a randomly sampled dataset. The expected loss for a decision c given data D, is

$$\begin{aligned} \mathbb{E}\Delta_0(c) &= \sum_{\tilde{c}\in K} \Delta_0(c,\tilde{c}) \Pr(\tilde{c} \mid D) \\ &= \sum_{\tilde{c}\neq c} \Pr(\tilde{c} \mid D) \\ &= 1 - \Pr(c \mid D). \end{aligned} \tag{7}$$

Noting that the probability $\Pr(c \mid D)$ is proportional to the likelihood $L(c \mid D)$, it follows that the Bayes classifier minimizing (7) is also the maximum-likelihood classifier. Hence the maximum-likelihood classifier maximizes the expected number of correctly classified cells.

A loss function which counts the number of chromosomes incorrectly classified was assumed by Slot (1979). The loss function can be written in terms of the components of c and $\tilde{c}$ as

$$\Delta_1(c,\tilde{c}) = \sum_{j=1}^{n} \delta(c_j, \tilde{c}_j) \tag{8}$$

where

$$\delta(a,b) = \begin{cases} 0, & \text{if } a = b, \\ 1, & \text{otherwise.} \end{cases}$$

The expected loss under this new loss function is

$$\begin{aligned} \mathbb{E}\Delta_1(c) &= \sum_{\tilde{c}\in K} \Delta_1(c,\tilde{c}) \Pr(\tilde{c} \mid D) \\ &= \sum_{\tilde{c}\in K} \sum_{j=1}^{n} \delta(c_j, \tilde{c}_j) \Pr(\tilde{c} \mid D). \end{aligned} \tag{9}$$

Inverting the order of the summation in (9), we obtain

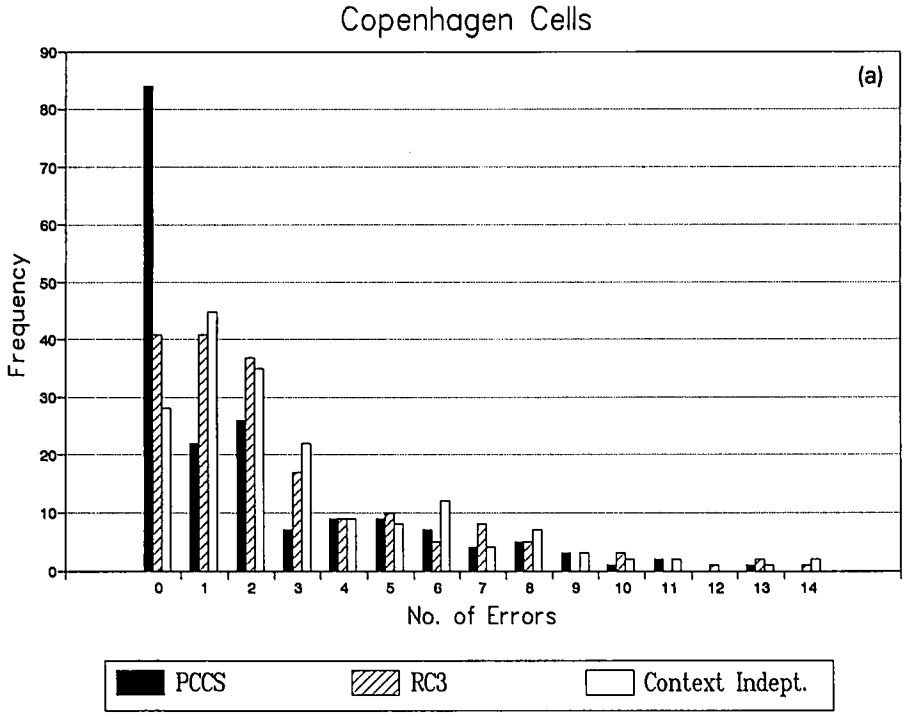
$$\begin{aligned} \mathbb{E}\Delta_1(c) &= \sum_{j=1}^{n} \sum_{\tilde{c}\in K} \delta(c_j, \tilde{c}_j) \Pr(\tilde{c} \mid D) \\ &= \sum_{j=1}^{n} \left[1 - \sum_{\tilde{c}_j = c_j} \Pr(\tilde{c} \mid D) \right]. \end{aligned} \tag{10}$$

The Bayesian classifier obtained by minimizing (10) over all classifications $c \in K$ minimizes the mean error rate, which is an intuitively reasonable measure of classifier performance. Slot (1979) proposes that the components of c should be chosen independently to minimize each component of the outer summation. However the resulting classification will not satisfy the karyotyping constraint that $c \in K$ unless a transportation procedure is applied. This requires that the cost matrix be computed whose elements are the inner summations in (10). The number of terms in each sum is in principle of the order of the total number of states in K, i.e., $\sim 46!/2^{23}$. However, it is known that in practice one or two size related features narrow the range of possible classifications for any chromosome and we would therefore expect Γ, the matrix of log likelihoods, to be quite sparse. If this is the case, then many of the terms in each sum will vanish and use of the transportation algorithm to compute this classifier may be feasible.

8. Discussion and conclusions

We have implemented an efficient transportation algorithm to perform a maximum-likelihood classification of a set of objects subject to constraints of the type occurring in the automatic karyotyping of human cell chromosomes. Use of the algorithm on three large datasets showed an improvement on previously published results.

The most noticeable improvement was observed on the high quality Copenhagen dataset where the



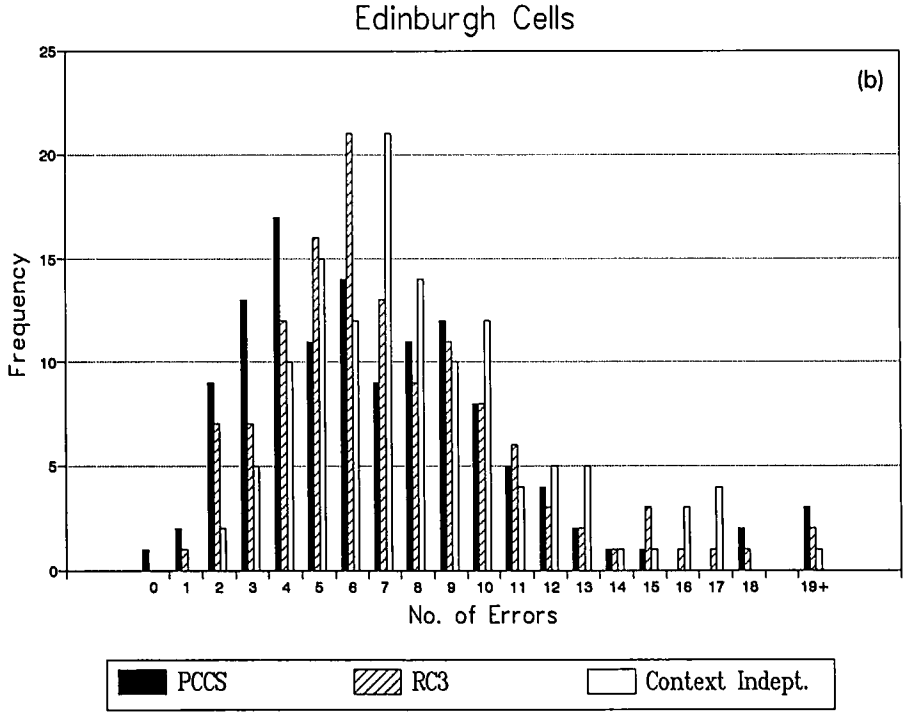


Figure 2. Distribution of number of errors made per cell.

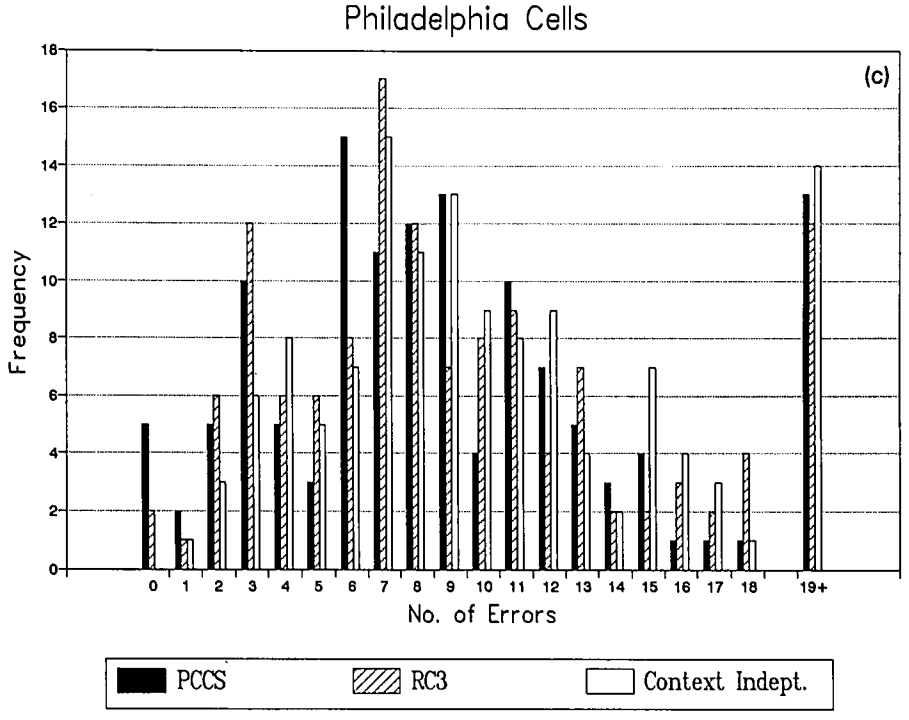


Figure 2 (contd.).

transportation procedure removed almost 30% of the errors remaining after context independent classification. The final recognition rate was over 95%—on average two misclassifications per cell. The percentage of cells that were completely correctly classified also rose from just below 20% to almost 50% for this dataset. Even on poorer quality cells a consistent improvement in the absolute error rate of some 2–3% was observed which removed some 10% of the errors remaining after context independent classification.

On poor quality cells, the modest improvement obtained is only to be expected since a process of balancing likelihoods will only work well when the likelihoods which have to be *estimated* from a training set are a good approximation to their *true* values, and this is more likely to be the case on well prepared samples where the inherent variability will be less. There is clearly no substitute for good data.

Through a Bayesian analysis, we have provided theoretical grounds for expecting that a transportation procedure aimed at finding a maximum-likelihood classification should obtain a high proportion of totally correct cells. This confirmed what was observed on the high quality Copenhagen data set. We have shown on the other hand that if the mean error count (number of misclassifications per cell) is to be minimized, then an alternative loss function should be adopted and that in consequence it may be possible to improve on maximum-likelihood classification if this is the criterion of interest. We have demonstrated that a transportation procedure is still appropriate for this problem with an appropriately redefined cost matrix.

We conclude that there are considerable advantages to be gained in viewing the problem of classification under class total constraints as a network optimization. For example, there is the possibility of applying sensitivity analysis to the results of classification, in order to identify assignments that are sensitive to perturbations of the likelihood matrix. There is also the possibility of applying algorithms for non-bipartite matching to extract homologous pairs, to develop an approach suggested by Zimmerman et al. (1986). It is hoped that the demonstrable usefulness of a

special purpose transportation algorithm for the particular pattern recognition task considered in this paper will provide an additional stimulus to the extensive research being carried out on such algorithms in an O.R. context.

Acknowledgements

Some of these results were presented at a Workshop in Besse-en-Chandesse, France in September 1989 held under the auspices of the EC Concerted Action on Automated Cytogenetics. We gratefully acknowledge the rôle played by this program in stimulating this collaborative work. We especially wish to record our thanks to Jim Piper of the MRC Human Genetics Unit in Edinburgh for generous assistance in providing the datasets on which this work is based.

References

Balinski, M.L. (1985). Signature methods for the assignment problem. *Operations Research* 33, 527–536.

Granum, E. (1982). Application of statistical and syntactical methods of analysis and classification to chromosome data. In: J. Kittler, K.S. Fu and L.F. Pau, Eds., *Pattern Recognition Theory and Applications* (Proc. NATO ASI Series). Reidel, Dordrecht.

Granum, E., M.G. Thomason and J. Gregor (1989). On the use of automatically inferred Markov networks for chromosome analysis. In: C. Lundsteen and J. Piper, Eds., *Automation of Cytogenetics*. Springer, Berlin.

Groen, C.A., T.K. ten Kate, A.W.M. Smeulders and I.T. Young (1989). Human chromosome classification based on local band descriptors. *Pattern Recognition Letters* 9, 211–222.

Kleinschmidt, P., C.W. Lee and H. Schannath (1987). Transportation problems which can be solved by the use of Hirsch-paths for the dual problem. *Mathematical Programming* 37, 153–168.

Piper, J. (1986). Classification of chromosomes constrained by expected class size. *Pattern Recognition Letters* 4, 391–395.

Piper, J. and E. Granum (1989). On fully automatic feature measurement for banded chromosome classification. *Cytometry* 10, 242–255.

Rutovitz, D. (1977). Chromosome classification and segmentation as exercises in knowing what to expect. In: E.W. Elcock and D. Michie, Eds., *Machine Intelligence 8*. Ellis Horwood, London, 455–472.

Slot, R.E. (1979). On the profit of taking into account the known number of objects per class in classification methods. *IEEE Trans. Inform. Theory* 25, 484–488.

Tso, M.K.S. and J. Graham (1983). The transportation algorithm as an aid to chromosome classification. *Pattern Recognition Letters* 1, 489–496.

Tso, M.K.S. (1986). Network flow models in image processing. *J. Oper. Res. Soc.* 37(1), 31–34.

Zimmerman, S.O., D.A. Johnston, F.E. Arrighi and M.E. Rupp (1986). Automated homologue matching of human G-banded chromosomes. *Comput. Biol. Med.* 16(3), 223–233.

11. **Resolution of composites in interactive karyotyping.** J. Graham, in *"Automation of Cytogenetics", C. Lundsteen and J. Piper (eds.), 1989, Springer-Verlag, Berlin, pp 191 - 203. doi:* 10.3233/978-1-60750-851-9-174

Resolution of Composites in Interactive Karyotyping

James Graham

Summary

The major requirement for operator interaction in automated karyotyping systems arises from the formation of "composite chromosomes" at the segmentation stage. These composites occur because chromosomes often touch closely or overlap so that they cannot be separated by simple, context free, segmentation methods such as thresholding, however carefully applied. This paper describes a method for automatically resolving the most commonly occurring type of composite, that due to touching chromosomes, by using simple contextual information about chromosome positions to drive a split and merge algorithm. A pilot trial shows that application of this method results in a significant reduction in the need for interaction in the analysis of metaphase cells of routine quality.

1. Introduction

Recently there has been an increase in interest in interactive karyotyping as a useful technology in the routine clinical chromosome laboratory. The system which I have described at previous workshops of the European Concerted Action in automated cytogenetics [1,2] has been available as a commercial product for some time. Other products with a similar functional specification are available from a number of manufacturers, both in Europe and the United States. Some of these products are packages of integrated metaphase finders and karyotypers. Some are karyotyping only systems. Lundsteen and Martin [3] have recently produced a useful review of the systems available and their various strengths and weaknesses.

Interactive karyotyping is the process of machine analysis of high resolution images of metaphase or pro-metaphase cells (usually from blood or amniotic fluid), supervised by a trained cytogeneticist or cytotechnician. The analysis consists of segmentation of individual chromosomes, measurements of size, banding pattern and centromere position, classification into pairs and, usually, production of a karyogram which can be reproduced on some hard copy device. At each stage in the analysis fairly standard pattern recognition techniques provide reasonably satisfactory results. However, in a system required to deal with the range of sample qualities obtainable routinely, and the occurrences of overlaps, twisted chromosomes, debris etc. that occur in eventhe best prepared

Automation of Cytogenetics Editors: C. Lundsteen J. Piper
© Springer-Verlag Berlin Heidelberg New York 1989

specimen, reasonably satisfactory results still leave a great deal of scope for error. Misclassifications occur, centromeres are incorrectly located and, most importantly for the purposes of this paper, chromosomes can be poorly segmented. All of these difficulties are resolved by allowing the skilled operator to intervene in the analysis and correct errors. From one point of view this is a very acceptable solution. It allows the skills of the operator to be employed efficiently in the analysis, but it requires that the interactions are straightforward and meaningful to an operator who has a deep understanding of the image, though not necessarily of the machine, and it also requires that the number of interactions needed should not be excessive.

Interactive karyotyping software may be implemented on moderately priced general purpose hardware, allowing a useful measure of automation to be developed and widely used in clinical laboratories which do not have large capital equipment budgets. The system, which is the subject of this paper [4] was developed, together with a metaphase finder [5] on a Magiscan image analyser from Joyce Loebl. The hardware architecture of this instrument has been fully described by Taylor et al [6] and its data structures and image processing software by Graham et al [7]. This integrated hardware and software architecture makes for straightforward and efficient implementation of the type of algorithm described here.

In a clinical assessment of this system, Philip and Lundsteen [8] found that interaction time is the most significant part of the analysis and that operator happiness decreases rapidly with the number of interactions. By far the largest number of interactions occurs in the separation of composites of touching chromosomes, exceeding, for example, those needed to separate overlaps or to correct centromere positions.

This paper addresses the problem of reducing the number of interactions required in the segmentation phase by identifying and resolving composites which occur where several touching chromosomes are segmented as one object.

2. The Image Segmentation Process

For the purposes of this paper it is necessary to describe the segmentation procedure and its limitations in some detail. The images consist of 512×512 pixels of 64 grey levels obtained by a TV camera mounted on a microscope. There are two phases to the segmentation - a "coarse" segmentation in which the positions of objects are determined and the chromosomes are counted, and a fine segmentation where boundaries are located as accurately as possible, forming the first stage in the detailed analysis of the cell. Both segmentation phases are achieved by binary slicing at thresholds obtained from grey level histograms.

At the coarse "counting" segmentation, the histogram is taken from the central quarter of the image. A global threshold is selected in order to underdetect objects. The exact boundary positions are not important at this stage;

it is more important to try to ensure that closely touching chromosomes are located separately. The result is an approximation to the chromosome count; an approximation, because in most cells not all touching chromosomes will have been separated by this simple procedure, and some single chromosomes may have been split along light bands. These errors are corrected by an interaction. The automatically counted objects are each marked with a dot and the count is displayed. The operator corrects the count by indicating undetected or incorrectly detected objects with a lightpen, causing either the removal of the object or the marking of a new object, with a corresponding adjustment of the displayed count. This method of counting is efficient in that it makes use of the strengths of both the machine and the operator. Marking an internal point in each chromosome not only provides visual feedback to the operator during interaction, but also provides useful data for the next segmentation phase.

Fine segmentation locates the boundary of each chromosome by using a threshold obtained from a histogram of the image in the region close to it. The localities of the regions to be histogrammed are determined from the chromosome positions found at the coarse segmentation phase. The boundaries obtained by local thresholding are usually very satisfactory from the point of view of separation of chromosomes from background. However, the region of background between closely abutting chromosomes is often darker than the best threshold, resulting in composite chromosomes at the fine segmentation stage (figure 1a). Notice that this is not just a thresholding problem. A threshold which will separate touching chromosomes will almost certainly not result in a good overall boundary (figure 1b). The interaction required to resolve such composites is straightforward; the object is selected by pointing at it with the lightpen and the correct boundary positions indicated by drawing a single line. These steps, and the rest of the automated karyotyping process, are described in detail elsewhere [4].

3. Resolving Composites

In the case of closely abutting chromosomes, there is usually a valley in the grey value landscape between the two significant objects. The composite is created because the grey value in the valley floor is higher than the selected threshold. The position of best separation of the chromosomes is at the grey level minimum along the valley floor. Finding this minimum is functionally identical to the process of finding "watersheds" in grey level landscapes as described by the proponents of mathematical morphology [9]. Many morphological operators are cellular logic operators, and most efficient if implemented on a machine where a number of pixels are processed in parallel. Friedlander and Meyer [10] have described an algorithm for finding watersheds by a propagation method, suitable for pixel-serial processing. This algorithm involves first finding positions of grey level minima as "seeds" and propagating the regions of these

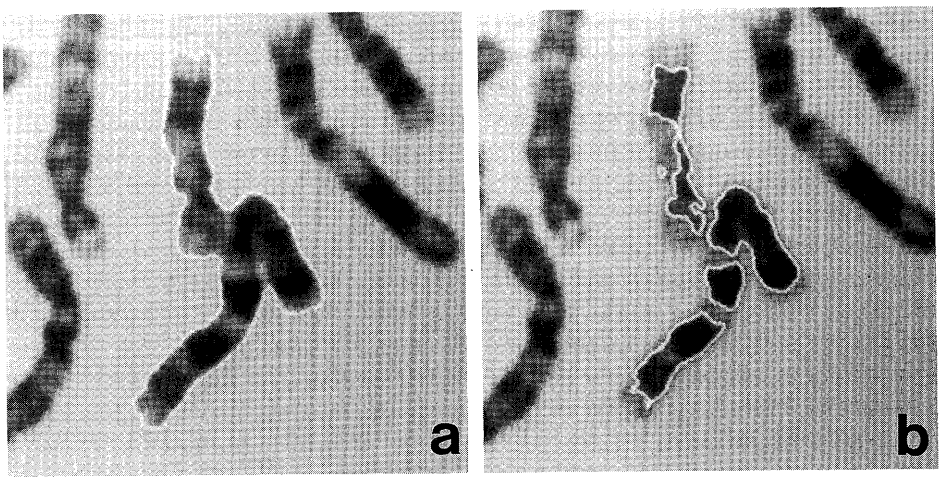


Fig. 1. A composite of two touching chromosomes. (a) The boundary obtained by local thresholding. (b) The boundaries obtained at a threshold which just separates the two chromosomes. Finding good boundaries for both chromosomes is not merely a thresholding problem.

seeds towards the watersheds. The second part of this algorithm is identical to a procedure described by Rutovitz in 1978 [11] for obtaining what he called (using the opposite topographical metaphor) the "fall-set" of some seed region; the fall-set is the set of points which can be reached by a downhill path from the seed region. The fall-sets of two neighbouring objects meet and terminate at the minimum of the grey level valley between them .

In the case of chromosomes, we wish to be careful about which minima we find, since the chromosome bands also give rise to grey level valleys, where we would certainly rather not split an object. Suitable seed regions are the regions obtained by binary slicing the image at a threshold which just separates the two neighbouring chromosomes (figure 1b). In many cases this will mean that internal minima will lie within the seed regions and will not be subjected to the fall-set finding procedures. There is, however, the problem of deciding which objects require to be "just separated" and how to obtain a threshold which will achieve this. This is described in the "splitting" procedure below.

It may not always be possible to keep internal minima within the seed regions. Sometimes the minima between dark chromosome bands are lighter than the minima separating objects (figure 1b). In these cases splitting the chromosome is unavoidable, resulting in neighbouring fall-sets being fragments of the same chromosome. These fragments have to be re-merged, as described in the "merging" procedure below. Both splitting and merging use information about expected chromosome positions obtained at the counting stage. These processes are illustrated by the resolution of the "cartoon" composite in figures 2 and 3.

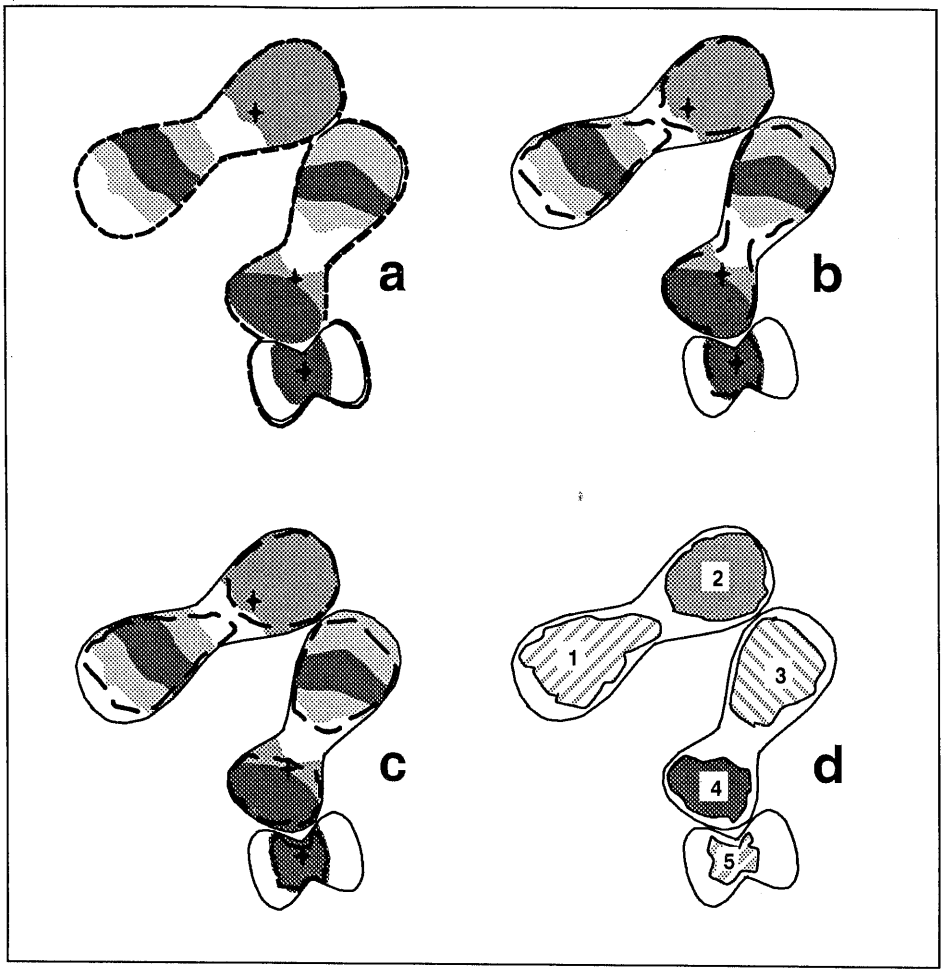


Fig. 2. Illustrating the splitting process. (a) The best threshold produces a boundary (dashed line) which encloses three chromosomes. The positions marked with a dot are those determined at the counting stage. (b) At a higher threshold the object starts to split up. The first split does not produce any new "core" fragments. (c) After several successive rethresholding passes, the object is split into five fragments, three of which are "core" fragments. (d) The "seed" regions obtained by the recursive splitting process, labeled for re-expansion.

3.1. Splitting

The boundary obtained by thresholding using a local histogram is taken as the best separation of chromosome from background. The data structures for describing image regions [7] allow this boundary and the region contained within it to be stored in such a way that image operations can be performed with maximum efficiency on only those points which form part of the boundary or region. One thing which can be easily checked is whether an identifying point

obtained at the count phase is contained within the boundary. If, for example, there are no such points, then the object was not included in the original count and can be rejected as a potential chromosome. If a single point is found inside the boundary, then the object detected is a single chromosome and can be placed directly on the chromosome list. If more than one point is found then the object must be split by re-thresholding.

The re-thresholding can be thought of as proceeding recursively. At each stage the threshold is increased by one grey level, and a new binary slice is obtained only from those points inside the object obtained at the previous threshold value. If an object is split by re-thresholding, each new component is re-thresholded in turn. The splitting stops for a particular object when the number of count-points it contains is either one or zero. Whether this process ends in a single chromosome being isolated or in the creation of several fragments depends on the relative grey levels between light bands and the inter-chromosome valleys and the positions of the count-points (figure 2). This splitting process can be described by a tree structure where the nodes represent splits and the leaves represent the final seed regions. The threshold at which a particular split occurs is recorded at the appropriate node in the split tree.

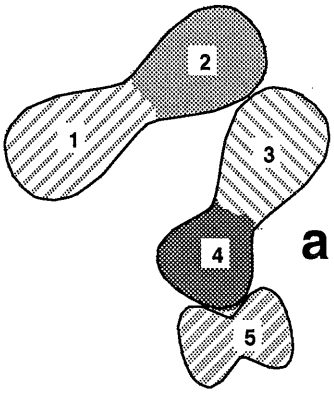
Each of the regions extracted by recursive re-thresholding acts as a "seed" for expansion to its fall-set using Rutovitz's algorithm, within the boundary of the original composite object. The algorithm begins by writing a label value for each seed region into a label image (figure 2d), and at the end this label value has been propagated over the entire fall-set (figure 3). The fall-set of a particular seed region can be obtained by binary slicing the label image at the appropriate label value.

3.2. Merging

The fragments of the original composite arising from re-thresholding and expanding can be divided into two types: "core" fragments, i.e. those which

---

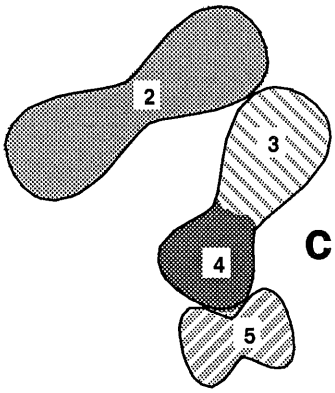
Fig. 3 (opposite). Illustrating the merging process. (a) Regions 1 - 5 are the results of expanding the "seeds" obtained in figure 2d using the fall-set algorithm. Not only do region boundaries occur between individual chromosomes, but two chromosomes have been split into abutting fragments. (b) The region adjacency matrix corresponding to this set of regions. The numbers indicate the likelihood of merging computed from the length of shared boundary and the image intensity at that boundary. "Core" fragments are marked with an asterisk. (c) The most likely merge (1 + 2) has taken place. Fragment 2 gains no new neighbours from this merge. The likelihood of a merge between 2 and 3 has gone down due to the decreased length of the common boundary as a fraction of the total boundary length. (d) The next most likely merge would involve two "core" fragments and is disallowed. Of the two possible merges involving fragment 3, the most likely is 3 + 4. (e) Fragments 3 and 4 are merged. Fragment 4 now gains fragment 2 as a neighbour. The likelihood of merging 4 and 5 is decreased due to the change in boundary length. The merging stops here as only "core" fragments remain.



a

	1	2	3	4	5
1	-	75	-	-	-
* 2	75	-	40	-	-
3	-	40	-	50	-
* 4	-	-	50	-	60
* 5	-	-	-	60	-

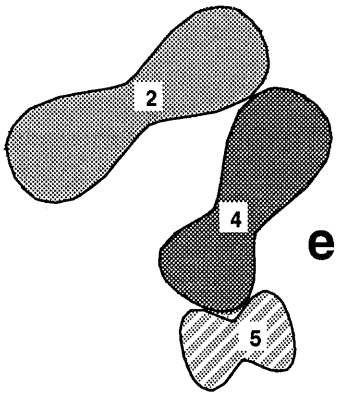
b



c

	1	2	3	4	5
1	-	-	-	-	-
* 2	-	-	30	-	-
3	-	30	-	50	-
* 4	-	-	50	-	60
* 5	-	-	-	60	-

d



e

	1	2	3	4	5
1	-	-	-	-	-
* 2	-	-	-	30	-
3	-	-	-	-	-
* 4	-	30	-	-	50
* 5	-	-	-	50	-

f

ended up containing a single count-point, and "non-core" fragments which ended up containing no count points. The core fragments may represent whole chromosomes or parts of a chromosome; the non-core fragments certainly represent parts of chromosomes and need to be re-merged with neighbouring core fragments to produce complete chromosomes. For each region it is possible to obtain from the label image the labels of neighbouring regions and the lengths of the common boundaries. From the splitting tree it is possible to determine at which threshold value that region separated from each of its neighbours. Both of these values are used in determining a likelihood for remerging each pair of regions.

A region adjacency matrix R is constructed in which each element R_{ij} is a record containing the common boundary length, separating grey level, and merge likelihood for regions i and j. As merging proceeds all of these values get updated. R is, of course, symmetric with $R_{ij} = 0$ for $i = j$. Each row is labeled as either a core or non-core fragment and adjacency values between neighbouring core fragments are initialised to zero to avoid re-merging two separated chromosomes. The matrix R not only indicates directly which merges are available, but can also be easily updated when a merge has occurred. If an object i is merged with an object j then the elements of row j are moved into row i so that previous neighbours of j become neighbours of i. If elements R_{ik} and R_{jk} are both occupied, this means that object k is a neighbour of both i and j. The new R_{ik} is updated by adding together the boundary lengths and substituting the minimum grey level separation in R_{ik} and R_{jk}. The merge likelihoods for row i are recalculated, row j has all adjacencies set to zero and the matrix is made symmetric again. The merge is recorded in the label image by relabeling object j as i (figure 3).

Merges proceed in order of likelihood. The most likely merges occur first since the adjacency values will then have to be updated, affecting the likelihood of other merges. Quite often a non-core region will only have a single core neighbour and merging can proceed without difficulty. In other cases a choice must be made between potential merges. The most likely merge is chosen, of course, but only if the difference in likelihoods exceeds some threshold; the local criteria of grey level and boundary length are not very discriminating and only clear choices are made.

Merging continues in this way until either all non-core fragments have been united with neighbouring cores or the remaining non-core fragments have approximately equal likelihoods of merging with two or more neighbours. The merge criteria do not allow an informed decision to be made in these cases and a safe solution is adopted which returns the difficult case to the operator to be resolved interactively. This is achieved by a second merge pass in which the significance criterion is abandoned and neighbouring core regions are allowed to re-merge. Thus, core fragments which abut via these undecided non-core regions are forced to unite to form composites. Other core fragments not

involved in these adjacency relationships remain separate, so that good work is not all undone. Figure 4 shows the result of applying this procedure to the composite object shown in figure 1.

3.3. Difficulties

Certain situations may occur which do not result in correct segmentations. Overlapping chromosomes, for example, do not fit the initial premise that the objects to be identified are separated by a grey level minimum. Splitting and merging will be attempted however, and, invariably, fail. The best outcome is a complete re-merge to come out with the original object again, but it is possible for the object to be split up into regions, none of which corresponds to a single chromosome and some of which may incorporate parts of more than one chromosome. Incorrect choices in re-merging can occasionally produce similar results in touching chromosomes . A special form of interaction is included in the karyotyping program to enable efficient resolution of this problem at the interaction stage (see below).

Very large composites can result in a large and complex adjacency matrix if the splitting proceeds to completion. The likelihood of erroneous remerging increases as the matrix becomes more complex, and the additional time required to perform the remerging is partially wasted. It is useful therefore to be able to stop the splitting at some controllable level of complexity. This is difficult to achieve if the splitting is truly recursive, as implied above, since recursive procedures are necessarily depth-first. An iterative, breadth first, algorithm is used here : all the nodes at a given level in the splitting tree are expanded simultaneously. The splitting can be stopped at any required level of complexity so that large composites are typically not resolved completely, but split into smaller composites. Useful side effects are that the iterative procedure is faster and the use of memory resources can be controlled.

4. A pilot trial

The results to be expected from the inclusion of this split and merge procedure in the karyotype analysis program are a reduction in the number of interactions required to achieve a correct segmentation, a reduction in the interaction time required and an increase in the time spent in automatic segmentation. The net result should be an overall decrease in the time taken to perform a complete analysis. To test the effectiveness of the method, fifty cells (25 blood and 25 amniotic fluid) were selected from lists provided by the metaphase finding component of the same package [5]. Some selection was applied to avoid analysing the one or two very poor quality cells presented in the list, but the twenty five cells used came from the first twenty nine as selected by the metaphase finder in the amniotic case, and the first twenty six in the case of blood and so can be claimed to be representative of those which might be analysed routinely.

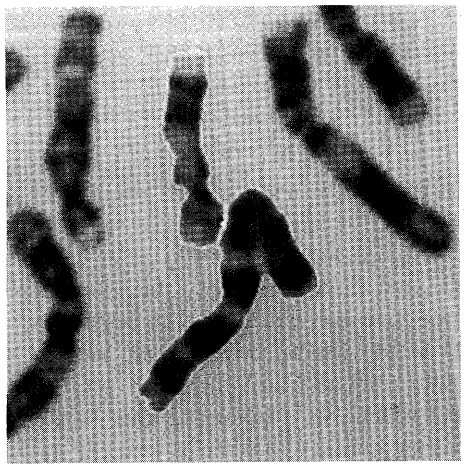


Fig. 4. The result of the split and merge process applied to the composite object in figure 1.

(Indeed in their clinical trial, Lundsteen et al [12] use 10 cells from the first 19 on average, only four of these being fully analysed. This is a much stricter testing environment.)

The cells were first analysed by a "thresholding only" version of the software and then by a "split and merge" version. Automatic segmentation time and interaction time were noted ,as well as the total number of composites for which interaction was required. The measurements applied only to the segmentation stage of the karyotype. Also noted were the number of overlaps which occurred, and which would not be expected to be resolved, and the number of cases of an incorrect merge requiring the use of a new "merge and redivide" interaction mode, not required in the "thresholding only" version. It is important to know that this new facility is not a net creator of interactions! The results are shown in table 1.

Two important figures determining the ease of interaction are the number of composites per cell and the number of chromosomes forming the composite (clump size). Small composites of up to three or four chromosomes may be separated with a single interaction The table shows that the number of composites per cell has been reduced by over 70%, the reduction being rather greater in the case of blood than amniotic fluid. The mean clump size has been reduced slightly, mainly due to the breaking up of very large composites. In fact only two composites consisting of ten or more chromosomes occurred with the split and merge software whereas thresholding alone resulted in ten of these large objects. Segmentation using thresholding alone resulted in the requirement for at least one interaction in all cells, and two or more interactions in all but two. Using the split and merge software, no interactions were required in 28% of cells and only one interaction in a further 18%.

Table 1. Summarised data on segmentation performance with and without the application of the split and merge procedures

	AMNIOTIC		BLOOD		ALL	
	Thres-hold only	Split and merge	Thres-hold only	Split and merge	Thresh-hold only	Split and merge
Total number of cells	25	25	25	25	50	50
Total composites	193	59	171	50	364	109
Average number of composites per cell	7.78	2.36	6.84	2.0	7.28	2.18
Average clump size	3.6	3.2	2.9	2.4	3.25	2.81
Number of cells with no interactions	0	5	0	9	0	14
Number of cells with one interaction	0	5	1	4	1	9
Number of composites which include overlaps	3	7	11	16	14	23
Number of "merge and redivide" interactions	0	6	0	11	0	17
Average automatic time	15.7	29.9	17.2	31.1	16.4	30.5
Average interaction time	85.5	40.7	59.4	34.9	72.5	37.8
Average segmentation time	101.2	70.6	76.6	66.0	88.8	68.3

The number of composites which included a chromosome overlap increased using the split and merge software. At first sight this may seem rather curious since the same cells were used in both tests. The explanation is that composites containing two or more overlaps have largely been broken up into composites containing only one. These comprise about 20% of the residual requirements for interaction. Of the seventeen cases in which the new "merge and redivide" interaction was required, six were due to the existence of overlaps. The remainder were due to incorrect decisions in touching chromosomes.

The total segmentation time has been reduced by some 20%-30%. Interestingly, the greatest time saving occurs in the case of amniotic fluid samples, where the remaining interactions are greatest in number. This must be attributed to the removal of the very large composites which require long interaction times.

5. Discussion

Karyotype analysis of metaphase cells has for a long time been a target for the application of automatic image interpretation methods. In recent years this effort has borne fruit in the form of a number of commercially available karyotyping systems, some with associated metaphase finders and some without. All of these systems require operator interaction to achieve complete analysis of all but the most perfect cells. Clinical trials on the system implemented on the Joyce Loebl Magiscan image analyser conclude that the main requirement for interactions is in the separation of touching chromosomes [8].

I have described here a method which automatically resolves a large proportion of the most commonly occurring cases of composite chromosomes and which can eliminate the need for interaction at the segmentation stage in significant proportion of cells of routine quality. The method has been aimed at resolving the composites which occur most frequently in blood and amniotic fluid samples, for which the analysis package was primarily intended, namely small clumps of touching chromosomes. The resolution of overlapping chromosomes almost certainly requires some syntactic shape analysis (see for example Ji [13] this volume) and may be the next most pressing target for reducing the number of interactions. Large agglomerations of touching chromosomes may be only partially resolved.

The results of a small pilot trial show that significant reductions in the requirement for interaction have been achieved and that this results in some time-saving in the overall analysis. The main advantage seems to be that the worst types of interaction for the operator, namely splitting large composites, have been effectively removed and that, in a substantial proportion of cells analysed, no interaction at all is required for segmentation. Lundsteen et al [12] have included a version of the software including the "split and merge" procedure in a test of clinical performance involving a large set of cells from the routine laboratory production. Several different software configurations were under test in this trial, and it is not easy to identify the improvement in interaction requirement due to segmentation alone. However, one can estimate that about 40% fewer interactions were required at the segmentation phase in their study. This improvement is somewhat less than would be expected from the pilot trial presented here. The difference may be due to a difference in the quality of preparation used in the two trials but this seems unlikely since the slides used here were obtained from the Rigshospital chromosome laboratory. The regime in the Rigshospital trial was to fully analyse four cells from the first nineteen detected and ranked by the metaphase finder. Thus the comparison consisted of analysis of relatively "easy" metaphases. The figures presented here suggest that this method produced a substantial improvement in the analysis of more difficult cells. One result of using this procedure may be a reduction in the number of cells which have to be located in order to perform the required number of full analyses.

The clinical trial did not produce figures on the number of cases in which no interaction was necessary at segmentation. The authors do however speculate on the possibility of using this version of the software as the basis of a "semi-hands-off" karyotyping system in which the operator only interacts at the count stage and to correct and verify the final karyogram.

The knowledge which drives this method is very straightforward - a list of points, each one internal to a single chromosome. This knowledge is available in the system for other purposes and its use here has been rather opportunistic. The simplicity of this knowledge is responsible for some of the residual errors. The re-merging process for instance would be much more reliable if the objects created by merging where tested for their plausibility as chromosomes, rather than simply using local context-free information about the shared boundary between merged fragments. There is some interest now in the use of "knowledge based" methods derived from the field of artificial intelligence in generating a truly automatic karyotyping system. Piper et al [14] give a very clear account of the complex set of relationships between objects and the knowledge about those objects which arise in the karyotyping problem. A complete specification of this knowledge and these relationships seems a daunting project. This study indicates that the application of explicit knowledge locally to specific tasks such as composite resolution has considerable promise for success.

References

1. Graham J. The Magiscan Interactive Chromosome Karyotyper - MICKY. Proc. V European Chromosome Analysis Workshop, Heidelberg (1983)
2. Graham J. Recent Developments on the Magiscan Karyotyping System. Proc. VI European Chromosome Analysis Workshop, Leiden (1985)
3. Lundsteen C, Martin AO. On the selection of Systems for Automated Cytogenetic Analysis. Am. J. Med. Genet. In press.
4. Graham J. Automation of Routine Clinical Chromosome Analysis I. Karyotyping by Machine. Anal. Quant. Cytol. Histol. 9:383-390 (1987)
5. Graham J, Pycock D. Automation of Routine Clinical Chromosome Analysis II. Metaphase Finding. Anal. Quant. Cytol. Histol. 9:391-397 (1987)
6. Taylor CJ, Dixon RN, Gregory PJ, Graham J. An Architecture for integrating Symbolic and Numerical Image Processing. In: Duff MJB (ed) Intermediate Level Image Processing (Academic Press, London, 1986) pp.19-34
7. Graham J, Taylor CJ, Cooper DH, Dixon RN. A Compact set of Image Processing Primitives and their Role in a successful Application Program. Patt. Recog. Lett. 4:325-333 (1986)
8. Philip J, Lundsteen C. Semi-automated Chromosome Analysis. A Clinical Test. Clinical Genetics 27:140-146 (1985)
9. Serra J. Image Analysis and Mathematical Morphology. (Academic Press, London, 1982) pp.456-463
10. Friedlander F, Meyer F. A sequential Algorithm for Detecting Watersheds in a Grey Level Image. Proc. 7th Congress of ISS, Caen, France (1987)
11. Rutovitz D. Expanding Picture Components to Natural Density Boundaries by Propagation Methods. The Notions of Fall Set and Fall Distance. Proc. IV IJCPR, Kyoto, Japan, (1978) pp.657-664
12. Lundsteen C, Gerdes T, Maahr J, Philip J. Clinical Performance of a System for Semi-Automated Chromosome Analysis Am. J. Hum. Genet. 41:493-502 (1987)
13. Ji L. Automatic Resolution of Overlapping Chromosomes. This volume.
14. Piper J, Baldock R, Towers S, Rutovitz D. Towards a Knowledge Based Chromosome Analysis System. This volume.

12. **Automatic karyotype analysis.** J Graham and J Piper, in *" Chromosome Analysis Protocols", J.R. Gosden (ed), Humana Press inc, Totowa NJ, pp141 - 185, 1994.* doi:10.1385/0-89603-289-2:141

Chapter 11

Automatic Karyotype Analysis

Jim Graham and Jim Piper

1. Introduction

1.1. A Warning to the Reader

This chapter differs from the majority in this book in that its subject matter is the application of computer image interpretation techniques to the analysis of metaphase chromosome spreads. Were we to follow the prescription of the remainder of the book, we might simply publish the code of a computer program together with a list of suitable equipment. This is not, however, a realistic option. The computer programs in current commercial systems for automated cytogenetics typically consist of approximately 100,000 lines of source code; in the case of automatic metaphase finders, the equipment may include proprietary mechanical or electronic components. Also, the rate of change in the performance and cost of cameras, displays, and computers are such that any list of equipment that is appropriate as we write in mid-1991 would most likely be nearing obsolescence by the time the book is published.

We have therefore decided to describe, in some detail, the main procedures that have to be implemented in software, and the minimum performance that would be required from commercially available computer and imaging hardware in order to run the software successfully. Both should be specified in sufficient detail that a working system could be built by an experienced computer programmer.

We will model our exposition on the best currently available technology. Commercially, we believe that this is represented by the Magiscan

From *Methods in Molecular Biology, Vol. 29. Chromosome Analysis Protocols*
Edited by J R Gosden Copyright ©1994 Humana Press Inc , Totowa, NJ

(Joyce Loebl, Gateshead, UK), Cytoscan (Image Recognition Systems, Warrington, UK) (both of these companies are now part of Applied Imaging International, Inc., Sunderland, UK) (the authors were, respectively, involved in the development of these two systems), AKS-2 (Amoco Technology Inc., Naperville, IL), Genetiscan (Perceptive Scientific Instruments Inc., League City, TX), and similar machines, together with recent developments that may well appear in commercial products in the near future.

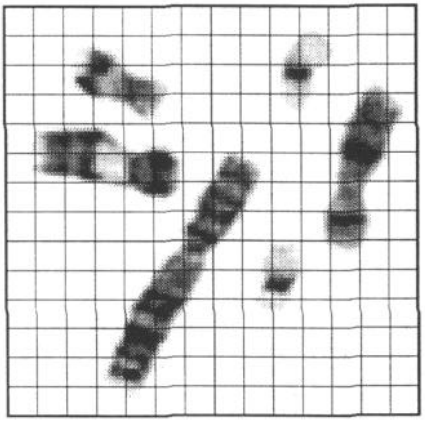
However, we think it essential to warn readers contemplating making their own system that this is a big undertaking. The time and effort required depend, of course, on the skills and equipment available, but we cannot imagine that a simple but usable interactive karyotyping system (without automatic classification) could be programmed with less than six months of effort from an experienced programmer; an automatic karyotyper would require at least a year, as would an automatic metaphase finder. Polished, reliable, and ergonomic versions might increase these times by a factor of between two and ten! Those who think that these estimates are excessive should bear in mind the widely held belief that a competent computer programmer can, on average, produce just ten lines of correct, bug-free, and documented program code per working day. We believe that with modern operating systems and software tools, this figure is an underestimate; however, even a simple "no-frills" automatic karyotyping system would most likely require several tens of thousands of lines of code.

1.2. A Brief Introduction to Image Analysis

Automatic karyotyping involves the analysis by computer of two-dimensional light microscope images. In outline, such analysis involves the following stages.

1.2.1. Image Capture

Light transmitted through the microscope slide is focused onto the target of an electronic camera and sampled on a regular two-dimensional grid. The resulting brightness values or "pixels" are stored as numbers in computer memory (Fig. 1). The pixel values and their geometric positions are the basic data for all subsequent analysis. The natural unit of measurement in image analysis is the spacing between pixels in the image (the sampling interval), and unless stated other-



A

1	1	1	1	1	1	1	1	1	3	2	1	1
1	1	8	2	1	1	1	1	1	6	1	1	1
1	1	5	6	6	1	1	1	1	3	1	2	2
1	2	2	1	1	1	1	1	1	1	1	6	5
2	6	4	4	8	1	1	2	1	1	1	9	2
1	1	2	4	9	1	6	7	1	1	1	5	1
1	1	1	1	1	1	7	2	1	1	6	5	1
1	1	1	1	1	4	6	1	2	1	4	3	1
1	1	1	1	3	6	1	1	5	1	1	1	1
1	1	1	1	6	6	1	1	3	1	1	1	1
1	1	1	7	6	1	1	1	1	1	1	1	1
1	1	1	6	2	1	1	1	1	1	1	1	1
1	1	1	3	1	1	1	1	1	1	1	1	1

B

Fig. 1. **A.** A simulated image with sampling grid superimposed. **B.** The measured pixel values. The darker pixels (shaded) comprise several distinct connected components.

wise, measurements and geometrical constructions will be assumed to be based on this pixel spacing unit.

1.2.2. Segmentation

The set of pixels corresponding to a single chromosome must be determined so that they may be processed together in order to make measurements about the chromosome separately from the other objects in the field. This can be achieved, for example, by choosing a "darkness threshold" that is a little darker than the mean pixel value of the clear field between the chromosomes. Then pixels that are darker than the threshold "belong" to chromosomes, and individual chromosomes can be separated by finding connected subsets of the darker pixels (Fig. 1). Of course, this procedure is applicable to images other than metaphase cells, and sometimes other dark objects will occur even in metaphase fields, for example, interphase nuclei, so we will refer in general to such segmented sets of pixels as "image regions."

1.2.3. Feature Measurement

Feature measurements on each image region are made by applying mathematical formulae to the set of pixels; in many cases (e.g., for

shape moments), the positional coordinates of pixels are also required. For example, the area of a chromosome may be estimated simply by counting the number of pixels. In practice, things are usually a little more complicated. In particular, in the case of chromosomes, most of the useful measurements depend on the position of the pixels relative not to the original Cartesian coordinate system of the pixel digitization grid, but to the chromosome's medial or symmetry axis, which may be imagined running between the chromatids, and so finding this axis is a necessary precursor to making such measurements, of which two obvious examples are the chromosome's length and centromeric index.

In order to make such feature measurements efficiently, each segmented image region or chromosome should be represented in an appropriate data structure. As will become clear in Section 3., the data structure must be capable of representing the arbitrary shapes, sizes, and orientations of chromosomes, and allow access to the pixel values so that computations may be made concerning the banding pattern. To describe such structures in detail is beyond the scope of this chapter, but examples may be found in refs. *1* and *2*.

1.2.4. Classification

Classification of image regions is typically made by applying statistical rules to a set of feature measurements. The rules are initially obtained either by introspection by the system designer, or more usually from a "training" or "design" set of image regions of predetermined class. Alternative classification schema, known variously as "syntactic" or "structural," are based on recognizing the "grammatical" arrangement of substructures of the image. More complicated systems based on artificial intelligence principles are the subject of current research.

1.2.5. Model

Crucially, a model or set of general principles that predict how a given biological entity, such as a metaphase chromosome, will be represented in digitized pixel values is essential to guide the search for meaning in the digitized image. The model has to accommodate both biological variability (for example, how to deal with touching chromosomes, bent chromosomes, the random position of chromosomes within the metaphase, or different metaphase contraction states), and image

degradation on account perhaps of noise from the camera or a less than optimally set up microscope. Typically in chromosome analysis, the models are implicit rather than explicitly stated and are simplistic in the extreme, partly accounting for the widespread reliance on operator interaction for many of the nontrivial decisions.

1.3. Metaphase Finding

An essential component in any investigation involving chromosome analysis is the location of dividing cells of sufficient visual quality to permit the assessment of the chromosomes. The required cells may be at metaphase, prometaphase, or prophase, but the visual task does not vary greatly, and we will speak generically of "metaphase finding."

Unautomated metaphase finding involves visually scanning the microscope slide fairly rapidly at low magnification. When a metaphase is seen, it is examined carefully to assess its suitability for detailed analysis. If it appears suitably compact, well stained, and well spread, it may then be reexamined (either visually or automatically) at high magnification, when its quality can be completely determined and analysis performed as appropriate.

The proportion of the total analysis time and effort devoted to metaphase finding depends on the goals of the chromosome analysis and the material used. In "classical" (randomly induced) aberration scoring, for example, where the material is normally peripheral blood and the mitotic index is high, good-quality metaphases are easily found. The subsequent examination of individual cells, however, is very rapid, since dicentrics, acentric fragments, ring chromosomes, and so forth, are easily identified visually. The time spent locating new cells can therefore contribute significantly to the total analysis time. Karyotyping of bone marrow cells for leukemia diagnosis or treatment monitoring provides an example of a different type of task. The analysis of individual cells is a difficult, time-consuming exercise, but the visual quality of the metaphase cells may be so poor and the mitotic index so low that it may be necessary to locate all the dividing cells in a sample, involving a thorough search of the entire slide. In this case also, finding the metaphases is a significant proportion of the total task.

In clinical karyotyping using amniotic fluid or peripheral blood, metaphase finding contributes less significantly. Mitotic indices are high, and good-quality metaphases are fairly easily found in most

routine material, although this is less true if direct chorionic villus samples are used. There are tasks of clinical interest, however, in which the metaphase finding component can be significant. Detection of fragile sites requires the examination of the order of 100 cells to be sure of correct diagnosis. The inspection of each cell can be fairly rapid, since the identification of the fragile site is often straightforward and the role of metaphase finding is similar to that in classical aberration scoring. Prophase analysis has similarities with cancer cytogenetics in the respect that a very thorough search might be necessary to locate a small number of cells in which the required bands can be identified on both homologous chromosomes.

1.3.1. Automatic Metaphase Finding

The central role of metaphase finding in all aspects of chromosome analysis and the fact that in some investigations it is a significant task in itself have led to the development of a number of automatic metaphase finders, some of which are associated with automated karyotyping systems. Finding metaphases automatically is a fairly typical image analysis task. However, the quantity of data is enormous. With the usual pixel size of about 1 μm^2, a coverslip area comprises about 10^9 pixels. Metaphase finders therefore aim to use simple, but fast, analysis methods.

It can be argued that metaphase finders have been more successful technically than karyotyping systems. Other than initial definition of the area of slide to be searched, they require no operator interaction and have been demonstrated to be highly efficient at identifying dividing cells *(3)*. Most metaphase finders also include a measure of metaphase quality that can be calculated when the cell is found, allowing the cells to be presented for analysis in ranked order. This is a useful feature in clinical karyotyping where only a small number of cells is required, but it is important that they should be of good visual quality. In particular, in prophase analysis, selection of cells in which the number of overlapping chromosomes is small may make a highly significant contribution to the efficiency of the overall process.

In the following paragraphs, we briefly discuss some of the features or properties of existing metaphase finders as they will be perceived by the user. In Section 3.1., we discuss some of the technical aspects of metaphase finding that influence these features. Detailed assessment

of some of these features for metaphase finders in use in Europe has recently been made by Korthof and Carothers *(3)*.

1.3.2. Speed

At first sight, it appears that a high scanning speed is an essential feature of a metaphase finder. It is certainly the case that running the metaphase finder should not result in a significant time overhead for the busy cytogenetic laboratory. High scanning speeds can be achieved by applying highly optimized image acquisition methods (as used by Cytoscan, for example, *see* Section 3.1.1.). Machines based on less application-targeted hardware, which acquire their images using a television (TV) camera, scan more slowly. In the Korthof and Carothers survey, Cytoscan achieved scanning speeds of 88 s/cm^2 of slide on average over a range of material, whereas Magiscan took 596 s/cm^2 and other TV-based systems (no longer available commercially) took over 1000 s/cm^2. Also using line scanning, the recently developed Geneti-Scanner from Perceptive Scientific Instruments is reported to achieve scanning speeds of about 66 s/cm^2 *(4)*.

Speed is clearly important. All other things being equal, it is better to find metaphases quickly rather than slowly. However, TV-based systems may compensate for their lower scanning speeds by running metaphase finding overnight on a number of slides, which may be as beneficial, or more so, to clinical laboratory throughput as a scan of a few minutes on each slide as required. Performance comes at a cost, and the impact of fast scanning speed on the entire automated analysis package must be considered in assessing the benefit obtained.

1.3.3. Accuracy

We can consider accuracy in terms of false-negative rates (undetected metaphases), false-positive rates (nonmetaphases classed as metaphases), and ranking ability (the number of good-quality metaphases placed early in the analysis queue). For many clinical applications, the latter may be the only measure of interest to the user. If material with a low mitotic index is being analyzed, as in cancer cytogenetics, false-negative rates do become important. Fairly high false-positive rates may be tolerated, provided ranking is sufficiently good that nonmetaphases are only rarely presented to the user.

1.3.4. Adaptability

Slide preparation for karyotyping varies widely from laboratory to laboratory, and of course among different types of specimen. The parameters used by a metaphase finder should therefore be adjusted to provide optimal performance for each type of material for each laboratory. If these parameters are adjustable by the user, changes in laboratory practice, such as improvements in preparation techniques or introduction of new types of investigation, can be accommodated conveniently. The most suitable method of adjustment by the user is through system training, in which the metaphase finder is used on the new material and the trained operator labels each of the objects found according to its quality. The system then uses these quality scores to match the measured parameters to the desired properties.

1.3.5. System Considerations

Metaphase finding is never an end in itself. Automatic metaphase finding is always a component in some overall analysis task, and the features of a metaphase finder must be considered as part of the overall system. A slow metaphase finder as part of a stand-alone karyotyping system may enhance the overall system efficiency at fairly little additional cost. A laboratory with a large throughput may find a fast metaphase finder to be a useful central resource servicing a number of independent karyotyping stations. The Geneti-Scanner is clearly intended to be used in this way, since it can be loaded with up to 60 slides. After a setup period of about 30 min, these can be scanned unsupervised *(4)*.

The usefulness of a metaphase finder in any karyotyping environment depends not only on its basic performance characteristics, but also on its user interface. It is important not only that every cell presented to the operator for analysis is analyzable, but also that the operator interaction in loading the slides and specifying the scan parameters should be minimal and straightforward.

1.4. Automatic Karyotyping

Automatic karyotyping aims at describing the chromosome complement or karyotype of a metaphase cell and producing an annotated karyogram (an arrangement of images of the chromosomes in a prescribed pattern). In the early days of research in automated cytogenetics, the goal was to produce a completely automatic system.

However, the end product of the process of karyotyping is a statement about the genetic constitution of an individual from the point of view of his or her health. The consequences of that decision are of great importance to the individual. The decision is influenced by a number of factors, some of which involve detection of subtle signs in the image and some of which involve information not present in the image at all. For these reasons, it is clear that, for the foreseeable future, the assessment of the data leading to that clinical decision will be made by highly trained human beings and not by computer systems. Whatever level of computer assistance is provided, it will be in the form of an aid to human decision making. That is to say the system will be interactive.

A karyotyping system will be involved in either counting, or counting and fully analyzing, the chromosomes in a metaphase. However, the initial image obtained from the camera, and thresholded and segmented as described in Section 1.2., usually contains objects other than isolated chromosomes, notably chromosome clusters (both of touching and overlapping chromosomes), interphase nuclei, and noise of one sort or another (for example, stained cytoplasm, stain particles, or other dirt). In order to complete the classification and produce a karyogram, or even simply to perform a count, the objects that represent nonchromosomal material must be rejected, and the clusters resolved as far as possible into individual chromosomes. For karyotyping, the isolated chromosomes must be measured and classified. Methods for carrying out these procedures automatically are described in Sections 3.4.–3.6.

The current generation of image analyzers fall well short of performing these tasks with 100% reliability, either because the necessary algorithms cannot be run quickly enough on currently available computers, or because sufficiently sophisticated algorithms have not been developed. The result of this is that intervention by the operator is needed to resolve difficulties in the detailed analysis. From the point of view of the system developer, this is embarrassing, but because of the fact that karyotyping is inherently interactive, useful systems can be provided, albeit requiring rather more input from the user than is ideally desirable. The important features from the point of view of system usefulness are the number of interactions, the ease of interaction, and whether the interactions intrude on the user's interpretation of the image of the chromosomes.

1.5. Human–Machine Interaction

1.5.1. Interface Design

The process of interaction involves communication of knowledge between the user and the machine. For this communication to take place, it is necessary to have a physical medium on which messages are passed and an agreed vocabulary.

1.5.1.1. The Physical Interface

The physical components of the interface are displays, keyboards, and pointing devices, such as lightpens, mice, graphics tablets, or trackerballs. There has been some experimentation in karyotyping systems with voice input, but this technology is insufficiently advanced at this time to provide appropriate interaction.

The most important item to be displayed is the metaphase image, although other items, such as menus and textual information, need to be displayed also. The image display needs to be of high quality, since the final decision is often based on fairly subtle image features. The minimum specification acceptable for image display is 512 × 512 pixels of 64 gray levels. With earlier systems, the visual quality of such an image was generally believed to be poor compared to that obtained directly from the microscope or on a photograph, but probably sufficient for diagnostic purposes *(5–6)*. Nowadays, cameras and display monitors are available that are capable of considerably higher spatial and gray level resolution.

The display of nonimage information, such as menus or text, may be considered intrusive if it occupies the same area as the image. One answer to this is to use a separate display for this information. Alternatively, if high-resolution displays are used, a section of the display may be devoted to textual information without intrusion on the image. Many computer systems provide display management software using "windows," which allows information from several sources to be displayed and manipulated independently on the same screen. Thus, the areas of the screen to be used for different purposes may be altered interactively. Areas may be used temporarily for special purposes, such as magnification of selected regions of the image without altering the underlying display. The AKS-2 system, which is based on a MacIntosh computer, uses the high-resolution display and windows environment very effectively in this way.

Pointing devices are needed, since interaction usually involves specifying particular objects or parts of objects in the image, or selecting menu options. A mouse is preferred for this purpose, being inexpensive, robust, and easy to use. There is an advantage in using a lightpen for some types of interaction, particularly if careful drawing is required, but in general, lightpens are less satisfactory than mice. For karyotyping, trackerballs and graphics tablets appear to offer no particular advantage.

It is inevitable that some textual input will be required, such as sample identifiers or comments on a karyotype, and for this purpose, the keyboard is indispensable. However, it need have no other role in interaction, and its use should be kept to a minimum.

1.5.1.2. User Model

By "user model" we mean the user's understanding of the objects displayed in the human–machine interface and his/her expectation of the system's behavior on interaction. A widely known example of a user model is that generated using the "desktop metaphor" employed by a number of office systems, where the user interacts with the system using concepts familiar from the everyday world, such as filing cabinets and wastepaper baskets. This model is successful, because it allows the user to express his/her requirements in terms of objects and activities that characterize the task, rather than the machine's implementation. In karyotyping, similarly, an appropriate interface should require the user to specify his/her requirements in terms of such objects as metaphases, chromosomes, centromeres, chromatids, or karyograms. It should not be necessary to require the user to think in terms of thresholds or pixels; these concepts are not difficult to cope with, but they introduce an element of opacity into a system that should be made as transparent as possible.

1.5.2. The Interaction Process

Here we outline some of the types of interaction that may be expected in automatic karyotyping systems.

1.5.2.1. Segmentation

Inadequacies in segmentation algorithms generally show up as the inability to separate touching or overlapping chromosomes. In the case of touching chromosomes, it is easy for an operator to indicate with the pointing device the place where the composite object should be cut,

either by drawing a separation line or indicating a few points around the cut location. Overlapping chromosomes call for more complex interaction, since the extent of each chromosome in the composite must be separately indicated. Cytoscan has a convenient method for achieving this by drawing a rough axis for each chromosome.

1.5.2.2. Axes and Centromere Positions

Defining a chromosome axis or centerline is a common step in extracting a number of important chromosome measurements (*see* Section 3.4.). For badly bent chromosomes, this may not be easy to define automatically. Centromere positions can also be difficult to measure, particularly in the case of highly elongated chromosomes. Since classification performance will be affected by errors in axis and centromere positions, correction of automatically generated positions may be required. This usually involves drawing a correct axis with the pointing device or indicating a correct centromere. However, it is often easier to accept errors of this type and correct the resulting classification errors at a later stage, and there is some evidence that this results in fewer interactions overall *(7)*.

1.5.2.3. Classification (Karyogram)

Classification errors occur whether or not all stages in the analysis of the image have proceeded correctly. All of these errors show up as misplaced chromosomes on the initially presented karyograms. Chromosomes may either be in the wrong locations on the display or allocated to a "reject" class. Since the chromosomes must be examined carefully at this stage, e.g., for small structural abnormality, interaction to correct these errors is not a serious overhead, provided that there are only a few corrections to be made. Such corrections are generally made by pointing to a chromosome on the display and indicating its correct position on the karyogram. Options will also be available for inverting a chromosome that has been presented the wrong way up or shifting chromosomes so that they are correctly aligned, in addition to other possible presentation facilities, such as rotation, chromosome straightening, or banding pattern enhancement *(8)*.

1.5.2.4. Counting

Counting chromosomes is the most easily described task in karyotype analysis, but it is one that is most difficult to automate. This is because the whole procedure must be carried out quickly (as quickly

as the chromosomes can be counted by eye in the microscope). Using the types of computers appropriate for karyotyping systems, this precludes the use of highly sophisticated segmentation algorithms. Automated karyotyping systems approach this problem in various ways. In the Magiscan system, each image is digitized before any analysis takes place, and the user interacts with the image at all times via the screen. The Magiscan system provides semiautomated counting, in which an approximate count is presented to the operator, with all chromosomes marked, for correction by pointing at false chromosomes or missed chromosomes with the lightpen *(9)*. Since the user may wish to examine every chromosome at this stage in any case, it is frequently found more convenient to disable the automated counting phase and to have the operator simply mark each chromosome in turn. In this way, the operator is guaranteed to look at each chromosome in the image, and the count is generated as a byproduct. The Cytoscan system also provides interactive, screen-based counting. Again, however, except in particular cases (e.g., when counting hybrid cells with very many chromosomes), in typical use, counting is done entirely by "eyeball" analysis, in this case, of the metaphase directly in the microscope, because it is faster and easier for the operator, and not all counted cells need be analyzed further.

Having a mark reliably placed on the interior of each chromosome at the counting phase provides information that can be used to cut down the number of segmentation interactions required *(10)*. This indicates that the provision of interaction in karyotyping systems is a matter that should be considered at the system level and not merely as a local "fix" to a processing problem. The need for a large number of interactions does not necessarily signify an inefficient combination of operator and machine. Interactions that advance the operator's understanding of the image do no harm and could even be beneficial to the overall process. What should be kept to a minimum is interactions in which the user is performing low-level tasks because of the machine's lack of competence.

1.5.2.5. Wholly Interactive Systems

The Magiscan, Cytoscan, and AKS-2 systems are intended to be fully automatic systems, requiring interaction only to assist the automatic process. Other systems take a different approach. Genetiscan, for

example, provides a completely interactive environment in which the operator indicates each chromosome with a pointing device and specifies its group, whereupon it is transferred to a karyogram. In systems of this kind, the method of isolating each chromosome varies, but may involve indicating several points to specify the axis of the chromosome. This style of interaction ensures that the operator examines the structure of the chromosomes. It may, however, require consideration of factors not normally of great interest to the operator, such as exactly where a chromosome bends, and can involve a large number of interactions for each image.

1.5.3. General Interaction Features

Whatever physical device is used to create the interface, and whatever the details of the user model, certain interaction features are essential to maintaining a "user-friendly" interface.

1.5.3.1. Instantaneous Response

Interaction frequently involves some new processing of the image or part of it following a user request. This processing should not be apparent to the user, who is interested only in the result. Being forced to wait for the response to a command can intrude on the course of an interaction, and it is important that responses should appear to be instantaneous. Thus, good user interaction can demand the use of a powerful computer.

1.5.3.2. Visual Feedback

Many interactions involve indicating significant image regions by pointing or drawing. It is important that the user is kept aware of the machine's interpretation of interactive requests by suitably highlighting regions or lines, and indicating the type of interaction being undertaken. In the case of Cytoscan, as the pointer is moved around the screen by the mouse, the nearest chromosome (or other object) is highlighted by a box surrounding it, and it is this chromosome that is selected for any subsequent command. This method has the added advantage of immediately confirming whether a chromosome is separated or is part of a cluster. Visual feedback should take account of the context of the interaction, for example, by warning of illegal or ambiguous user requests. The need for instantaneous response is particularly important here.

1.5.3.3. Undoing Interactions

A feature so useful as to be indispensable is the ability to undo the results of an interaction by returning to a previous state by at least one, but preferably several, steps. This is needed not only because interaction errors occur and need to be corrected, but also because it is sometimes unclear exactly what the outcome of an interaction should be. For example, it may be difficult to decide on the correct way to divide up a collection of overlapping chromosomes, and several trials may be needed.

1.5.3.4. Consistent Interface

It is much easier to interact efficiently if the style of interaction is kept as consistent as possible across different types of interaction. For example, all interactions might begin by selecting an object from the image. Many interactions involve drawing lines for different purposes. Line drawing should always be invoked and executed in the same way. If a mouse is used, the functions of the mouse buttons should be consistent. Interactions that are similar to each other should involve the user in similar actions.

1.5.3.5. User Assistance

It should be possible to obtain on-line descriptions of the available options by the provision of a "help" facility at each interaction stage.

2. Materials

2.1. Equipment Required for Karyotype Analysis

The minimum equipment required by those intending to build their own karyotyping system comprises a microscope, a computer, a camera to acquire the images, and some means of display. The computer requires reasonable power and a good program development environment, ideally a scientific workstation running UNIX or a top-end PC. In either case, if the computer has a high-resolution monitor, then the display of images (digitized metaphase, karyogram) may be adequate without further equipment. The display should have at least 800×600 pixels and at least 6-bit (64 level) gray-scale resolution. Color is not particularly important, but some means for displaying graphical overlays is desirable. Some means of interactive control apart from the key-

board is needed; the ubiquitous mouse is ideal. Windows software is highly desirable, but not essential. You will need adequate disk space (at least 50 Mbyte) and some means of backing up images and software (e.g., magnetic tape or optical disk).

In order to digitize the image from the camera, a frame grabber will be required. In the past, these have had a display capability that required a separate monitor (in which case the requirements of the computer itself are obviously less in this respect than those specified above); nowadays, however, it is possible to display the digitized image directly in a window on the work station and possibly also the live image prior to digitization. The frame grabber will acquire a rectangular array of pixels by digitizing the camera signal; this frame should be at least 512×512 pixels, with 6-bit gray-scale resolution. It is desirable that the pixel spacing be the same in each direction (square or 1:1 aspect ratio).

The camera should be a high-quality monochrome camera, either vacuum tube (Chalnicon or Newvicon tubes are best) or a CCD with resolution (number of pixels) similar to the frame grabber. There are several important aspects to bear in mind. First, CCD cameras are usually most sensitive in the near infrared part of the spectrum, and an infrared filter either within the camera or on the microscope will be essential for good image contrast. Second, an electrical low-pass or antialias filter with a cutoff frequency of about 6 MHz should be included between the camera and frame grabber (unless either already incorporates one), in order to reduce high-frequency white noise, and the "aliasing" effects that can arise between a CCD camera and a frame store unless their pixel clocks are synchronized (which is usually not possible). Third, the sensitive area of most CCD cameras is smaller than that of "1-in." vacuum tube cameras, resulting in an apparently larger magnification of the digitized image and correspondly smaller field of view. If you intend to use a "$^2/_3$-in." tube or CCD camera, you should consider acquiring a 63× objective instead of (or in addition to) the 100×, which is more suitable for use with a 1-in. camera. An alternative is to use a zoom attachment, available for most good microscopes. The camera can be attached to a standard microscope by using a C-mount adaptor in the photography port. Fourth, you will need to experiment with microscope color filters in order to obtain the best contrast (e.g., of G-bands) with the chosen camera. It is worth bearing in mind that the camera will most likely be placed at the

primary focus of the objective, where, depending on the microscope, the image may not be fully color-corrected, in which case a fairly narrow bandwidth filter may well improve the image sharpness.

A hard-copy device is essential for printing karyograms. There are two main types. Video printers attach directly to the composite video signal fed to a monitor; digital printers are capable of better resolution, but require a digital interface and will generally be slower in use. Relatively cheap video printers give reasonable results, with perhaps 64 levels of gray on thermal paper. Digital thermal printers are a little more expensive, whereas photographic laser printers (not to be confused with the widely used xerographic laser printers, which are unsuitable) are capable of superb reproduction quality, but are expensive to buy and to run.

2.2. Additional Hardware Requirements for Metaphase Finding

To search for metaphases autonomously, the metaphase finding computer must be in control of the microscope. That is, it must be able to move the microscope stage along its x and y axes with adequate speed and accuracy; it must maintain focus during its search and so must be able to move the stage along its z axis. This is usually achieved by using a microscope stage fitted with stepper motors that are driven from the computer. The step size of these motors should be no greater than about 5 μm. A stepper motor can be attached to the focus control and should be geared to produce movements of the stage in the z direction in steps of about 0.1 μm. It is also useful to have computer control of the microscope lamp to maintain a suitable constant illumination level during the search.

3. Methods

3.1. Metaphase Finding

In this section, we consider some of the technical issues that must be addressed in constructing a metaphase finder and that determine the system's performance characteristics.

3.1.1. Image Capture

As outlined in Section 2.1., image capture for karyotype analysis is generally done using a television camera, and this is also true of most metaphase finders. Scanning for metaphases involves moving

the microscope stage in fixed size steps and allowing it to stabilize between image captures. An alternative approach has been adopted by Shippey et al. *(11),* who exploited the fact that scanning is inherent to metaphase finding. The Fast Interval Processor, later to be made commercially available as the Cytoscan metaphase finder, uses a single-line CCD detector rather than the two-dimensional array of a television camera. The second image dimension is obtained from the continuous motion of the stage underneath the detector. A fast hardware preprocessor analyzes the input line by line and generates descriptions of the imaged objects that are passed on to a computer for analysis. Although inherently less flexible than systems using TV cameras, this strategy provides very high speeds of data capture and processing in scanning tasks. Line scanning is also used in the Geneti-Scanner *(4).*

3.1.2. Features Used for Metaphase Recognition

Each image that is captured as the stage is moved must be analyzed to detect potential metaphases. A metaphase can be recognized as a region of rather granular image texture of a certain expected size. These regions may be automatically identified in various ways, and since each field needs to be analyzed rather quickly, the method used in any machine tends to exploit the strengths of that machine's particular hardware. However, the image properties that are measured are essentially the same from one system to another. A number of separate objects must be found within a predictable size range and with some optimum separation.

The image analysis problem is to estimate the properties of object number, size, and separation given that individual chromosomes are not well resolved by the approx 1-μm^2 sized pixels typically used for metaphase finding. The method adopted by Graham and Pycock *(12)* and implemented in the Magiscan system was to threshold a suspected metaphase region using a locally determined threshold and to use methods from the field of Mathematical Morphology *(13),* involving erosion and dilation of the binary image, to estimate the sizes and separations of individual touching objects (Fig. 2). Regions are selected for this analysis according to the output of a rapid local texture measurement applied over the entire image and tuned to the variation in brightness expected within the region of a metaphase.

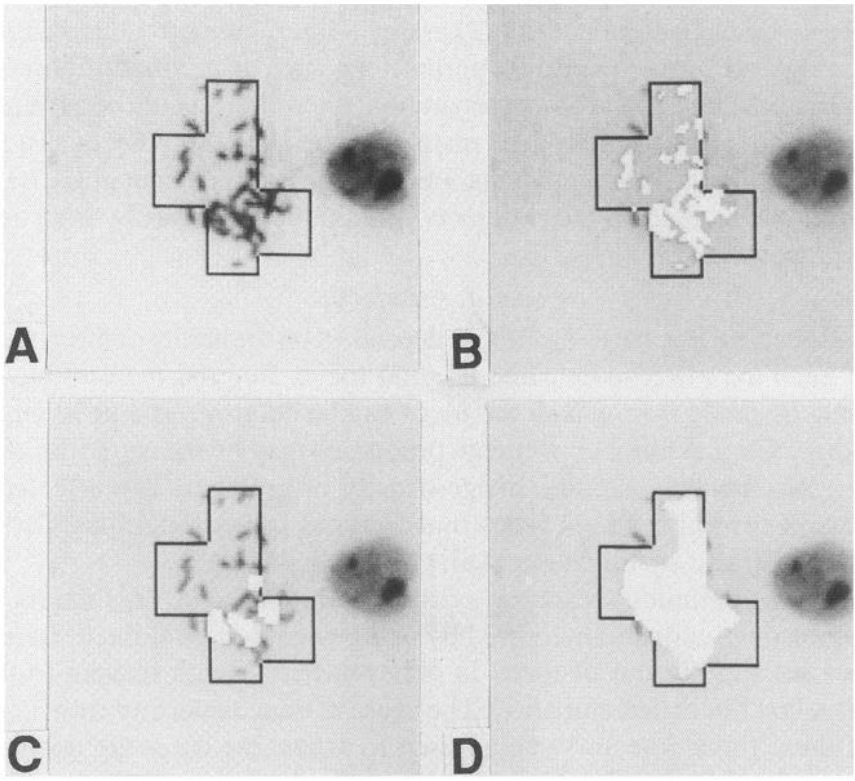


Fig. 2. Analysis of a suspected metaphase using morphological sizing *(12)*. **A.** Extent of an irregular area with image texture appropriate for a metaphase. **B.** Result of applying a locally determined threshold within the area to detect the chromosomes. **C.** The detected regions have been subjected to an opening (erosion + dilation) operation. Most isolated objects have been removed, but clumpy areas remain. Application of this operation with differently sized erosion/dilation structuring elements allows a size distribution of possibly touching objects to be determined. **D.** The detected regions from (B) have been subjected to a closing (dilation + erosion) operation. The areas between the detected regions have been filled in, allowing object separation and overall object size to be measured. (Reproduced by permission of Science Printers and Publishers Inc., St. Louis, MO.)

A method of achieving essentially the same measures appropriate for linear scanning is described by Shippey et al. *(11)*. The clustering in this case uses measured distance between detected "limbs" (simply connected above threshold regions that may touch each other).

Both of these systems assign a quality index to the detected metaphase. Graham and Pycock *(12)* assign a figure of merit consisting of a weighted sum of measured parameters, such as number of objects, image brightness, and object separation, the weights being determined by training on samples of the material to be used. Shippey et al. *(11)* use a similar measure, applying a box classifier (*see* Section 3.6.1.) to eliminate at an early stage objects that are highly unlikely to be useful metaphases.

3.1.3. Autofocus

Reliable metaphase detection is dependent on the ability of the system to keep the microscope slide in good focus. Regular measurements must be made from which the focus can be determined and adjusted if necessary. A number of image properties may be measured for this purpose, such as average image density or gradient. The effectiveness of several different focus functions has been assessed by Groen et al. *(14)* and by Firestone et al. *(15)*.

In the continuous scanning system of Shippey et al. *(11)* the focus can be continuously monitored by using two additional linear detectors set slightly out of focus in either direction with respect to the principal linear detector array. The average image intensity from each of these three detectors can be used to determine the stage position for optimal focus. Usually in TV-based systems, focus is maintained by driving the stage through the focus position at regular intervals in the scan area to find the best value of the focus measure. Graham and Pycock *(12)* use a method of determining a sparse grid of correct focus positions before scanning for metaphases and interpolating between neighboring grid points as the scan proceeds.

3.1.4. Building Your Own

Commercially available metaphase finders are generally based on fairly specialized computers. These may have been designed specifically to optimize scanning processes, such as Cytoscan, or may be machines intended to address a wider range of image processing tasks. In either case, the machines are architecturally rather specialized and therefore expensive.

Up to about the time of writing this chapter, this use of specialized processors was necessary to allow the application of sufficient computing power to achieve realistic scanning speeds. This situation has

changed recently by the introduction of very powerful, inexpensive personal computers and scientific work stations. These machines deliver sufficient performance for image processing, while retaining a general-purpose architecture. This chapter is intended in part to act as a guide to those who wish to build their own systems, and these developments open up the possibility of a do-it-yourself metaphase finder built of commercially available components.

The performance that might be expected from such a system can be judged by the recent implementation of a metaphase finder on a MacIntosh IIfx computer by Vrolijk and his colleagues at the University of Leiden *(16)*. They have used commercially available hardware for image acquisition and microscope control, and achieved scanning speeds of about 360 s/cm^2. The method of detection and assessment of metaphases used by these workers is similar to that described in *(12);* regions detected by a specially designed texture filter are measured and ranked using a combination of mathematical morphology operators. The caveat of our introduction should, however, be particularly emphasized here. To achieve the scanning speeds they report, these workers have used their long experience in this field to develop methods that are not only highly specific to the images, but also highly optimized for the particular computer architecture they chose. Any laboratory considering an *ab initio* implementation of a metaphase finder should bear in mind the investment in time and expertise necessary to build a working system with reasonable performance.

3.2. High-Resolution Image Capture

For karyotype analysis, it is particularly important that the quality of the image captured by the camera is as high as possible. To this end, care should be taken with the optical setup of the microscope lamp and condensor, cleanliness of the lenses, the level of illumination, the filter color (both of the latter may need to be different for the camera than for the human eye), and the overall magnification, taking the camera pixel size into account. As mentioned above, an infrared filter is essential for a CCD camera, and an electrical antialias filter is recommended. Image degradation can be caused by incorrect electrical termination of video signals and also by electrical noise pickup. To reduce radio-frequency noise, it is wise to link all metal components, *including the microscope frame,* to a common earthing point.

A number of factors can be optimized by appropriate image analysis programs. Light level can be measured by identifying the background peak in a histogram of digitized pixel gray values; the lamp should be adjusted so that this peak lies near the "white" end of the frame grabber's range of pixel gray values. Focus is best adjusted by looking at the live image as seen by the frame grabber (since the camera and eyepieces may not be parfocal). Most microscopes show nonuniform illumination, in that they are typically less bright toward the edge of the field; this can be compensated for by taking a "shading map" of a clear field and using it for shading correction, and also to compensate for any camera nonuniformity.

Captured images will comprise a large amount of data, typically between 250 and 500 kbyte. Storing such images will soon fill up whatever disk storage is available. Since most of the image is "background" (clear field), removing background (segmentation; *see* Section 3.) before storing to disk may reduce file sizes by about 90%. Your operating system may provide a file compression program; this can typically reduce the size of the thresholded images by a further 30%.

There will always be some exceptionally well-spread metaphases that do not fit within the camera frame. The solution is to fuse parts of multiple digitized fields under software control; here again, it is sensible to use segmented images as the basis of the operation.

3.3. Metaphase Image Segmentation

3.3.1. Initial Image Segmentation

This is invariably done by thresholding; a darkness value is chosen, and "background" pixels lighter than this threshold value are discarded. Finding the connected sets of darker pixels results in an initial division into image regions that may represent individual chromosomes, chromosome clusters, or unwanted objects, such as stain particles and interphase nuclei.

The threshold may be chosen automatically, by analysis of the histogram of density values (Fig. 3). Such a histogram has a pronounced "background" peak, and the threshold must be chosen at a value a little above the upper end of the background region.

Some authors *(9)* have described methods of local thresholding to take account of a nonuniform background, caused, for example, by slight cytoplasm staining. However, the most usual cause of nonuniform back-

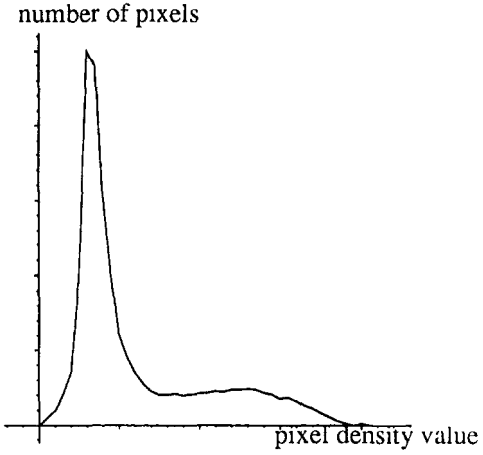


Fig. 3. Histogram of pixel densities from a digitized metaphase image. The pronounced peak comes from the large area of clear background, whereas the chromosomes give rise to the relatively small number of darker pixels.

ground is uneven microscope illumination; this and any camera non-uniformity is best compensated for by preliminary shading correction *(8,17)* (Section 3.2.). In an unpublished experiment, one of us (J. P.) found that local thresholding led to significant deterioration of the per-class coefficients of variation of relative chromosome size compared with simple global thresholding.

In an interactive system, the operator should have control over the final detection level, whether obtained from global or local analysis. As noted in Section 1.5., it is inappropriate for the operator to be asked to specify a threshold numerically. A rapid visual method can easily be provided if the display facilities include a look-up table (LUT). The effect of any particular threshold choice (specified, say, by pointing at a "slider bar") may be simulated by setting LUT values below the proposed threshold to some uniform, non-natural color such as mid-gray or pale blue. The visual effect on the metaphase image is as if the operator had a bird's eye view of the tide rising or falling around a number of islands. If the automatic threshold selection has been done properly, then interactive adjustment should in any case be required only rarely. After setting the threshold, the values of the remaining pixels may be "stretched" to make use of the full dynamic range of the display, resulting in a useful increase in image contrast.

3.3.2. Interactive Segmentation

The user interface has been described above (Section 1.5.2.). From the programming point of view, the requirement is then quite straightforward, namely, to take an image region and some graphical object generated by the interaction system (such as a polygon along the desired split path), and return the image regions of the segmented chromosomes. Further discussion on semiautomatic segmentation can be found in Note 1.

3.4. Feature Measurement

The purpose of feature measurement is to obtain information in numerical form that is useful for classification of a chromosome into its correct class, or to decide that it is abnormal in some way. The features commonly used include relative size, centromeric index, and some numerical description of the banding pattern.

3.4.1. Chromosome Size

To a cytogeneticist, size usually means the length of a chromosome, and the relative length is, of course, an important discriminator of chromosome class. However, it has been found that chromosome area is an equally reliable size measure, which has the advantage of being easy to compute, simply by counting the number of pixels, and in particular, does not depend on correctly locating the chromosome's axis.

3.4.2. Relative Density

Some chromosomes are, overall, paler than others, and the relative density of a chromosome may be obtained simply by adding its pixel values and dividing by the area.

3.4.3. Medial Axis, Orientation, and Polarity

For all the other measurements that we wish to make, whether the chromosome is bent or straight, or whatever its orientation in the metaphase plate is irrelevant, and such geometric variation is compensated for by making all measurements in a non-Euclidean coordinate frame determined by the chromosome's axis of symmetry or centerline. Since one aim of a karyotyping system is to produce a karyogram presentation of the metaphase, the orientation must also be found explicitly, so that the chromosome can be displayed vertically in the karyogram. Finally, it is conventional to display the short arm upper-

most, and so the chromosome polarity (which arm is the short arm) must also be determined.

Although in principle the chromosome orientation could be defined as the average direction of the medial axis, it turns out that finding the orientation is rather more reliable than finding the axis, and indeed, most axis-finding methods rely on an initial estimate of orientation. Possible methods include taking second-order moments of gray values *(18),* finding the least-squares straight line fit to boundary coordinates *(9),* or finding the minimum width enclosing rectangle (MWR) *(7).* The MWR method makes use of the fact that the MWR is parallel to one chord of the convex hull or minimum enclosing convex polygon. Assuming an appropriate data structure for the chromosome image, the convex hull can be found extremely rapidly *(19),* and finding the MWR is then straightforward and rapid.

Axis finding is still an unsolved problem; looked at another way, existing methods are prone to error in a significant number of cases. The consequence of finding a curve that is not in fact the chromosome's symmetry axis is that subsequent measurement of length, centromere position, and banding features may well be erroneous. Since the system under discussion is intended to be automatic, such errors will not be apparent until the chromosome is misclassified and presented in the wrong location in the karyogram. Axis errors are an important cause of classification errors, and as was mentioned in Section 1.5.2., in some systems, the operator is given the opportunity to correct the axis interactively.

Given the orientation, the axis can be found initially as a straight line *(9,18).* Of course, very few chromosomes are truly straight, and a curved axis is usually required. For relatively straight, well-formed chromosomes, the set of midpoints of chords perpendicular to the major orientation direction (Fig. 4) suffices *(7).* In the case of more seriously bent chromosomes, various methods have been proposed: fitting a cubic to the chord midpoints *(9),* a piece-wise linear fit *(18),* or use of the "skeleton" of the chromosome image region *(7).*

As yet, no known method copes well with metaphase chromosomes that have an acute bend at the centromere (or elsewhere), since the true centerline of the chromosome no longer corresponds to the midline of the segmented "shape" of the object. The problem could be solved if it were known *a priori* that the object was an acutely bent chromo-

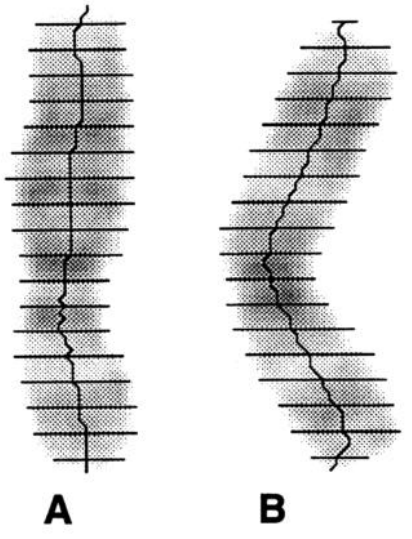


Fig. 4. The axis derived from the midpoints of chords perpendicular to the major orientation of a chromosome is usually satisfactory in the case of relatively straight chromosomes **A**, but can be significantly in error at (particularly) the ends of bent chromosomes **B**. For clarity, only every fourth chord has been shown.

some. However, even visually, such an object can often only be recognized as a single chromosome, because, in effect, its two arms are recognized independently as belonging to the same chromosome class. In an automatic system, recognition follows segmentation and axis finding, and such circular reasoning has not yet been proven possible *(20)*. Similarly, chromosomes whose chromatids are not parallel will typically confuse an axis algorithm. These are illustrations of a general principle that an automated system for biomedical image analysis will only work satisfactorily if care is taken to ensure that the biological preparation is of the highest possible quality and conforms to the system's implicit model of such material.

3.4.4. Profiles

Having obtained the chromosome's symmetry axis, the first stage in obtaining the remaining features is to reduce the two-dimensional chromosome image to a one-dimensional form known as a profile. A profile represents the distribution of some property of the chromo-

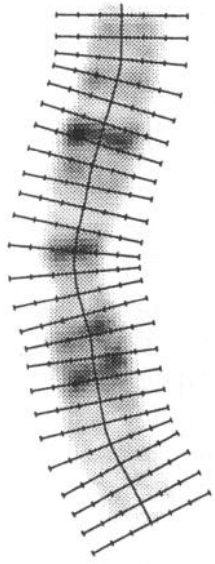


Fig. 5. Non-Euclidean coordinate system for computing profiles, based on lines perpendicular to the chromosome axis. For clarity, the scale used has been reduced by 4×.

some, for example, its width or the intensity of staining, as it varies along the chromosome in a direction determined by the medial axis. More precisely, we define a non-Euclidean coordinate space (Fig. 5) and make measurements at points in this space. The points of interest are obtained in the following way.

First, points are found on the medial axis that are unit distance (one pixel spacing) apart. Note that such points themselves do not usually have integer coordinates. Next, at each such point, a line is constructed perpendicular to the axis, and points are found on each such line at unit distance spacing (Fig. 5). Again, these points do not have integer coordinates, nor are they usually at unit distance from all of their neighbors. At each such point, an appropriate pixel intensity value is computed from the values of the neighboring original pixels. This can be done most simply by finding the nearest original pixel and choosing its value (nearest-neighbor method, Fig. 5), but it has been shown that values obtained by bilinear interpolation among the four surrounding

original pixels (Fig. 5) lead to more accurate measurement *(21),* and this is now the commonly used technique.

Profiles are used both to represent the banding pattern and also to reduce the chromosome's shape to a one-dimensional representation, in order to find the centromere. For band pattern representation, the "integrated density" profile is computed by taking the sum of all pixel values inside the chromosome boundary on each of the transverse lines across the chromosome (Fig. 6). Shape may be represented by the "width" profile (Fig. 6), computed by finding the number of connected pixels with above-threshold values in each transverse slice *(9).* The width may well either be noisy on account of closely adjacent chromosomes or stained cytoplasm, or the width at the centromere may not be significantly less than elsewhere. An alternative profile that partly overcomes these problems by integrating information across the chromatid structure is the moment or "shape" profile *(7,22),* computed as shown in Fig. 7. Shape profiles are compared with width profiles in Fig. 8.

3.4.5. Centromere

Usually, a metacentric chromosome's centromere appears as a pronounced minimum in either the width or shape profile, whereas that of an acrocentric is represented only by a smaller than usual gradient at one end of the profile (Figs. 6 and 8). Since it is not known *a priori* whether a particular chromosome is in fact metacentric or acrocentric, the essential problem to be solved is how to compare the properties of the gradient at an end of the profile with properties of a profile minimum. Various solutions have been proposed, none of which is entirely satisfactory:

1. Groen et al. *(18)* used the width profile, truncated at either end, and chose the overall minimum width. The assumption is that this will be at the correct end in the case of an acrocentric.
2. Graham *(9)* found the end with lower gradient if there was no profile minimum in the central 60% of the width profile.
3. Piper *(22)* took the convex envelope of the profile and found the most "significant" chord. The centromere was then assumed to be at the point furthest beneath this chord.
4. Piper and Granum *(7)* deliberately constructed minima at either end of the profile and then chose the "best" minimum, which in the case of metacentric chromosomes was expected not to be one of those at the end.

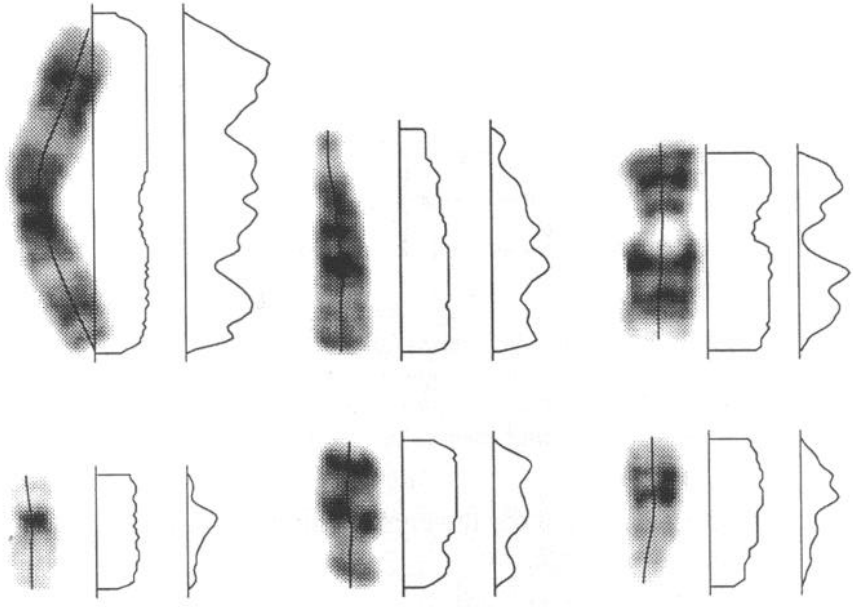


Fig. 6. Chromosome width (left) and density profiles, determined from the central axes shown.

3.4.6. Shape Features

Centromeric index is the most useful shape measure yet discovered. It may be computed on the basis of length, area, or total pixel intensity. Although the former two are highly correlated, the last introduces some additional information about the chromosome. In order to compute the centromeric indices by area and total pixel intensity, the chromosome image must be divided by a line perpendicular to the medial axis that passes through the centromere, and pixels on either side of this line must be determined.

In ref. *7,* some other shape features were proposed, computed by applying Granum's weighted density distribution (WDD) functions *(23)* to the shape profile. However, these tend either to be quite highly correlated with centromeric index or to have rather little class discrimination ability. The centromere position is the best-known automatic determinant of chromosome polarity (which end is the short arm),

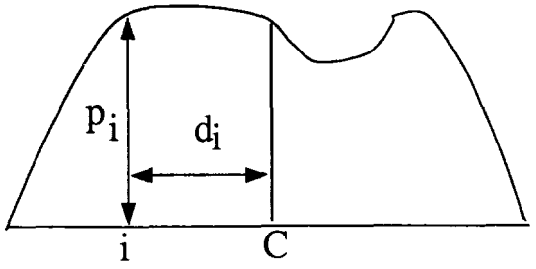


Fig 7 Computation of single point of "shape" profile. The profile shown is of the distribution of pixel values of a single "slice" across the chromosome (Fig. 5), perpendicular to the axis If C is the centroid of the slice, and point i at distance d_i from C has pixel density p_i, then the profile value for this slice is $\Sigma p_i|d_i|\Sigma p_i$.

which in turn is required for the computation of some of the following band pattern features.

3.4.7. Band Pattern Features from Density Profile

Band pattern features divide into two overall classes, "global" and "local." A global feature is a single number computed from the entirety of the density profile by a uniform arithmetical procedure. On the other hand, a local feature describes some particular structure in the density profile, for example, the location of the most intense band. For comparison, the area of a chromosome is also a global feature, whereas the centromeric index is a local feature, determined by the centromere position.

The view has long been held that local band pattern features will be required in order to detect and describe abnormality in banding patterns, and attempts at local band pattern descriptions have been made for more than two decades. However, from the point of view of classification of normal chromosomes into the normal classes, it has been found that the global features are superior. Active research continues in a number of laboratories into local description methods.

3.4.7.1. Global Band Pattern Features

Two approaches are used to obtain single numbers that represent some aspect of an entire density profile. First, the profile values may be treated as samples from a distribution, and parameters of the distri-

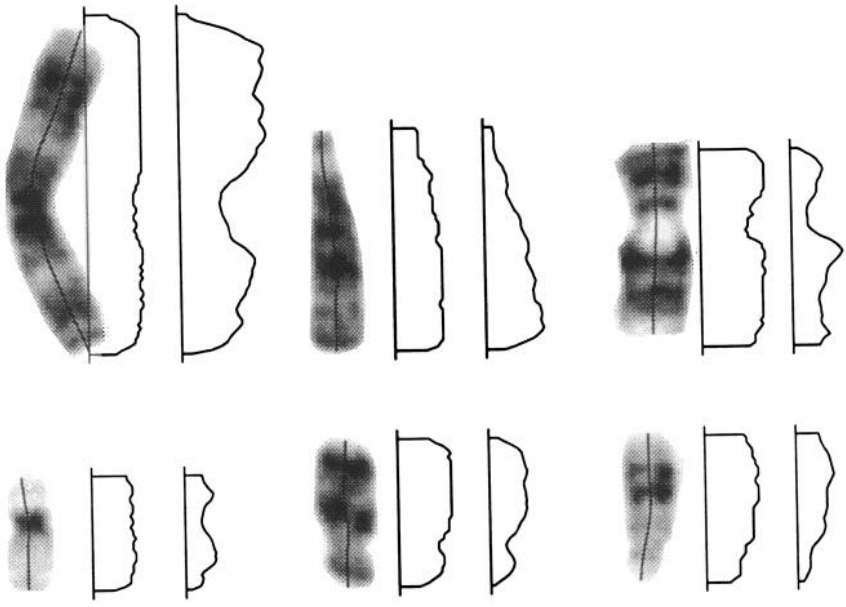


Fig. 8. Width (left) and "shape" profiles of metacentric and acrocentric chromosomes.

bution estimated *(7)*. A variation is that the density differences between adjacent profile points are taken as the sample *(23)*.

The second main method is to multiply the profile by each of a set of basis functions, resulting in a corresponding set of feature values. If appropriate sinusoidal basis functions are used, then the resulting set comprises the Fourier transform of the profile *(24)*. However, the lack until recently of affordable floating point hardware has led to widespread adoption of Granum's WDD basis functions, which are triangular rather than sinusoidal *(23)*. Typically, the first four functions are used. Granum *(23)* recommends the use of these functions both on the entire profile, and also on the p and q portions separately in order to obtain further features. The usefulness of the latter may, however, not be great in a system in which the machine-found centromere is left uncorrected. Piper and Granum *(7)* showed that two additional WDD functions (Fig. 9) were also valuable. They also showed that if the WDD basis functions were applied either to the

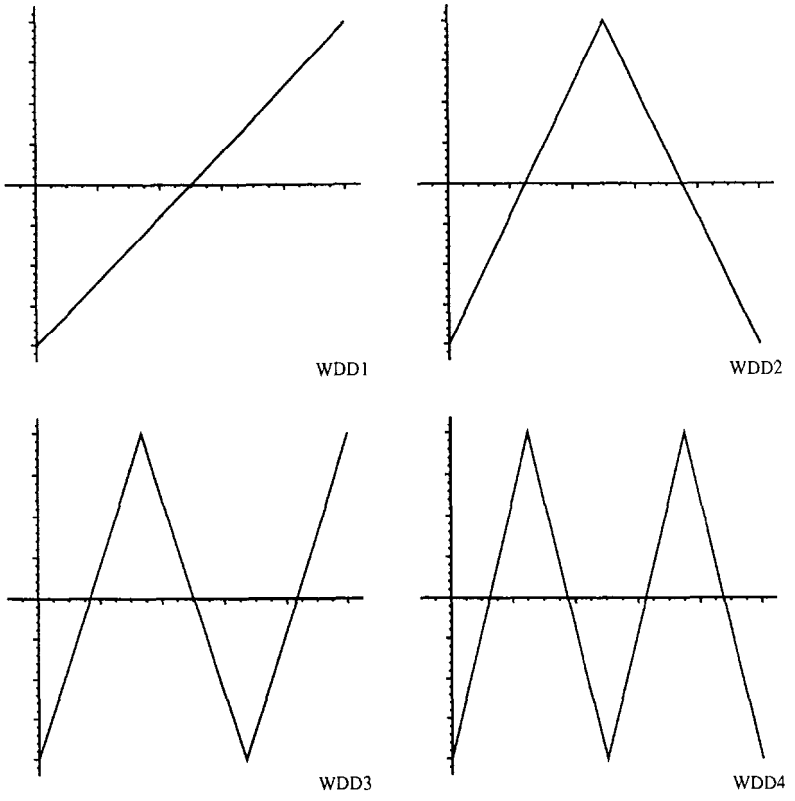


Fig. 9. The first four weighted density distribution functions *(23)*.

"shape" profile or to the profile of differences between adjacent points of the density profile, then further valuable features resulted.

3.4.7.2. Local Band Pattern Description

By analogy with the centromere, it is possible to use the location of certain landmark bands as classification features. Thus, the locations of the darkest band in either arm, and of the dark bands nearest to the centromere and to the telomeres provide usable features *(17,18)*. Thus far, such features have been used with a conventional statistical classifier, but together with a structural/syntactic classifier, they could well provide the means of detecting some band pattern abnormalities. A more rigorous approach to addressing that problem is provided by the

work of Granum and his associates *(25)* in the use of hidden Markov models for chromosome band patterns; so far, however, they have only presented data from normal material.

3.5. Feature Normalization

In a prometaphase cell, a number 1 chromosome may be two or three times longer than a number 1 in a midmetaphase. Indeed, the midmetaphase number 1 may be shorter than, say, a chromosome 10 in the prometaphase cell. Thus, "raw" size is not a very helpful measure. What remains true is that in either cell the length of a number 1 comprises about 8% of the length of a haploid set.

The transformation to relative size by, perhaps, division by the sum of the sizes of all chromosomes in a cell is an example of a process known as normalization *(7,9,18,23)*. Unfortunately, for the majority of features except size, there is no clear theoretical reason to justify normalization or to guide how to perform it. Centromere index varies only slightly with cell contraction and is usually left unnormalized. Other features may be normalized by standardizing the distribution within a cell (i.e., obtaining zero mean and unity standard deviation) *(7)*, and in many cases, this appears to improve their discrimination capability, but the reason for this improvement is not well understood in most cases.

3.6. Classification

In Sections 3.4. and 3.5., we described ways in which attributes of chromosomes may be measured and represented numerically by an automated karyotyping system. Here we describe how these measurements can be used to produce an automatic classification. This subject is often referred to as Pattern Recognition. We will begin with a brief introduction to Pattern Recognition methods and then consider how these are applied to chromosome classification.

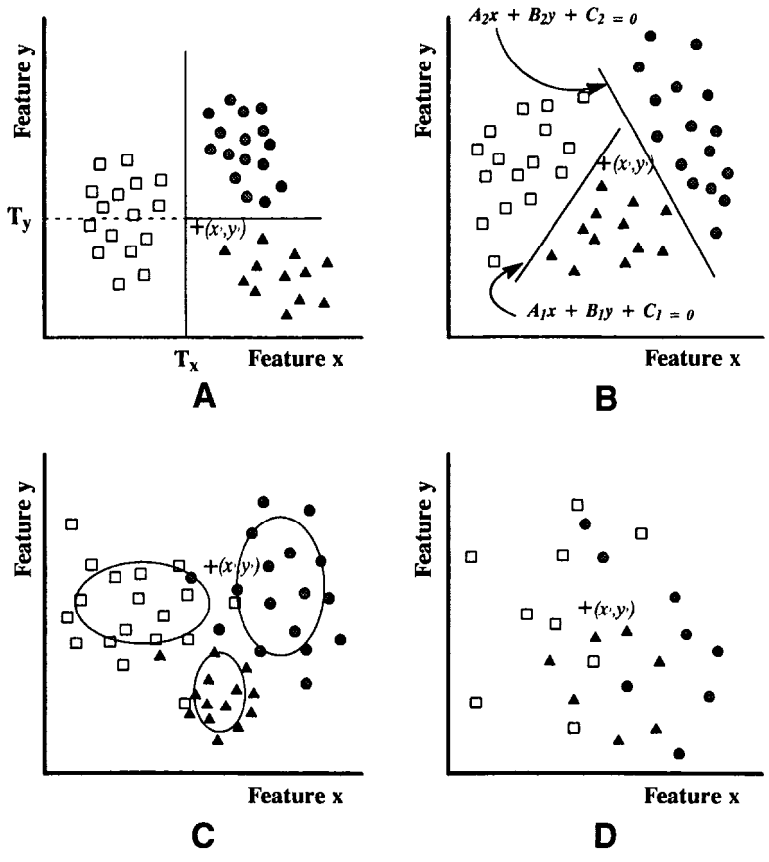
3.6.1. Pattern Recognition

This is a necessarily brief introduction to a highly developed subject, and the reader is referred to one of a number of excellent texts for a full account (e.g., *26,27*). In any classification or pattern recognition task, objects are to be assigned to classes on the basis of a set of measured attributes, or features, such as those described in Sec-

tion 3.4. Features should be numeric (i.e., they must be measurements), they should have some discriminating power (i.e., the features measured from objects in different classes should be different), and they should ideally be independent (i.e., they should represent truly different properties of the classes).

Consider a case where we measure two features x and y (say chromosome length and centromeric index, as used for the classification of unbanded chromosomes), and we wish to distinguish three classes. There are a number of methods that may be applied to discriminate the classes on the basis of these features. To determine which method to choose and the appropriate parameters, features must be collected from a representative set of objects, and these objects assigned to their classes by an expert in the appropriate domain (a cytogeneticist in our case), in a process known as "classifier training." We can plot the feature values on a space whose axes correspond to the features (Fig. 10). This space is known as a *feature space.* The feature values (x,y) for a particular object constitute its *feature vector.* The task of the classifier is to partition the feature space, so that when the feature vector for a new (unknown) object is plotted, it can be assigned to one of the classes. Figure 10 shows some common methods of partitioning the space.

Fig 10 *(opposite page)* Scattergrams representing the two-dimensional feature vectors for a number of training examples of three classes. Depending on the distribution of these vectors in feature space, different strategies may be applied for allocating an unknown object with feature vector (x',y') to a class **A.** Box classifier. The classes can be distinguished by thresholds on each of the feature axes The thresholds may define upper and lower bounds for class membership. In the example shown, a single threshold on each of the axes is sufficient to partition the feature space. **B.** Linear discriminant. In this case, the classes would be poorly discriminated by thresholds, but can be easily separated by straight lines in feature space. An unknown vector is assigned to one of the classes according to its position with respect to these lines, i e , according to the value of $A_1x' + B_1y' + C_1$ and $A_2x' + B_2y' + C_2$ **C.** Parametric classifier. The feature vectors of the training set form recognizable clusters that may overlap and so cannot be separated by a decision line. The clusters may be modelled by a bivariate (in general, multivariate) normal distribution The ellipses represent the area contained within one standard deviation of the mean values of these distributions. An unknown feature vector is assigned to the class according to its distance from each of the class means, the distance being weighted by the covariance of the appropriate class. **D.** Nearest neighbor classifier. If the feature vectors in the training data are few in number, or not well clustered, a new vector may be assigned to the class of the nearest vector in the training data



3.6.1.1. Box Classifier

This is the simplest form of partition. "Boxes" are defined around each cluster by setting thresholds on the feature values. A new object will be classified according to which "box" its feature vector falls into. The advantages of a box classifier are its simplicity and ease of computation. Its principal disadvantage is that it requires the clusters to be well separated along at least one of the dimensions of feature space. This is rarely the case.

3.6.1.2. Linear Discriminant

If classes cannot be separated by thresholds on the features, they may be separated by a straight line in feature space. If the line $Ax + By + C = 0$ separates the classes, then a measure of $Ax_1 + By_1 + C$ for some new

object with feature vector (x_1,y_1) will be greater than zero for (x_1,y_1) on one side of the line and less than zero on the other. If more than two classes are present, several lines can be used.

3.6.1.3. Parametric Classifier

In this case, the clusters arising from the classes are represented in some parametric form. Commonly, the clusters are modeled by multivariate normal distributions (bivariate in the two-dimensional case considered here). The classes are parameterized by the mean and variance of the corresponding clusters along each of the dimensions. A new object may be assigned to one of the classes by measuring its distance from the mean vector of each class, weighted by the variance of that class. In general, this measure can be used to assign a likelihood of a new object belonging to each of the classes. The likelihood that an object with feature vector (x,y) belongs to class i is given by:

$$[1/2\pi(\sigma_{xi}^2 + \sigma_{yi}^2)^{1/2} \exp -^1/_2] \{[(x - \mu_{xi})^2/\sigma_{xi}^2] + [(y - \mu_{yi})^2/\sigma_{yi}^2]\} \quad (1)$$

Since, in general, the object will be assigned to the class with highest likelihood, this method is often referred to as a maximum likelihood classifier.

As described above and in Fig. 10, the features are assumed to be independent. This assumption is usually unrealistic and can be dispensed with by including the covariances in the calculation of distances. This, however, adds greatly to the computational cost, but has been shown to add little to the overall classification accuracy, at least for classification of chromosomes *(28,29)*.

3.6.1.4. Nearest Neighbor Classifier

If the feature vectors do not fall into compact clusters or the training set is very small, none of the above methods may be suitable. In this case, an unknown feature vector may be assigned to the class of the nearest object in the training set or, more robustly, to the class to which the majority of the k nearest objects belong (a k-NN classifier). The k-NN classifier can be highly effective, its principal disadvantage being that the entire set of training data must be available and searched for each classification.

3.6.2. Classifying Chromosomes

3.6.2.1. Classifier Capability

Karyotyping is a rather challenging classification problem. There are few other applications in which it is necessary to assign objects to as many as 24 classes. In current implementations, the assignment may be based on anything from five to 16 features *(7,9,18);* that is to say that the two-dimensional feature spaces shown in Fig. 10 should be visualized in up to 16 dimensions. The topic of automatic selection of suitable features is considered in Note 2. To our knowledge, all currently available karyotyping systems use maximum likelihood classification. In the case of the Magiscan and Cytoscan, the likelihood estimates are based on a parameterization of the observed distribution of the training set as in Fig. 10 C. In the AKS-2 system, the probability density functions are estimated directly from the training set, rather than being expressed as a small number of parameters *(18).*

Because of a number of factors, perfect chromosome classification is never achieved in practice. Errors in feature measurement, inadequacies in the feature sets used, and imperfect separation of the classes in feature space result in misclassification rates on the order of 5–20% for routine quality preparations of material used for clinical karyotyping *(7,30,31).* We should be a little careful about what we mean by classification error rates. The percentage misclassifications just quoted give us a measure, over several different studies, of the total number of chromosomes not assigned to their correct classes by an automatic classifier. However, if we examine classifier performance from a system point of view, this is not necessarily the most meaningful number we could derive. In a working karyotyping system, a classification error will be corrected interactively by the operator. This is much more easily done if the misclassified chromosomes can be identified and placed together in a special group, rather than being assigned to the wrong groups and scattered all over the karyogram. This can be achieved to some extent by the inclusion of a "reject class" to which chromosomes are assigned if their likelihood of belonging to any real group is below some threshold. There will, of course, remain a number of chromosomes that are wrongly classified with high likelihood, and these are the most serious classifica-

tion errors from a system standpoint. Lundsteen et al. *(30)* find that their residual misclassification rate is reduced from 9.0 to 5.5% by the inclusion of a reject class. We should, of course, be somewhat circumspect about this. We could easily achieve near-zero error rates by setting the likelihood threshold high enough at the expense of rejecting most of the chromosomes.

3.6.2.2. Constraints on Chromosome Classification

The classification rates reported above refer to context-free classification; that is to say, the class to which a chromosome is assigned depends entirely on the features measured for that chromosome, with no account being taken of the features or classification likelihoods of other chromosomes in the cell (other than by normalization; Section 3.5.). The fact that almost all chromosomes in a cell will be paired as homologs of very similar appearance provides constraints on classification that can be exploited to improve the overall classification performance. (This is of use in most cases, although when karyotyping cells from bone marrow or solid tumors it is often the case that not all chromosomes in a cell are visible or the cell is in any case highly aneuploid, and this type of constraint is of limited use.)

The most useful constraint arising from this source is the knowledge that no class should contain more than two chromosomes (except in the rare, but important case of numerical abnormality). Context-free classification may well assign more than two chromosomes to the same class on the basis of their maximum likelihoods. The classification must therefore be "rearranged" by assigning some chromosomes to "second choice" classes, possibly displacing other chromosomes from these classes, and so on. Piper *(32)* tested a number of algorithms for achieving such a rearrangement and showed that the overall classification rate can indeed be improved in this way. All of the methods improve classification, but none guarantee an optimal rearrangement in the sense that the maximum overall likelihood is obtained subject to the constraint. Tso and Graham *(33)* showed that an optimal assignment of chromosomes to classes can be achieved using a method derived from Operations Research. An efficient algorithm for performing this assignment has recently been described and tested *(34)*. Application of this technique results in a further small improvement in assignment over the suboptimal methods in ref. *32*. It should be

noted, however, that the overall performance improvement obtained by the use of such rearrangement algorithms is small, of the order of 2–3% in a misclassification rate of 5–20% *(32,34)*. This change alone would almost certainly pass unnoticed by the user of a karyotyping system. The methods do, however, result in an increase in the number of cells in which no misclassifications occur or in which there is only one misclassification. These cells will be exactly those of high visual quality, which should be specifically selected by a competent metaphase finder. Thus, the effects on the efficiency of the system will be greater than might be expected from consideration of misclassification rates alone.

Some consideration has also been given to the fact that two homologous chromosomes assigned to the same class should look similar. Zimmerman et al. *(35)* showed that chromosomes could be matched to their homologs with high accuracy. Recently, this constraint has been incorporated into a classifier with promising results, but the high computational cost makes it unsuitable at present for inclusion in a practical automated karyotyping system *(36)*.

Karyotyping, as we have said, is a difficult classification task, and looked at in isolation, existing chromosome classifiers do not do particularly well by pattern recognition standards, but in combination with other system components, they can contribute to useful semiautomatic cytogenetic analysis systems. Some further aspects of karyotyping classification are considered in Notes 3 and 4.

4. Notes

1. Semiautomatic segmentation: Although the purely interactive segmentation methods described in Sections 1.5.2. and 3.3.2. are adequate, at least some automation is highly desirable, particularly when analyzing difficult material, such as bone marrow preparations. Fully automatic segmentation is still an area of active research, and existing techniques are both not particularly accurate and extremely complex, so that only the barest outline will be given here, since a detailed review would comprise a long chapter all by itself. However, even rather straightforward automatic strategies can assist considerably, and a few of these will be mentioned here. The problem may be regarded as consisting of two stages, (a) recognition of objects, i.e., individual chromosomes, clusters, nuclei, or other nonchromosome material, and (b) resolution of clusters. Since nonchromosomal material may be involved in clus-

ters, the two stages may iterate. The operator must still be allowed to review and change a machine decision, or initiate an action that the machine has not suggested.

In some material, there is a large amount of particulate debris that appears in the image as many small spots. These may be highlighted in some way and classified as noise, unless the operator cancels the highlighting. The operator's involvement remains essential, since frequently the satellites of a D-group chromosome will have been separated by the initial segmentation by thresholding. Similarly, large objects cannot possibly be isolated chromosomes. If these are also highlighted and the operator prompted that something must be done about them, then the scope for missing some necessary decisions is reduced.

There are two different possible approaches to the resolution of clusters. Either the cluster detection and segmentation can both be automatic, with operator review of the final result *(10,37,38)*, or the operator can be relied on to point at an object, which is then resolved automatically *(39,40)*. In either case, the system will make errors in a substantial proportion of cases (usually fewer in "better" material), so some means of highlighting machine actions is important so that the operator can review and correct the decisions.

Automatic cluster recognition and decomposition can be based on one of two techniques. Graham thresholded the image at two levels, and found an optimal grouping of the higher-threshold particles by a region-based split and merge technique that is guided by the expected position of chromosomes obtained by a previous count-by-pointing phase *(10)*.

The other approach to cluster recognition and segmentation is largely based on an analysis of the shape of the boundary compared with the expected shape of a single chromosome. Those boundaries with complex curvature are most likely to belong to clusters, and concavities on the boundary are likely end points for a path that separates the cluster. Such an analysis may be used both for cluster recognition *(37,38,41)* and for splitting *(39,40)*. The more successful systems construct splitting paths by "valley following" (regarding pixel value as topographical height) and compare several potential splitting paths, choosing the most probable on the basis of a set of measured features *(39,41)*.

2. Feature selection for classification: Not all features have the same discriminating power, and in some circumstances, a feature may contribute little or nothing. Consider, for example, the case of the band-pattern features if classifying homogeneously stained chromosomes. In such cases, inclusion of a feature may make the classification error rate higher,

whereas omitting it will also reduce the computational cost. Thus, it makes sense to evaluate the discriminating power that each feature contributes, in order to select a useful subset.

Methods for automatic feature selection for chromosome classification are discussed and illustrated in refs. *7,23,* and *28.* Essentially, these depend on both the discrimination capability of the feature taken alone, which may be estimated from the classifier training data by a simple formula, and on the correlation between a particular feature and others already selected as "useful," since the inclusion of an additional feature that is highly correlated with one already selected will most likely contribute little.

Alternatively, one can run full classification experiments with the training data (split into separate "training" and "test" sets *[7]*), using a variety of different, but plausible sets of features, and choose the best on the basis of minimum error rate. However, to do this thoroughly is extremely (computer) time-consuming, since the number of possible feature sets explodes combinatorially as the number of features increases.

3. Future classifier developments:

Classifier design: Classifiers and their associated feature sets in current systems were designed for use with fairly contracted metaphase chromosomes. The tendency in cytogenetics laboratories to analyze cells with longer chromosomes, showing more bands, may result in decreasing performance of these classifiers. Two areas of development in classifier design address this problem. Granum and his coworkers *(25)* have developed an approach to chromosome classification based on syntactic analysis, that is a structural description of the banding pattern, which is in principle extendible to high resolution banding. The application of an artificial neural network to chromosome classification has been described by Errington and Graham *(42).* Classification performance is similar to that which can be obtained using statistical classifiers, and the flexibility inherent in neural networks should make them easily adaptable to changes in the appearance of cells being analyzed.

Automatic detection of abnormalities: Another notable feature of the current generation of classifiers is that they are designed only to classify normal cells. In the main, they have no inherent definition of any specific abnormality. At first sight, this appears curious, since the object of karyotyping is to identify specific abnormalities. It can be understood by noting that a cytogeneticist spends most of his/her time examining normal cells. Since karyotyping systems are intended to be used interactively, a normal cell classifier is a useful system component.

This is not to say that a classifier capable of recognizing abnormalities is undesirable, but only that its design would be more difficult and that little work has been done in this area. Lundsteen et al. *(43)* discuss some possible ways of approaching this and include a description of a small pilot study suggesting that analysis of the feature values for banded chromosomes in a corrected karyotype could provide an indication of the existence of abnormalities that are difficult to perceive by eye. Carothers et al. *(44)* showed that it is theoretically possible in a multiple-cell karyotyping system (*see* Note 4 *below*) to detect numerical abnormalities completely automatically by processing 16–32 cells from a particular specimen, assuming the use of a classifier whose performance is not very different from those in current systems. If enough cells are analyzed, an aneuploidy should be detectable above the noise of the system error rate. The use of such a system would require a regime in which tests for aneuploidies would be carried out separately from analysis for structural abnormalities. This would involve a radical change in the normal practice of most cytogenetic laboratories.

4. Multiple cell karyotyping: It has been suggested several times during the development of automated cytogenetics systems that multiple-cell karyotyping would be a highly useful application of computer technology. The original idea, in the era when chromosome analysis research aimed at complete automation, was that the final karyotype (description of the chromosome complement) would be produced in a statistical fashion from several cells *(45,46)*. Alternatively, it has been proposed that the chromosomes in each group from all cells be displayed together *(47)*. In either case, the aim is not only to provide a useful means of presenting the information from several cells, but also to reduce the need for operator interaction. Those analysis errors resulting in incorrect segmentations or wrongly positioned axes or centromeres would simply be ignored, and result in objects that would be either wrongly classified or impossible to classify. Provided the number of these errors is fairly small, there should be enough examples of correctly identified chromosomes in each class to generate a karyotype.

   Carothers et al. *(44)* have specified the conditions under which a fully automatic system could be successful. Lundsteen et al. *(43)* have shown that a multiple-cell karyotyping display system can be made to work, in which the only interactions required are the initial interactive count for each cell and an inspection of the final composite karyogram. The interesting point here is that all the interactions involve the operator understanding the chromosomes. This kind of interactive facility is

not yet offered by any of the suppliers of karyotyping systems, but is well within the capability of current technology. Its introduction would require a change in the working practice of cytogenetics laboratories, which may not be possible until there is a wider acceptance of machine-assisted karyotyping.

References

1 Piper J. and Rutovitz D. (1985) Data structures for image processing in a C language and Unix environment *Patt. Recog. Letts.* **3,** 119–129.

2. Graham J., Taylor, C J., Cooper, D H , and Dixon, R N (1986) A compact set of image processing primitives and their role in a successful application program *Patt Recog Letts* **4,** 325–333.

3. Korthof, G. and Carothers, A. D (1991) Tests of performance of four automatic metaphase finding and karyotyping systems *Clin Genet.* **40,** 441–451.

4. Castleman, K. R (1992) The PSI automatic metaphase finder. *J. Radiation Research,* **33 (suppl.),** 124–128.

5. Martin, A. O. (1985) My life with two automated systems. Automated Chromosome Analysis Workshop, Leiden

6. Daker, M (1985) The detection of chromosome abnormalities using Magiscan 2 Automated Chromosome Analysis Workshop, Leiden

7. Piper, J and Granum, E. (1989) On fully automatic feature measurement for banded chromosome classification. *Cytometry* **10,** 242–255.

8. Lloyd, D., Piper, J., Rutovitz, D , and Shippey, G. (1987) Multiprocessing interval processor for automated cytogenetics. *Appl. Optics* **26,** 3356–3366.

9 Graham, J (1987) Automation of routine clinical chromosome analysis I Karyotyping by machine *Anal Quant. Cytol. Histol.* **9,** 383–390.

10 Graham, J (1989) Resolution of composites in interactive karyotyping, in (Lundsteen, C and Piper, J , eds.) *Automation of Cytogenetics* Springer-Verlag, Berlin, pp 191–203

11. Shippey, G , Carothers, A. D., and Gordon, J. (1986) Operation and performance of an automatic metaphase finder based on the MRC fast interval processor *J Histochem. Cytochem.* **34,** 1245–1252.

12 Graham, J. and Pycock, D. (1987) Automation of routine clinical chromosome analysis. II Metaphase finding *Anal. Quant. Cytol Histol.* **9,** 391–397.

13 Serra, J (1982) *Image Analysis and Mathematical Morphology.* Academic, London, UK

14. Groen, F C A , Young, I T , and Ligthart, G (1985) A comparison of different focus functions for use in autofocus algorithms *Cytometry* **6,** 81–91.

15 Firestone, L., Cook, K., Culp, K., Talsania, N., and Preston, K. (1991) Comparison of autofocus methods for automated microscopy *Cytometry* **12,** 195–206

16 Vrolijk, J., Sloos, W C. R., Verwoerd, N. P., and Tanke, H. J. (1991) A MacIntosh based system for metaphase finding. Poster contribution to EC Concerted Action on Automated Cytogenetics Workshop, Leiden, September 2–3.

17. van Vliet, L J , Young, I T , and Mayall, B H (1990) The Athena semi-automated karyotyping system *Cytometry* **1,** 51–58

18. Groen, F. C A., ten Kate, T. K , Smeulders, A. W. M , and Young, I. T. (1989) Human chromosome classification based on local band descriptors *Patt Recog. Letts.* **9,** 211–222
19 Rutovitz, D. (1975) An algorithm for in-line generation of a convex cover *Computer Graphics Image Processing* **4,** 74–78.
20. Piper, J. and Lundsteen, C. (1987) Human chromosome analysis by machine. *Trends in Genet* **3,** 309–313.
21. Groen, F. C., Verbeek, P. W , Zee, G. A., and Oosterlinck, A. (1976) Some aspects concerning computation of chromosome banding profiles, in *Proceedings of the 3rd International Joint Conference on Pattern Recognition,* Coronado, CA, pp 547–550.
22. Piper, J. (1981) Finding chromosome centromeres using boundary and density information, in *Digital Image Processing.* (Simon, J -C and Haralick, R M , eds.) D. Reidel, Dordrecht, Netherlands, pp 511–518
23. Granum, E. (1982) Application of statistical and syntactical methods of analysis and classification to chromosome data, in *Pattern Recognition Theory and Applications.* (Kittler, J., Fu, K. S., and Pau, L F eds) NATO ASI (Oxford 1981), Reidel, Dordrecht, pp. 373–398.
24. Caspersson, T., Lomakka, G., and Moler, A. (1971) Computerised chromosome identification by aid of the quinacrine mustard fluorescence technique *Hereditas* **67,** 103–109.
25. Thomason, M. G and Granum, E (1986) Dynamic programming inference of markov networks from finite sets of sample strings. *IEEE-Transactions on Pattern Analysis and Machine Intelligence (PAMI)* **8,** 491–501.
26. Devijver, P A. and Kittler, J (1982) *Pattern Recognition, A Statistical Approach* Prentice-Hall International, London, UK.
27. Duda, R. O. and Hart, P. E. (1973) *Pattern Recognition and Scene Analysis.* Wiley, New York
28. Piper, J. (1987) The effect of zero feature correlation assumption on maximum likelihood based classification of chromosomes *Sign Proces.* **12,** 49–57.
29. Kirby, S. P. J., Theobald, C. M., Piper, J, and Carothers, A. (1991) Some methods of combining class information in multivariate normal discrimination for the classification of human chromosomes. *Statistics in Medicine* **10,** 141–149.
30. Lundsteen, C., Gerdes, T., and Maahr, J (1986) Automatic classification of chromosomes as part of a routine system for clinical analysis. *Cytometry* **7,** 1–7.
31. Piper, J. (1992) Variability and bias in experimentally measured classifier error rates. *Patt. Recog. Letts* **13,** 685–692
32 Piper, J. (1986) Classification of chromosomes constrained by expected class size. *Patt. Recog. Letts.* **4,** 391–395.
33 Tso, M. K S. and Graham, J. (1983) The transportation algorithm as an aid to chromosome classification. *Patt Recog Letts.* **1,** 489–496
34 Tso, M , Kleinschmidt, P , Mittereiter, I , and Graham, J. (1991) An efficient transportation algorithm for automatic chromosome karyotyping. *Patt. Recog. Letts.* **12,** 117–126.

35 Zimmerman, S. O., Johnston, D A , Arrighi, F. E., and Rupp, M E (1986) Automated homologue matching of human G-banded chromosomes *Comput Biol. Med* **16,** 223–233.
36 Piper, J , Carothers, A., and Guest, E (1991) Chromosome classification incorporating similarity constraints. *Digest of the World Congress on Medical Physics and Biomedical Engineering, Kyoto, Japan, 1991, Medical and Biological Engineering and Computing* **29 (suppl),** 221.
37. Wu, Q., Snellings, J , Amory, L., Suetens, P., and Oosterlinck, A (1989) Model-based contour analysis in a chromosome segmentation system, in *Automation of Cytogenetics,* (Lundsteen, C. and Piper, J , eds ,) Springer-Verlag, Berlin, pp. 217–229.
38 Vossepoel, A M. (1989) Separation of touching chromosomes in *Automation of Cytogenetics,* (Lundsteen, C. and Piper, J , eds) Springer-Verlag, Berlin, pp. 205–216.
39 Ji, L (1989) Intelligent splitting in the chromosome domain *Patt. Recog* **22,** 519–532
40 Ji, L. (1989) Decomposition of overlapping chromosomes, in *Automation of Cytogenetics,* (Lundsteen, C. and Piper, J., eds) Springer-Verlag, Berlin, pp 177–190
41. Ji, L (1991) Fully automatic chromosome segmentation. *Cytometry,* in press
42. Errington, P. A and Graham, J (1993) Application of artificial neural networks to chromosome classification *Cytometry,* in press.
43. Lundsteen, C , Gerdes, T , and Maahr, J. (1989) Cytogenetic analysis by automatic multiple cell karyotyping, in *Automation of Cytogenetics* (Lundsteen, C. and Piper, J., eds) Springer-Verlag, Berlin, pp. 263–274.
44. Carothers, A. D., Rutovitz, D. and Granum, E. (1983) An efficient multiple-cell approach to automatic aneuploidy screening. *Anal. Quant Cytol.* **5,** 194–200.
45. Granlund, G. H., Zack, G. W., Young, I. T , and Eden, M (1976) A technique for multiple-cell chromosome karyotyping. *J Histochem. Cytochem.* **24,** 160–167.
46. Hilditch, C J. (1969) The principles of a software system for karyotype analysis, in *Human Population Cytogenetics* (Jacobs, P. A., Price, W. H , and Law, P., eds.) Edinburgh University Press, Edinburgh, Scotland, pp 297–325.
47. Lundsteen, C. (1978) A proposed format for system output: the combined karyotype *Proceedings of the 1978 European Workshop on Automated Human Cytogenetics* Electronics Laboratory, Technical University of Denmark, Lyngby, pp. s3.3,s3 4.

13. **A neural network approach to automatic chromosome classification.** A.M. Jennings and J. Graham, *Phys. Med. Biol 38: 959-970, 1993*. doi:10.1088/0031-9155/38/7/006

Phys. Med. Biol. **38** (1993) 959–970. Printed in the UK

A neural network approach to automatic chromosome classification

Anne M Jennings and Jim Graham
Department of Medical Biophysics, University of Manchester, UK

Received 5 October 1992, in final form 1 April 1993

Abstract. Classification of banded metaphase chromosomes is an important step in automated clinical chromosome analysis. We have conducted a preliminary investigation of the application of artificial neural networks to this process, making use of a natural representation of the banding pattern. Two different network architectures have been compared: the Kohonen self-organizing feature map and the multi-layer perceptron (MLP). For each of these a search of their respective parameter spaces over a limited range has resulted in configurations of modest dimension which achieve creditable classification rates. The MLP in particular shows promise of being a useful classifier. When size and shape features are supplied as inputs to the MLP in addition to a low-resolution banding profile, misclassification rates are obtained which are comparable with those of a well developed statistical classifier.

1. Introduction

In a normal human cell there are 46 chromosomes which, at an appropriate stage of cell division (metaphase), can be observed as separate objects using high-resolution light microscopy. Appropriately stained they show a series of bands along their length and a characteristic constriction called the centromere. Figure 1(*a*) shows a typical metaphase cell, stained to produce the most commonly used banding appearance (G banding). Chromosome analysis, which involves visual examination of these cells, is routinely undertaken in hospital laboratories, for example for pre-natal diagnosis of genetic abnormality or monitoring of cancer treatment.

This visual analysis involves counting the chromosomes and examining them for structural abnormalities. To determine the significance of both numerical and structural abnormality it is necessary to classify the chromosomes into 24 groups on the basis of their size, the pattern of bands and the centromere position. Twenty two of these groups normally contain two homologous (structurally identical) chromosomes. The other two groups contain the sex chromosomes X and Y. In the case of a normal male cell, the X and Y groups contain one chromosome each; in a female cell there is a homologous pair of X chromosomes and the Y group is empty.

The time consuming nature of chromosome analysis has resulted in considerable interest in the development of automated systems based on machine vision. A number of such systems are now in routine use in many hospitals (see, e.g. Graham 1987, Graham and Pycock 1987; for a review see Lundsteen and Martin 1989). The chromosome classification performance of these systems depends on the type of material used, but is at best in the range 6–18% misclassification (Piper and Granum 1989). This compares poorly with visual classification by cytotechnician, which was estimated by Lundsteen etal (1976) to result in

0031-9155/93/070959+12$07.50 © 1993 IOP Publishing Ltd

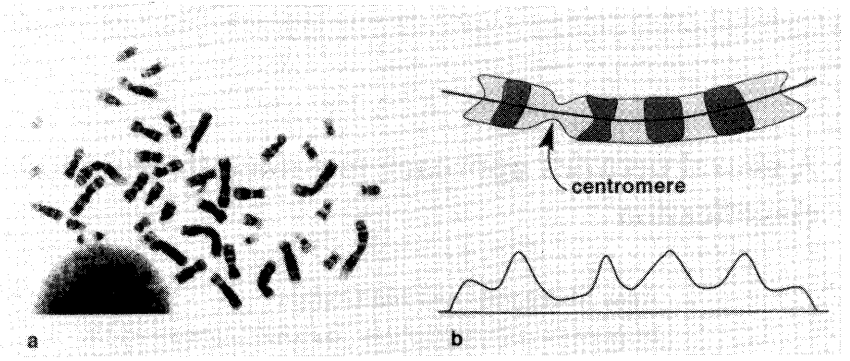


Figure 1. Chromosomes and chromosome features. (*a*) A cell at metaphase. The individual chromosomes show the banding pattern (G banding) produced by staining. (*b*) Schematic drawing of a chromosome showing the position of the centromere. The density profile (below) is formed by projecting the density onto the curved centreline.

a classification error rate of 3% for inspection of isolated chromosomes, dropping to 0.1% when all chromosomes in a cell could be examined together. All automated systems in clinical use operate interactively, allowing an expert operator to correct machine errors in image segmentation, feature extraction and classification, resulting in useful performance (Graham and Piper 1993). However, there is clear scope for improvement in automatic classification.

An important issue for automatic classification is the representation of the banding pattern. Several different classifiers have been reported using statistical or syntactic approaches (Granlund 1976, Granum 1982, Lundsteen *et al* 1981, Groen *et al* 1989, Thomason and Granum 1986). Each of these involves the extraction of a number of intuitively defined features, usually associated with the chromosome's density profile. The density profile is a one-dimensional pattern obtained by projecting the chromosome's density onto its centre line (figure 1(*b*)), and reflects the largely linear organization of the chromosome structure.

The processing involved in extracting features from the profiles involves the risk of losing information, a risk which may be eliminated by using the density profile itself as the banding representation. This type of one-dimensional pattern is a natural form of input for artificial neural networks, in which the classification features are selected automatically by a process of training from examples. Neural networks have been successfully used for a range of pattern recognition applications (Kohonen 1988, Sejnowski and Rosenburg 1987, Haykin and Deng 1991). Their advantages lie in their ability to correctly classify noisy or incomplete data and the relative ease with which they can be retrained for classification of new types of data. These properties are likely to be useful for chromosome classification. Specimen preparation techniques in routine use evolve very rapidly, resulting in changes in chromosome appearance. In particular, there is an increasing clinical requirement to use higher-resolution banding for diagnostic purposes, resulting in routine examination of longer (prometaphase) chromosomes. This will result in the need for greater adaptability in automated karyotyping systems.

In this paper we present a preliminary investigation into the feasibility of using neural networks as chromosome classifiers, using the density profile as the representation of the banding pattern. We compare two commonly used network configurations and assess their suitability for this problem. The data used for this study are a set of 2904 profiles from G-banded chromosomes from Rigshospitalet, Copenhagen (Gerdes and Lundsteen 1981), each annotated with the chromosome's length, centromere position and classification

assigned by an experienced cytogeneticist. The profiles were carefully measured by densitometry from photographic negatives, care being taken to avoid overlapped or badly bent chromosomes. For this reason many of the cells in this data set are incomplete (there are, on average, 38.2 chromosomes per cell) which does not present a problem for this study as we are interested solely in classification of isolated chromosomes. However, the sex chromosomes, particularly the Y chromosomes, are represented by a very small number of examples. For this reason we have omitted the Y chromosomes from the study, only using data from the remaining 23 classes.

2. Neural nets

There is a variety of different network architectures, each with its appropriate training algorithm (see, e.g. Wasserman 1989, or the review in *Physics in Medicine and Biology* by Clark (1991)). Two types of network have been predominantly used for classification problems: the Kohonen self-organizing feature map or Kohonen net (Kohonen 1990) and the multi-layer perceptron (MLP). These nets accept continuous-valued rather than binary inputs and have been shown to generate low classification error rates in a range of other application domains (Kohonen 1988, Bisset *et al* 1989).

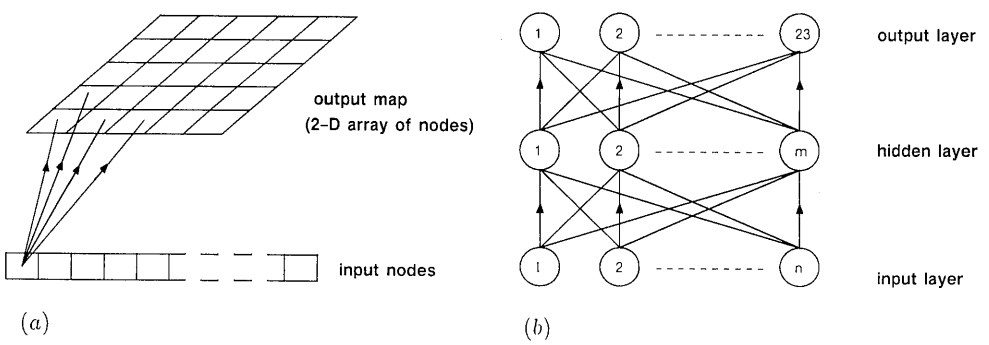


Figure 2. Neural network architectures. (*a*) The Kohonen self-organizing feature map. Each input is connected by a weighted link to each of the nodes in a two-dimensional array of outputs. (*b*) The MLP. The nodes are arranged in layers. Each node in a given layer is connected by weighted links to each node in the succeeding layer. The example shows a network with n input nodes, 23 output nodes and a single hidden layer, as used in this study.

Figure 2(*a*) shows the topology of a Kohonen net. All input nodes have links to all nodes in the two-dimensional array of outputs. Associated with each link is a weight, initially assigned a random value. The value of an output node is the sum of the weighted inputs on each link and these output values are manipulated by a competitive learning algorithm to arrive at an unsupervised clustering of output nodes (Kohonen 1990). As each training pattern is presented at the input, the output node whose pattern of weights most closely matches the input pattern is selected as the winner in the competition; its weights, and those of its neighbours, are updated to be closer to the input pattern, making it more likely to be selected and updated when a similar pattern is presented. In this way after a number of presentations of a representative set of training patterns, the output nodes form a map of regions each of which respond to different patterns in the training set. These regions can be

labelled according to the true class of the patterns to which they respond. The learning is unsupervised since the correct classifications of the input pattern are not used in adjusting the net parameters, merely in labelling the clusters after training.

The MLP topology is illustrated in figure 2(*b*). Nodes are connected in layers, all the nodes in each layer being connected via weighted links to all nodes in the succeeding layer. Each node performs a non-linear transformation of the sum of its inputs, and the output of this transformation is sent along all of the node's links to the next layer. As it passes along each link, this value is multiplied by the weight on that link before becoming one of the inputs to the next node. Each node may have an additional input, in the form of a bias or threshold which ensures that the summed inputs lie within an appropriate range for the transforming function. The input pattern is fed into the input layer, passing through several transformations before generating a pattern of outputs. There may be either one or two hidden layers of nodes between the inputs and the outputs, which act as feature detectors. In the experiments which we report below, only one hidden layer is used. For a given input, the output pattern is determined by the pattern of weights on the links. This pattern of weights is obtained by supervised training; each input pattern on training generates an output which is compared with the correct output pattern for its class, and the weights are adjusted so that the observed output is closer to the desired output. The bias values are also adjusted as part of the same process. The most commonly used algorithm for adjusting the weights is error back propagation (Rumelhart *et al* 1986). The error signal at the output is used to propagate adjustments to the weights back through the network so that the resulting pattern of weights produces outputs closer to those appropriate for the input pattern.

3. The classification experiments

The classification experiments involved training each network with approximately half of the chromosome data and using the remainder as 'unseen' data to test the network's classification performance, reversing the roles of the training and unseen set and averaging the classification rates. Several network parameters were varied for each network to investigate the effect on classification performance. One feature common to the two networks was the form of the input. We wished to use the banding profile to represent the banding pattern, i.e. to assign one profile sample to each input node.

Cells are measured at different stages of contraction during metaphase, so that the sum of the lengths of all the chromosomes in a cell varies considerably. Furthermore, chromosomes do not contract uniformly along their lengths so that not only do longer chromosomes show more bands than shorter ones but this increase in banding resolution occurs unevenly along the chromosome. There is therefore considerable inter-cell variation in the appearance of chromosomes of the same class. One aspect of this variation is in the number of samples making up a profile, as these are taken at equal spacings along the centreline, typically at pixel positions in the metaphase image. This variation can be reduced by normalization of the sum of the chromosome lengths. In our experiments we use the median cell length as a normalization measure to compensate for the fact that many of the cells had missing chromosomes.

To ensure that each input to the network corresponds to a consistent feature in the profiles, all chromosomes were presented using the same number of profile values. The number of profile value inputs needed to achieve best classification performance was one of the parameters investigated for each network. The longest profiles in our data set consisted of 116 elements. Shorter input vectors were generated by adding consecutive elements

together without averaging, a procedure which retains the total chromosome density in the profile, to produce profiles 58, 29 and 15 elements long. Additional inputs representing chromosome length and centromeric index were used in the case of the MLP.

In each experiment the measure of performance was the percentage of all chromosomes misclassified in both the training set and the unseen test set. A misclassification in this case was taken to be the assignment of a chromosome to a class other than that to which it had been assigned by the cytogeneticist when the data set was collected. The important misclassification rate is, of course, that for the unseen portion of the data, which provides information on the network's ability to generalize; it is this value we seek to minimize by selecting appropriate network parameters.

3.1. Kohonen self-organizing map

3.1.1. Methods. The Kohonen map was investigated using only the banding data as input. Three experiments were conducted.

(i) The number of input nodes. In this experiment the size of the map was fixed at 81 nodes (a 9×9 array). Each network, with 116, 58 and 29 input nodes, respectively, was trained using 30 examples of each of the 23 classes and tested on 50 unseen examples noting the classification performance. The number of output nodes was chosen to be similar to that which had been used successfully in a problem of similar magnitude, in which 96 output nodes were used to recognize 18 phonemes (Kohonen 1988).

(ii) The number of training examples. The selection of 30 examples for training in experiment (i) was rather arbitrary, and the effect on the selected network of altering the number of training examples was therefore investigated. Classification performance was noted using training sets of 10, 20, 30 and 40 chromosomes.

(iii) The size of the output map. Finally the effect on classification performance of the size of the output map was considered. The size of the output map was varied from 7×7 to 20×20 nodes using eight passes of 40 examples of each chromosome for training.

3.1.2. Results. (i) The number of input nodes. Table 1 shows that the classification error rate on both training and unseen data is relatively insensitive to the number of inputs used (i.e. to the resoluton at which the banding profile is presented). As both training time and labelling time are highly dependent on the number of inputs (and hence the number of weighted links), there is a clear advantage in the use of shorter input vectors. Experiments (ii) and (iii) were conducted using 29 input nodes.

Table 1. Variation in the average classification error of a Kohonen net as the number of input values reprsenting the banding pattern is varied. Output map size: 9×9; training set: 30 examples of each chromosome class; test set: 50 examples of each class.

Number of input values	Average classification error (%)	
	Training set (690 chromosomes)	Unseen set (1150 chromosomes)
116	24.0	37.4
58	22.8	37.0
29	22.4	37.1

(ii) The number of training examples. Figure 3 demonstrates the effect of the number of training examples on the net's ability to classify and to generalize to unseen data. Training

passes using different numbers of example chromosomes show that when 40 examples are used (figure 3(*d*)) not only is best classification performance achieved, but the performance on unseen data matches that on training data, indicating that the net is 'generalizing', or responding to general features of the training data, rather than specifically modelling the training examples.

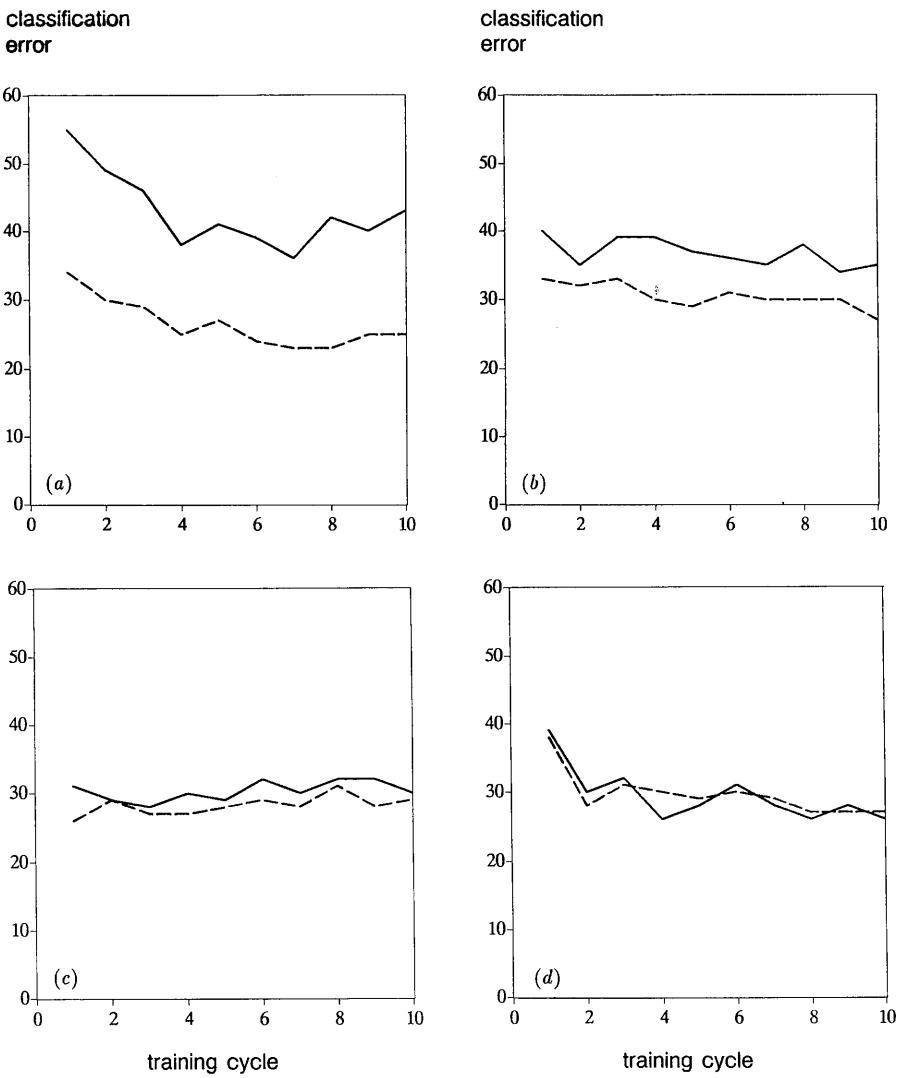


Figure 3. Training the Kohonen net with different numbers of examples of each class. (*a*) 10, (*b*) 20, (*c*) 30, (*d*) 40 examples in the training set. The curves show the percentage of chromosomes incorrectly classified in the training set (broken) and the unseen set (full) as the number of passes of the training data is increased. There is a clear improvement in both absolute classification performance and ability to generalize.

(iii) The size of the output map. The effect on classification of the size of the output map is shown in table 2. A consistent training regime (eight passes of 40 examples, following the result of experiment (ii)) was used for each map, and under this regime, best

performance was obtained using an 18 × 18 array of output nodes. As with the smaller maps, only a few passes of training data were required for the network to settle down to a fairly stable configuration (figure 3), after which the misclassification rates oscillate as the number of training passes is increased. The minimum misclassification rate observed with this configuration occurred after six passes of the training data, giving values of 11.8% on training data and 16.7% on unseen data.

Table 2. Variation in the average classification error of a Kohonen net with the size of the output map. Input vector size: 29 values; training set: 40 examples of each class.

Size of output map	Average classification error (%)	
	Training set (920 chromosomes)	Unseen set (920 chromosomes)
7 × 7	35.4	39.5
9 × 9	26.7	25.9
11 × 11	22.6	23.6
13 × 13	17.2	23.3
15 × 15	15.4	20.3
17 × 17	14.2	20.1
18 × 18	12.7	18.9
19 × 19	14.1	22.3
20 × 20	8.9	20.5

3.2. Multi-layer perceptron

3.2.1. Methods. The MLP net was trained using the back-propagation algorithm of Rumelhart *et al* (1986), the weights and bias values being updated after every presentation of a training profile. The back-propagation algorithm provides two parameters, gain (or learning rate) and momentum, controlling the rate at which weights change in response to error signals. We wished to investigate the effect on classification of varying the gain and momentum values, as well as the size of the input vector and the number of hidden nodes. For simplicity we restricted our study to the use of a single layer of hidden nodes. As we were not aware of any similar studies in the literature which might give us reasonable starting values for our experiments (as had been the case with the Kohonen net) we conducted some pilot experiments, similar to those reported below using chromosomes from groups 1–5 only (the longest chromosomes). The results of these experiments indicated that the number of inputs representing the banding pattern could be reduced to 15 without loss of classification ability, and that 15 hidden nodes were adequate. Experimental determination of appropriate gain and momentum values gave figures of 0.3 and 0.7 respectively (values which have been used successfully in other net simulations; see, e.g. Lippmann 1987).

All the nets tested had 23 output nodes, one for each class (Y chromosomes being omitted). The desired output is therefore close to unity at the output node corresponding to the correct class, and zero elsewhere. The classification produced by the network on being presented with a test pattern was taken to be the node with the highest output, even if this was significantly lower than unity.

(i) Gain and momentum. The gain term η determines the rate at which weights are altered in response to an observed error signal. A large value of gain (near 1.0) produces fast training and is useful in moving weights rapidly away from their initial random values. However, once the weight configuration is close to a minimum in error space, relatively large weight shifts can have the effect of 'jumping' out of a global minimum and into a nearby

local minimum. Small values of gain are more appropriate later in training and a useful approach is to successively reduce the learning rate as training proceeds. Two measures of network performance may be used to determine the points at which the gain term is reduced. These are the total net error (the sum of the absolute values of the differences between the desired and observed net outputs over a training pass) and the classification performance on training data. Both of these measures reflect the net's classification ability, but they are not identical, due to the fact that only the output node with the highest value affects classification. At different stages of the learning process one may be more useful than the other. This difference between these two measures has been noted in other applications (Dahl 1987). The weight change at a node at a given training cycle, determined by the observed error signal and the gain, may be further augmented by a fraction of the weight change which occurred at the previous cycle. In this way successive weight changes in the same direction are encouraged, whereas weight changes in successively different directions are damped, reducing oscillatory behaviour. The momentum term α specifies the fraction of the previous weight change which is added at each cycle. We investigated the best choices for momentum and initial gain, as well as the effect of reducing gain using the two criteria described above.

(ii) The number of input nodes. The result of our pilot experiment, that the profiles could be reduced to 15 samples (one eighth of the length of the longest profile), was tested by measuring classification performance on all 23 classes using a network with 15 hidden nodes, using gain and momentum values determined in experiment (i).

(iii) The number of hidden nodes. The classification performance was determined using different numbers of nodes in the hidden layer.

(iv) The use of additional chromosome features. Experiments (i)–(iii) were designed to examine the feasibility of configuring an MLP to use the density profile as a banding representation for classification. As noted in the introduction, chromosome size and centromere position are also important classification features. To test the overall classification performance of our network, two additional input nodes were included for the normalized length and the centromeric index. (The centromere divides the chromosome into a long 'arm' and a short 'arm' (figure 1(*b*)). The centromeric index is the ratio of the length of the short arm to the whole chromosome length.) The effect on classification of adding these features was examined.

3.2.2. Results. (i) Gain and momentum. The effect of reducing gain during training is demonstrated in figure 4, which shows a succession of training passes for a 15–15–23 network. Each training cycle involved the presentation of 1150 chromosomes. Figure 4(*a*) compares the network convergence on training with and without a reduction in gain. In this case gain was halved whenever a 10% increase in net error occurred or if the total number of correctly classified chromosomes had not increased by at least two over the previous presentation of the training data. The gain reduction method clearly results in improved convergence. Figure 4(*b*) demonstrates the effectiveness of using both total net error and classification performance as criteria for reducing gain. The starting value of gain was 0.3 and momentum was 0.7, as determined by the pilot study. Figure 5(*a*) and (*b*) shows training passes using different values of momentum and initial gain confirming that 0.7 and 0.3 are the most appropriate for this configuration.

(ii) The number of input nodes. Table 3 confirms that the classification performance is fairly insensitive to the number of values used to represent the banding profile (and hence, the numbers of inputs). No loss in classification accuracy is observed when an input vector of 15 values is used.

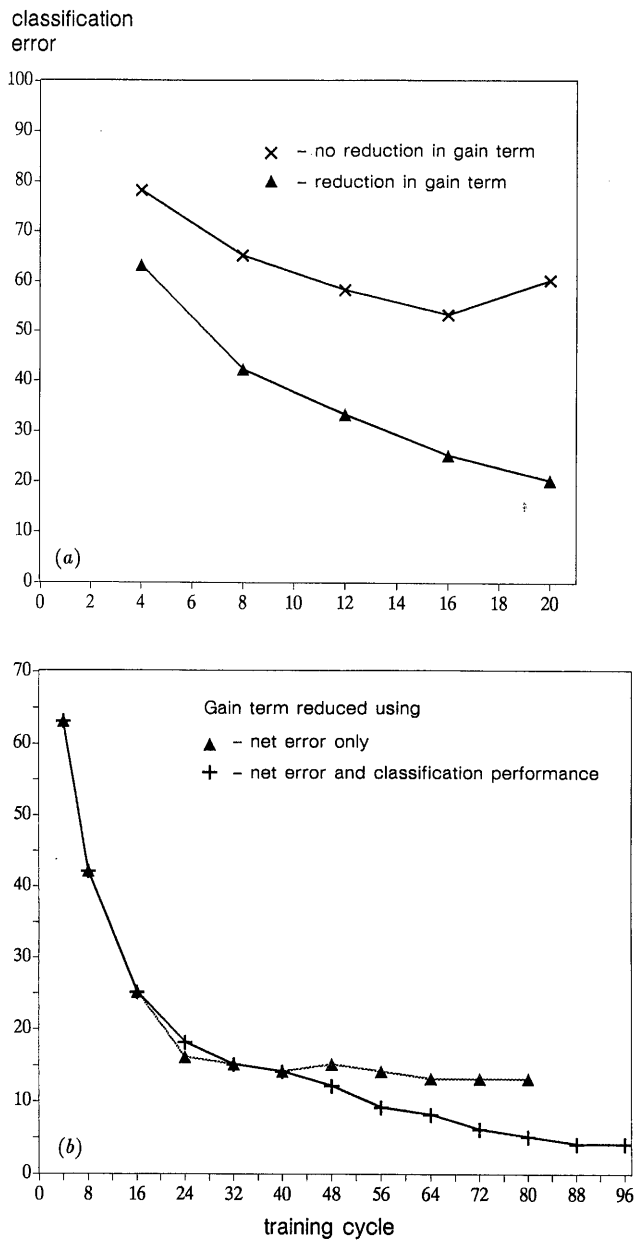


Figure 4. The effect of reducing gain during training of the MLP. A 15–15–23 net was used with momentum $\alpha = 0.7$ and initial value of gain $\eta = 0.3$. The graphs show the percentage of the training set incorrectly classified as the data are successively presented to the network. (*a*) Comparison of no gain reduction with gain reduction based on total network error. (*b*) Comparison of gain reduction criteria. Total network error alone compared with network error and classification performance combined.

(iii) The number of hidden nodes. The choice of 15 nodes in the single hidden layer is confirmed by figure 5(*c*) which shows training passes of 15–10–23, 15–15–23 and 15–20–23 networks.

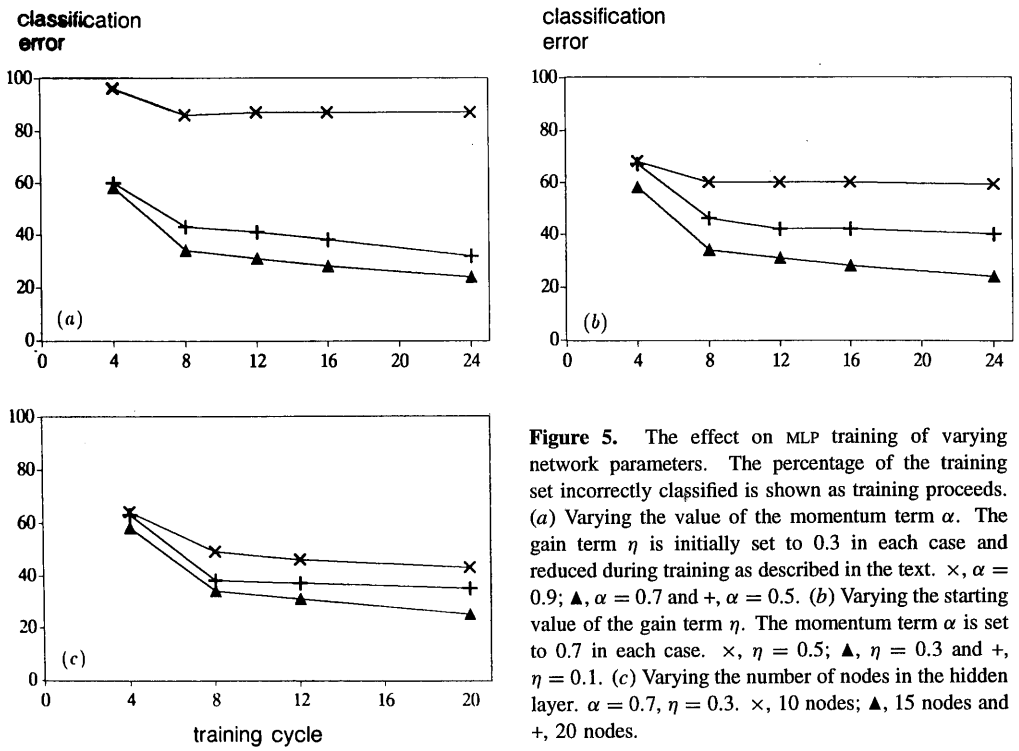


Figure 5. The effect on MLP training of varying network parameters. The percentage of the training set incorrectly classified is shown as training proceeds. (*a*) Varying the value of the momentum term α. The gain term η is initially set to 0.3 in each case and reduced during training as described in the text. ×, $\alpha = 0.9$; ▲, $\alpha = 0.7$ and +, $\alpha = 0.5$. (*b*) Varying the starting value of the gain term η. The momentum term α is set to 0.7 in each case. ×, $\eta = 0.5$; ▲, $\eta = 0.3$ and +, $\eta = 0.1$. (*c*) Varying the number of nodes in the hidden layer. $\alpha = 0.7$, $\eta = 0.3$. ×, 10 nodes; ▲, 15 nodes and +, 20 nodes.

Table 3. Variation in classification error rate of the MLP with the number of input values. The network had 15 hidden nodes and 23 output nodes.

Number of input values	Average classification error (%)	
	Training set (1150 chromosomes)	Unseen set (690 chromosomes)
15	2.2	10.1
58	3.9	10.0
116	3.0	10.3

(iv) The use of additional chromosome features. Table 4 shows the effect on classification rates of using additional inputs for normalized chromosome length and centromeric index using (16 or 17)–15–23 networks. These features clearly contribute significantly to classification performance. The overall best misclassification rates observed were 2.6% on training data and 6.6% on unseen data.

Table 4. The contributions of chromosome features to classification by the MLP.

Features used for classification	Average classification error (%)	
	Training set (990 chromosomes)	Unseen set (990 chromosomes)
Grey level profile	4.3	10.9
Grey level profile and normalized length	4.0	8.8
Grey level profile, normalized length and area centromeric index	2.6	6.6

4. Discussion

This preliminary study has investigated the feasibility of using neural network classifiers as part of an automated chromosome analysis system. The motivation for investigating neural networks in this application is their potential adaptability to changes in the data arising from evoluton in the sample preparation techniques and the desire to use a classifier making use of a more natural representation of the banding pattern than has been used in previous studies.

The MLP shows considerable promise as a classifier. By varying the operating parameters we have found a configuration which is capable of achieving useful classification rates. Our search of the parameter space was not broad and, even within its own narrow limits, not exhaustive. The final configuration of parameters may not therefore be the best obtainable, but it is gratifying to note that a 'good' combination was found fairly rapidly. It will be the subject of further study to investigate whether a more thorough search of parameter space will improve the network further. For comparison, the best published classification rates to be found in the literature (to our knowledge) are those reported by Piper and Granum (1989) who achieve 5.9% misclassifications on a superset of the profile data used here. The additional data used by these authors included a number of badly bent and some overlapping chromosomes (excluded from our data). Their results, however, include the effect of a further classification step which imposes the constraint that each chromosome class contains at most two chromosomes. While our results are not directly comparable with theirs, we are encouraged to achieve similar classification performance on similar data. Our experiments with the MLP demonstrate that considerable advantage is to be gained by careful adjustment of network parameters and training conditions.

Table 5. Comparison of the results obtained for chromosome classification using a Kohonen net with parameters and results for two other studies. The classification accuracy reported for phoneme recognition including post-processing using a context-sensitive grammar.

	Chromosome classification	Recognitition of the spoken words 'yes' and 'no'	Recognition of phonemes
Number of inputs	29	15	15
Output map size	18 × 18	7 × 7	8 × 12
Number of classes	23	2	18
Number of training examples	40	798	50
Error for unseen set (%)	16.7	6.9	3–8

It is unsurprising that the performance of the Kohonen map is inferior to the MLP in this application. For one thing, supervised training is much more appropriate to this data than unsupervised training. Table 5 shows a comparison of our results with two other applications of the Kohonen map on phoneme recognition (Kohonen 1988) and on recognition of the spoken words 'yes' and 'no' (Lucas and Kittler 1989). The results on phoneme recognition include a final supervised training phase, and refer to the classification accuracy of speech after processing the neural net output using a context-sensitive grammar (similar in its effect to the rearrrangment included in the results of Piper and Granum). It is tempting to speculate that better classification results could be obtained using the Kohonen map for chromosome classification, for example by increasing the size of the training set. This option was not available to us in this study, and the more promising performance of the MLP suggests that development of that route is more appropriate as a practical approach to this problem.

Acknowledgments

This work was supported by funding from the Science and Engineering Research Council. It was greatly facilitated by the exchange of materials within the Concerted Action of Automated Cytogenetics Groups, supported by the European Community (project II.1.1/13).

References

Bisset D L, Filho E and Fairhurst M C 1989 A comparative study of neural network structures for practical application in a pattern recognition environment *1st IEE Int. Conf. on Artificial Neural Networks (London, 1989)* (London: IEE) pp 378–82

Clark J W 1991 Neural network modelling *Phys. Med. Biol.* **36** 1259–317

Dahl D E 1987 Accelerated learning using the generalised delta rule *IEEE 1st Int. Conf. on Neural Networks* vol 2 (New York: IEEE) pp 523–30

Gerdes T and Lundsteen C 1981 *Documentation of the Rigshospital Chromosome Density Profile Data Base* Department of Obstetrics and Gynaecology YA and Department of Paediatrics, Rigshospitalet, University of Copenhagen

Graham J 1987 Automation of routine clinical chromosome analysis I. Karyotyping by machine *Anal. Quant. Cytol. Histol.* **9** 383–90

Graham J and Piper J 1993 Automatic karyotype analysis *Chromosome Analysis Protocols* ed J R Gosden (Clifton, NJ: Humana) at press

Graham J and Pycock D 1987 Automation of routine clinical chromosome analysis II, metaphase finding *Anal. Quant. Cytol. Histol.* **9** 391–7

Granlund G H 1976 Identification of human chromosomes using integrated density profiles *IEEE Trans. Biomed. Eng.* **BME-23** 183–92

Granum E 1982 Application of statistical and syntactical methods of analysis to classification of chromosome data *Pattern Recognition Theory and Application* ed J Kittler, K S Fu and L F Pau NATO ASI (Dordrecht: Reidel) pp 373–98

Groen F C A, tenKate T K, Smeulders A W M and Young I T 1989 Human chromosome classification based on local band descriptors *Pattern Recognition Lett.* **9** 211–22

Haykin S and Deng C 1991 Classification of radar clutter using neural networks *IEEE Trans. Neural Networks* **NN-2** 589–600

Kohonen T 1988 The neural phonetic typewriter *Computer* **21** 11–22

—— 1990 *Self Organisation and Associative Memory* 3rd edn (Berlin: Springer)

Lippmann R P 1987 An introduction to computing with neural nets *IEEE ASSP Mag.* **4** 4–22

Lucas A E and Kittler J 1989 A comparative study of the Kohonen and multiedit neural net learning algorithms *1st IEE Int. Conf. on Artificial Neural Networks (London, 1989)* (London: IEE) pp 7–11

Lundsteen C, Gerdes T, Granum E and Philip J 1981 Automatic chromosome analysis II. Karyotyping of banded human chromosomes using band transition sequences *Clin. Genet.* **19** 26–36

Lundsteen C, Lind A-M and Granum E 1976 Visual classification of banded human chromosomes I. Karyotyping compared with classification of isolated chromosomes *Am. J. Hum. Genet.* **40** 87–97

Lundsteen C and Martin A O 1989 On the selection of systems for automated cytogenetic analysis *Am. J. Med. Genet.* **32** 72–80

Piper J and Granum E 1989 On fully automatic measurement for banded chromosome classification *Cytometry* **10** 242–55

Rumelhart D E, Hinton G E and Williams R J 1986 Learning internal representations by error propagation *Parallel Distributed Processing: Explorations in the Microstructures of Cognition* vol 1, Foundations ed D E Rumelhart and J L McCelland (Cambridge, MA: MIT Press) pp 318–62

Sejnowski T J and Rosenburg C R 1987 Parallel networks that learn to pronounce English text *Complex Systems* **1** 145–68

Thomason M G and Granum E 1986 Dynamically programmed inference of Markov networks from finite sets of sample strings *IEEE Trans. PAMI* **PAMI-8** 491–501

Wasserman P 1989 *Neural Computing–Theory and Practice* (New York: Von Nostrand Reinhold)

14. **Application of artificial neural networks to chromosome classification.**
P.A. Errington and J. Graham, *Cytometry 14: 627-639, 1993.*
doi:10.1002/cyto.990140607

© 1993 Wiley-Liss, Inc.

Cytometry 14:627–639 (1993)

Application of Artificial Neural Networks to Chromosome Classification[1]

Phil A. Errington and Jim Graham
Department of Medical Biophysics, University of Manchester, Manchester M13 9PT, United Kingdom

Received for publication August 7, 1992; accepted January 18, 1993.

This work presents an approach to the automatic classification of metaphase chromosomes using a multilayer perceptron neural network. Representation of the banding patterns by intuitively defined features is avoided. The inputs to the network are the chromosome size and centromeric index and a coarsely quantized representation of the chromosome banding profile. We demonstrate that following a fairly mechanical training procedure, the classification performance of the network compares favourably with a well-developed parametric classifier. The sensitivity of the network performance to variation in network parameters is investigated, and we show that a gain in efficiency is obtainable by an appropriate decomposition of the network. We discuss the flexibility of the classifier developed, its potential for enhancement, and how it may be adapted to suit the needs of current trends in karyotyping.
© 1993 Wiley-Liss, Inc.

Key terms: Automated karyotyping, context free classification, Multi-Layer Perceptron

Inspection of chromosomes is an essential procedure in many fields of investigation for detecting genetic abnormality, damage due to environmental factors, or diagnosis of cancer. In particular, for clinical purposes, a karyotype is required in which chromosomes must be assigned to one of 24 classes (29). This task, although highly skilled, contains substantial elements of a tedious and repetitive nature, resulting in some interest in recent years in developing automated karyotyping systems (6,9,10,20,23,30,41). A number of such systems are available commercially and in use in clinical laboratories (reviewed in 24). They have been shown to contribute positively to laboratory efficiency (23). A central element in automated systems is the classification of chromosomes based on features that can be measured from the digitised image, such as that in Figure 1.

A number of approaches to automatic chromosome classification have been described (5,6,11,12,14,20,22, 23,27,28,31,32,37). In addition to using the important features of chromosome size and centromere position, all make use of some representation of the chromosome banding pattern, often in the form of a density profile projected onto the chromosome's centre line (see Figs. 7 and 8).

Chromosome size and overall density vary significantly between cells, but these differences can be compensated in a straightforward manner. The banding patterns also vary considerably in detail in chromosomes of the same class from different cells. It is to accommodate these differences that various classification methods have been applied such as template matching (27,28), Fourier analysis (5,6,20), Gaussian decomposition (11,12), the use of band transition sequences (22), weighted density distributions (14,23, 31,32), structural band descriptions (15), and Markov networks (37). All of these classification methods make use of an intuitive transformation of the density distribution into a set of features to be used by some sort of statistical discriminator. In this study, we present a new approach based on an artificial neural network. Neural network classifiers have been shown to be highly adaptable and capable of generalising about classes based on training data (3,21,35). They have been applied in classification tasks where classical pattern recognition methods have not been applied or have been unsuccessful (1,16,26,36). In applying a network to chromosome data, we not only have the opportunity to develop a novel and potentially superior classifier but to compare the performance of a neural network with that of a more conventional classifier in a well-studied domain.

[1]This work is partially funded by the U.K. Science and Engineering Council (SERC), Grant No. 90310105.

FIG. 1. Image of a metaphase cell showing G-banded chromosomes.

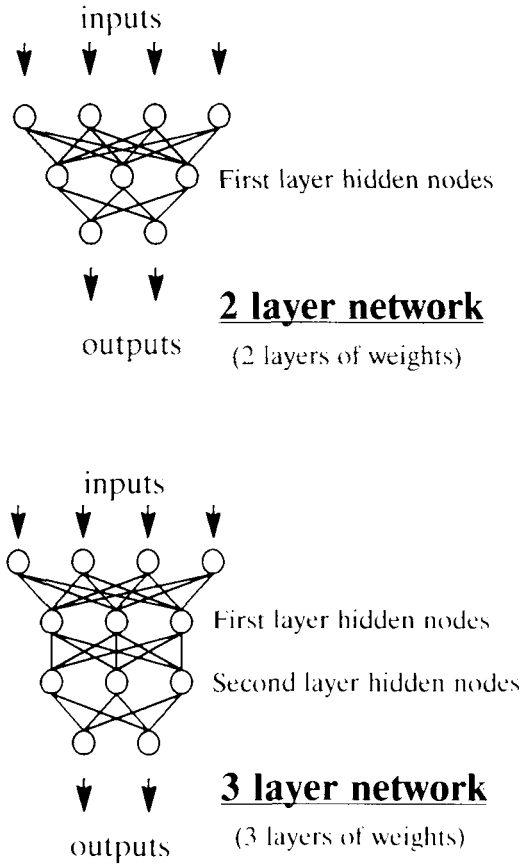


FIG. 2. Examples of 2- and 3-layer multi-layer perceptrons (MLPs). Weights are applied on links between each node.

MATERIALS AND METHODS

Neural Network Classifier

The Multi-Layer Perceptron (MLP) is a design of artificial neural network that consists of a number of nodes and interconnecting weights (Fig. 2). The nodes are arranged in layers, such that every node in one layer is connected to every node in a succeeding layer by a weighted link. No connections exist between nodes in the same layer. The MLP is capable of solving complex decision problems after sufficient training (34,36).

Each node in the network performs a weighted sum of all its inputs (Fig. 3). This sum is then passed through a nonlinear transfer function, commonly the sigmoid (equation 1), to produce an output in the range of 0 to 1 (Fig. 4). The value of this output is transmitted to all the nodes in the next layer. A special feature of a node is a trainable bias threshold, which allows different magnitudes of input sums to produce high or low responses at a node's output.

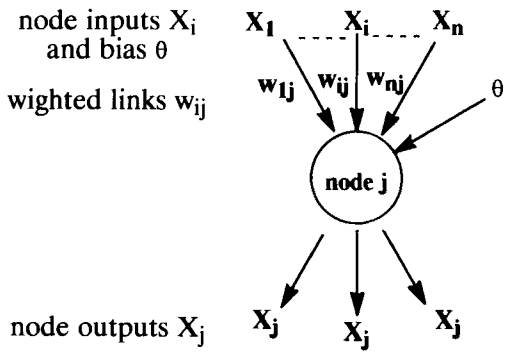
$$f(y) = \frac{1}{1 + e^{-y}} \quad (1)$$

The roles of the nodes in separate layers of the network are as follows. The input nodes serve to accept the data on which classification is to be based. Nodes in successive hidden layers discriminate between these inputs, acting as feature detectors, whereas output nodes map the features detected to output categories. Any number of hidden layers may be used, although two layers are sufficient for most classification tasks, as they can form arbitrarily complex decision regions in classification space (21).

For a given topology, the output of the network is a function of the input and the pattern of weights on the links between nodes. That is, the network's pattern of weights can act as a pattern recogniser, the relative strengths of each weight effectively constituting the pattern recognition algorithm. The pattern of weights appropriate to a particular classification task can be determined by training.

Network training algorithm. To train the network, we use the classical error back-propagation algorithm developed by Rummelhart et al. in 1986 (34). Other algorithms (reviewed in 8) are known to have improved properties, such as faster convergence to a stable set of network weights in some circumstances. We have not yet investigated whether their use would be appropriate in this case.

In error back-propagation, weights are initially set at small random values (so that summed outputs will initially lie on the steepest portion of the sigmoid curve; see Fig. 4). Node biases are also initialised to



$$X_j = f\left[\sum_{i=1}^{n} w_{ij}x_i - \theta\right]$$

where **f** is the sigmoid function
see equation 1

FIG. 3. Operation of a node in the network.

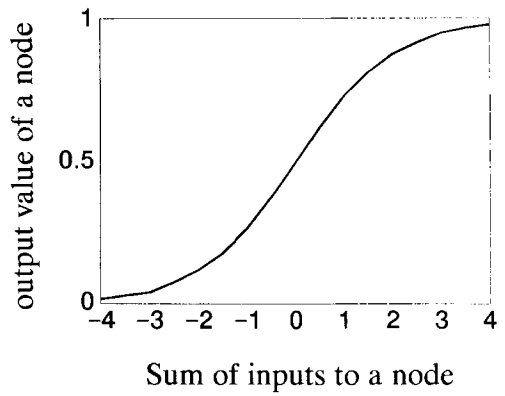


FIG. 4. Input to output transformation at a node in the network.

small values. Each time a training example is presented to the network, the values of all of the network's weights are altered in order that the values at output nodes move closer to desired target output values. As training proceeds, the weights are adjusted so that the response at an output signifying the category of training example is near 1, whereas outputs signifying other categories are near 0 (Fig. 5). After many applications of the training algorithm, the network weights should be such that if a training pattern is presented as input, a good approximation to the target response for the category of the pattern should be produced at the output nodes.

The key to the operation of the training algorithm is the back-propagation of error signals through the network, altering the values of weights to minimise each error signal. Although the error at an output node i is known, being the difference between an observed response o_{pi} and its target response t_{pi} for pattern p, the error for hidden nodes is not, being a function of the errors at each output node connected to the hidden node and the sigmoid transfer function. However, because the sigmoid function has a simple derivative, by application of the chain rule of differential calculus, the values of error at each hidden node can be calculated. (For specific details, refer to 34).

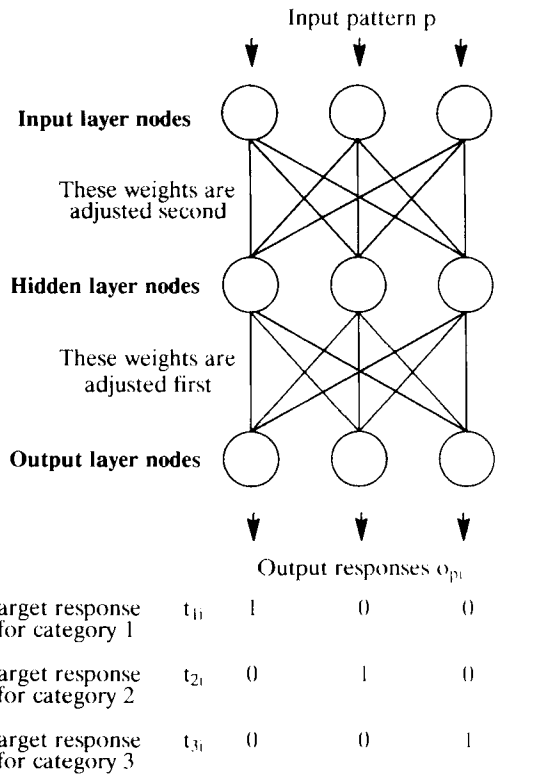


FIG. 5. Example of an MLP showing the ideal target responses and the order in which weights are adjusted during training.

The mechanics of the algorithm are as follows:

1. Initialise the network weights and thresholds to small random values.
2. Present each training pattern p in turn to the network.
3. Calculate the summed weights at each node passing the results to all nodes in the succeeding layer connected to that node, until values are produced at output nodes.
4. Calculate the error signal at each node in the output layer and use this to alter the values of the incoming weights to that node (using equations 2 and 4 below). Alter the node bias toward a value appropriate for the magnitude of the input sums.
5. Consider nodes in previous layers consecutively altering the values of incoming weights to these nodes using equations 3 and 4, adjusting the node biases appropriately.
6. Repeat steps 2–5 for all training patterns.

7. Repeat steps 2–6 until either all patterns produce the target responses at the output nodes or until the network weights do not appreciably change on each iteration.

The error signal δ_{pj} for pattern p at node j is calculated as follows:

For output units $$\delta_{pj} = (t_{pj} - o_{pj})\, o_{pj}\, (1 - o_{pj}) \quad (2)$$

For units in previous layers $$\delta_{pj} = o_{pj}\, (1 - o_{pj}) \sum_{k} \delta_{pk}\, w_{kj} \quad (3)$$

where

t_{pj} is the target response at node j for input pattern p.

o_{pj} is the observed response at node j for input pattern p.

w_{kj} are the weights to a succeeding layer k.

δ_{pk} are the error signals at nodes in a succeeding layer k.

The term $o_{pj}\, (1 - o_{pj})$ is the derivative of the sigmoid function with respect to its inputs.

Using the error signal δ_{pj}, the weights are altered using:

$$w_{ji}\, (t + 1) + w_{ji}\, (t) + \eta\, \delta_{pj}\, o_{pi} + \alpha\, (w_{ji}\, (t) - w_{ji}\, (t - 1)) \quad (4)$$

where $w_{ji}\, (t)$ is the weight value from node j to node i at iteration t

η is the gain term (see below)

α is the momentum term (see below)

o_{pi} is the value of the output of node i for pattern p

δ_{pj} is the value of the error signal for pattern p at node j

It has been shown that, using the above algorithm, the weights will eventually converge to a stable pattern as long as their initial values are nonidentical (34). However, because the method of weight adjustment is gradual in a gradient descent manner, the resulting stable pattern of weights may represent a local rather than a global error minimum. Convergence to a local minimum can often be avoided by the correct choice of the training parameters gain and momentum and the selection of a better suited network topology.

Gain and momentum. During training the weights to any node are adjusted in proportion to the error signal at that node and to the size of weight changes in previous iterations (see equation 4). The parameters controlling this adjustment are the gain and momentum.

The gain term in a network describes the amount by which the values of weights are changed by the error and output value of each node. Large values for gain result in large changes in weight values on each consideration of a training example. This is sometimes useful, as initially the random weight values on links may need to be altered considerably. However, so that weight values eventually stabilise to produce the lowest network error for training examples, smaller weight changes are more appropriate. The conflicting requirements may be resolved by reducing the magnitude of the gain as training proceeds. Other work (34) suggests that a fixed small gain can achieve the same final result as a reducing gain mechanism, but previous experiments with the data we use here showed that training times are improved by progressive reduction in gain (18).

The momentum term in a network is a mechanism for adding in previous weight changes to the current weight changes, producing the effect of smoothing the changes in weights made on each training pass. This mechanism reduces oscillations caused by presentation of two consecutive examples that seek to alter the weights in different directions. Similarly, successive weight changes in the same direction are amplified, allowing the network weights to reach a minimum error configuration faster. In general, inclusion of a momentum term aids the network in achieving a stable configuration of weights at a faster rate (34).

The gain in our experiments is initially fixed at a value between 0.1 and 0.9. Every 4 training passes (at step 7 in the above algorithm), a check is made to decide whether the current gain value should be reduced. If either the network error has significantly increased from its value on the previous pass or if the classification performance on training data is the same or worse, the gain value is halved. Training is stopped when the gain term is so low that no further weight adjustments occur (in our experiments when the gain drops below 0.0001). Figure 6 shows a typical error minimisation graph over a number of passes of the training data and the corresponding improvement in classification performance which accompanies this training.

Network topology. The performance of an MLP classifier is highly dependent upon its topology: the number of nodes it possesses and in how many layers they appear (Fig. 2 shows the two types of topologies generally chosen).

The number of output nodes is set by the number of possible output categories. For our application, 24 outputs are required, one for each chromosome class. The number of input nodes to the network is dependent upon how many inputs are to be fed to the network. In our experiments these inputs consist of a number of samples of the chromosome banding pattern together with features representing size and centromere position.

The use of two hidden layers of nodes, rather than one, can result in improved discriminatory ability (21) at the cost of increased training and classification time. For some applications, a single layer is sufficient (34,36).

Whereas theoretical guidelines can be derived concerning the number of nodes required for the first hidden layer in an MLP (3,17,34), these involve knowing something about the expected variability of the input data. As this is generally not known prior to experimentation, the number of nodes in each hidden layer and the number of layers required are usually deter-

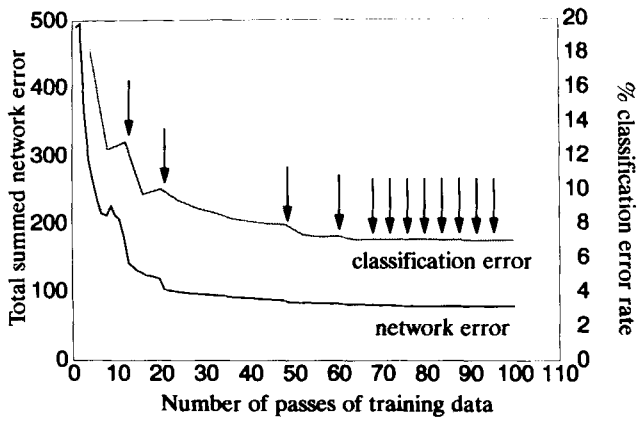


FIG. 6. Improvements in network error and classification rates as training proceeds. Arrows show positions where gain was halved.

mined empirically (regrettably not always very thoroughly).

In common with other authors, we have adopted a shorthand notation for network topologies. A single hidden layer (i.e., a 2-layer) network with 15 input nodes, 100 hidden nodes, and 24 output nodes will be written as a 15-100-24 network, for example.

Classification of Chromosome Data

For our experiments we have used three databases of annotated measurements from G-banded chromosomes that have been used in previous classification studies (13,14,18,27,31,37,41). Their details are summarised in Table 1.

In the case of the Copenhagen data set, chromosomes were carefully measured by densitometry of photographic negatives from selected cells of high quality. The other two data sets were taken from routine material. Each data set includes a number of severely bent and touching but not overlapped chromosomes. The nature of the slide preparation methods results in direct chorionic villus samples providing cells of significantly poorer visual quality than in the case of peripheral blood. The data sets therefore represent a range of data quality. (See Fig. 8 for some example density profiles from the Copenhagen, Edinburgh, and Philadelphia data sets.)

In each data set, the data for an individual chromosome consist of up to 140 grey level profile samples taken along the medial axis of the chromosome (see Fig. 7). This is supplemented with values for each chromosome's length and centromere position. The chromosome orientations were determined by the centromere finding algorithm (see (31)) and assigned before grey level profile samples were taken. As a result some profiles were sampled backwards, due in part to incorrect centromere location and in part to variation in the centromere position of meta-centric chromosomes. Manual correction of orientations was not applied.

Standardisation of data. The banding data in the raw data sets are not standardised for chromosome length or grey level variation between cells (caused by such factors as stain uptake and illumination), and the intensities of characteristic band sequences in a chromosome's profile vary considerably (Fig. 8). The number of banding samples also varies between cells, even for the same class of chromosome. Standardisation of these values was therefore required before they were presented to the network as inputs.

Density values were standardised by scaling the integrated density of the whole cell to a constant value. Chromosome lengths were similarly standardised to a constant cell length. The lengths of the individual profiles were further standardised by scaling each profile to a fixed length, stretching or compressing the profiles accordingly. In this way each input node in the network is presented with a profile value that consistently represents the same location in the banding pattern. Another approach would have been to adopt different standards for different classes of chromosome. This was, however, dismissed as unnecessarily difficult, as the classes of the chromosomes when presented to the classifier are unknown. By adopting a standard number of samples and re-introducing the length of the original profile as a feature, no information was lost.

Following results of a preliminary study (18), 15 profile values were used, extracted from the full profile by local averaging. The optimal number of profile element samples for each data set was determined empirically (see below). Figure 7 illustrates the extraction of a profile for a schematic chromosome; Figure 8 shows profiles both before and after the standardisation and sampling process.

Two other features used in our study were the centromeric index and standardised length of each chromosome. The centromere is a characteristic constriction that divides the chromosome into a long "arm" and a short "arm" (Fig. 7). The centromeric index is defined as the ratio of the length of the short arm to that of the whole chromosome.

Network Training and Testing

After extraction of the profiles and standardisation, the coarsely sampled profiles can be presented as inputs to the network. Approximately half the data in each set was used for training the network. The other half was later used as unseen test data. The roles of the training and test data partitions were exchanged in a subsequent experiment. Classification values for both experiments were then averaged to produce a mean classification error rate for training and test data for all of the data set.

The following experiments were conducted to identify the optimal configuration of network parameters the topology. Network training typically required more than 100 passes through the training data before further training had no effect (Fig. 6). There was no rearrangement of data between training passes. The initial network weight values for all three data sets were the

Table 1
Three Chromosome Data Sets Used

Data set	Tissue of origin	Digitisation method	Number of chromosomes	Data quality
Copenhagen	Peripheral blood	Densitometry from photographic negatives	8,106	Good
Edinburgh	Peripheral blood	TV camera	5,469	Fair
Philadelphia	Chorionic villus	CCD line scanner	5,817	Poor

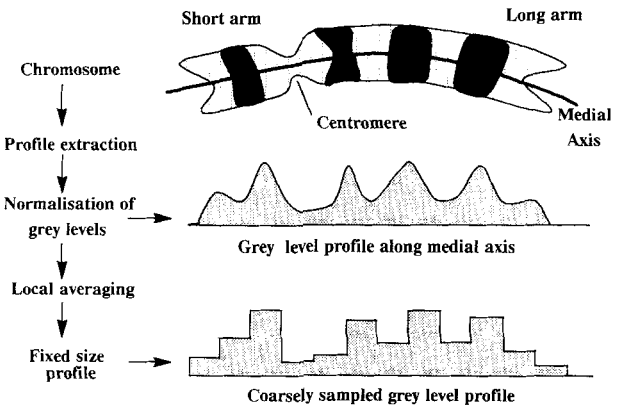


FIG. 7. Example of an extraction of a fixed length profile from a chromosome.

same, having been selected at random. The exception to this was experiment 5 in which the effect of varying the initial weight values was investigated. All of the classification results we present refer to classification of unseen test data.

Experiment 1: Choice of optimum gain and momentum values. The selection of initial gain and momentum terms is a matter infrequently addressed in descriptions of network applications. The effect on classification rates of these two parameters was tested by varying their values within their range of 0 to 1.0 and training a network with each value combination. After each network was fully trained, its classification performance on unseen data was evaluated. The experiment was performed using the Copenhagen data set.

Experiment 2: Selection of network toplogy. To test the performance of different toplogies, input profiles of 15 grey level samples were used. Twenty-four outputs were used correponding to the 24 possible chromosome classes in the data sets. Classification performance of the network was measured using different numbers of hidden nodes in both 2 and 3 layer networks.

Networks with 10, 24, 50, and 100 hidden nodes in the first hidden layer were initially tested for each data set, and the best performer of these networks for each data set was then tested with a further layer of nodes in a three layer network. The number of nodes in this third layer was varied over the same range. Notice that not all possible combinations of topology were explored, as that would have involved a prohibitively large number of trials. We have made the assumption, common in hill-climbing optimisation, that the "topology space" is sufficiently smooth that optimising sequentially along each dimension is adequate.

Experiment 3: Inclusion of centromeric index and length. Centromeric index (CI) and length values are known to be powerful classification features (7, 31,38). Three methods of incorporating these two features along with the banding profile were investigated (Fig. 9). Method 1 involved using the banding profile inputs with either a CI input or a length input; method 2 incorporated both CI and length as dual inputs, whereas method 3 included the CI and length information after it had been processed by a pre-classifier.

Centromeric index and length are features that have been used to classify chromosomes into seven "Denver" groups (7). Automatic classification using statistical classifiers have been shown to produce acceptable discrimination into 10 groups (38). The correspondence between the 10 "Denver" groups and the 24 pairs of the Paris convention (29) are shown in Table 2.

For our "Denver" classifier we have use a MLP with two inputs, 10 outputs and a single hidden layer of nodes. The optimum number of hidden nodes for this network was determined in a similar manner to that described in experiment 2. The 10 outputs correspond approximately to probabilities of belonging to each of the "Denver" groups (33). As our experiments were conducted first with one-half the data in each data set as training data and then, in a following experiment as unseen data (see above), two separate preclassifiers were required for each data set.

The three methods of inclusion of centromeric index and length features were evaluated using a (15 + X)-100-24 network, where X corresponds to 1 input (length or CI), 2 inputs (length and CI) or 10 inputs (from the "Denver" preclassifier).

Experiment 4: Effect of the number of profile samples. The decision to use 15 input samples arose from preliminary investigations (18). Since optimisation of the other network parameters may have altered the validity of this finding, we have evaluated the effect on network performance of varying the number of input nodes, and hence the profile sampling density. In this evaluation a two-layer network with 100 hidden nodes and 24 outputs was used. The number of input samples was varied between 1 and 100 (i.e., the network toplogies ranged from 1-100-24 to 100-100-24). These sampling rates were chosen to represent a range from the coarsest possible (one sample measures mean

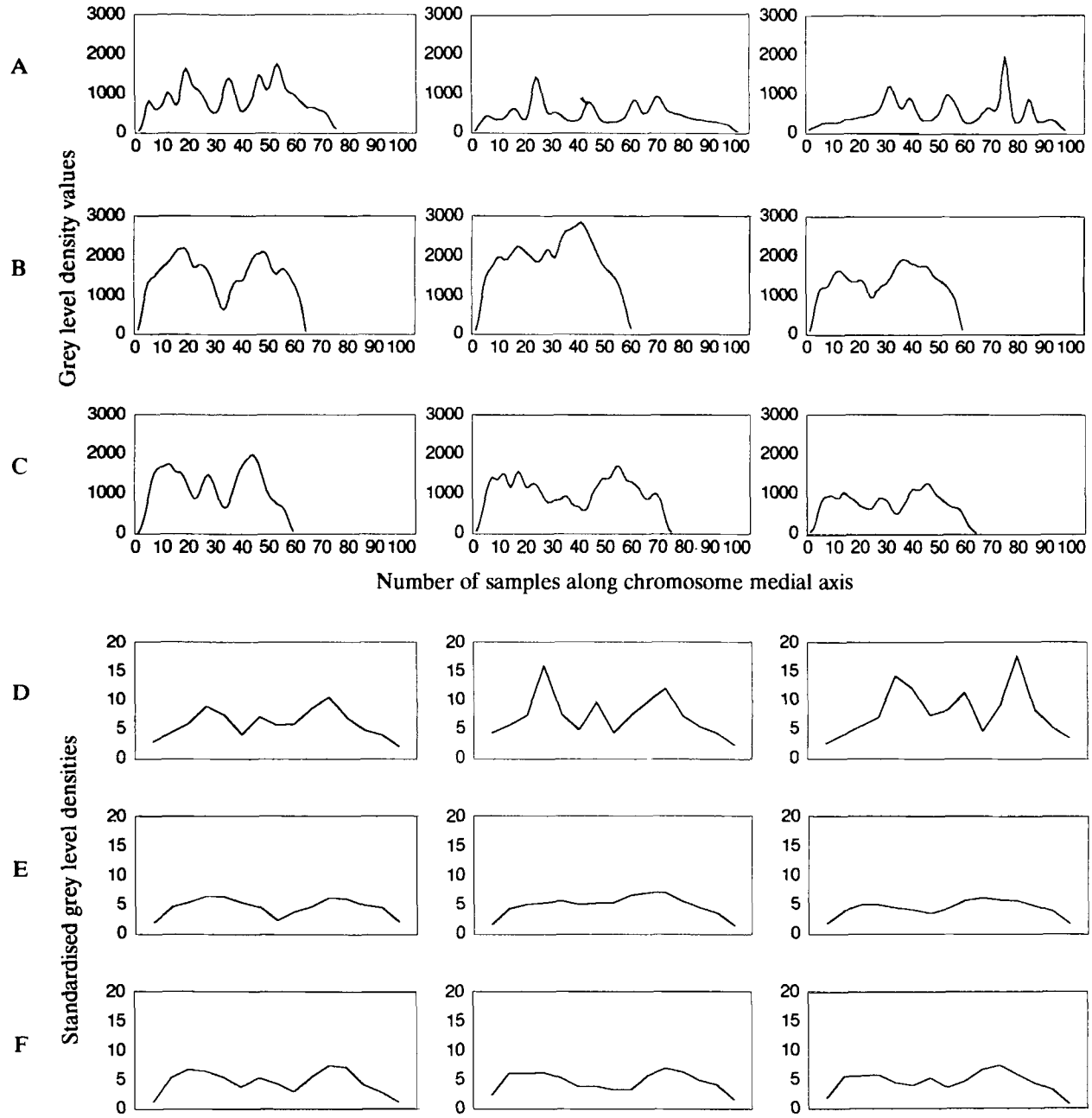
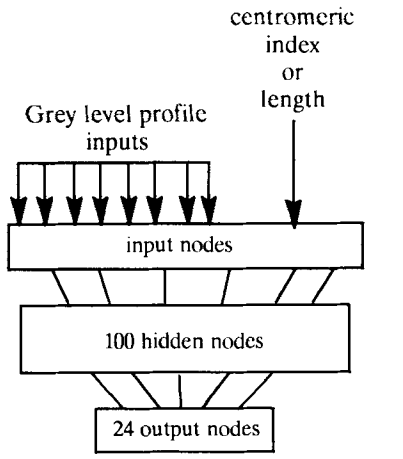


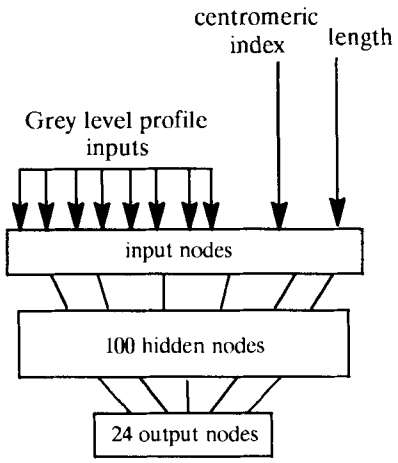
FIG. 8. Chromosome density profiles of chromosomes from group 1. **A,B,C** show examples of profiles from the Copenhagen, Edinburgh, and Philadelphia data sets, respectively, demonstrating the variability of the data that may be encountered. The rightmost example in A shows a profile that has been sampled in the wrong direction due to an incorrectly assigned centromeric index. **D,E,F** show the same profiles as 15 coarse samples after standardisation for length and cell grey level content has been applied.

grey level along the chromosome) to a number approximating to the average number of samples available for each chromosome.

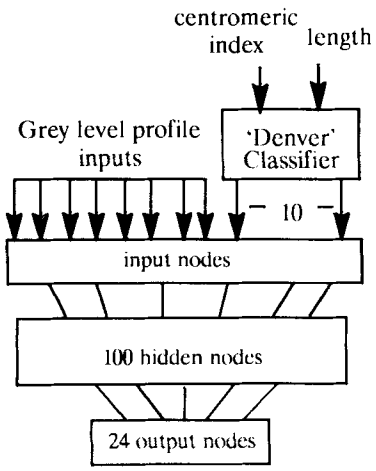
Experiment 5. Effect of the initial random weight settings in the network. One remaining factor that needed investigation was the effect of starting the network with different random weight values. This is necessary due to the gradient descent nature of back propagation, which may result in different final weight configurations and a corresponding classification performance variation. To investigate the possible variability in performance, five experiments were conducted for each data set, training a 15-100-24 network with different initial random weight settings. The per-



Method 1



Method 2



Method 3

FIG. 9. Different methods of inclusion of the centromeric index and length features.

formance of each resulting network, once trained, can be used to judge the precision of all the results we present.

RESULTS

Experiment 1: Choice of optimum gain and momentum values. Figure 10 shows the effect on classification performance of varying the gain and momentum parameters using the same network. This demonstrates that the variation in network performance can be severely affected by inappropriate choices of values for these parameters. High parameter values need to be avoided, particularly in combination. However, there appears to be a broad valley of possible gain and momentum combinations that produce similar low classification error rates. Classification performance of the network is therefore not critically dependent on which combination of the gain and momentum value is selected in this region. However, the rate of convergence of the network weights during training varies with the different parameter combinations. The values producing the most rapid convergence were 0.1 for the gain and 0.7 for the momentum, and these values were selected for use in all subsequent experiments.

Experiment 2: Selection of network topology. Figure 11 shows the variation in performance of various 2-layer network topologies. As can be seen, the best performance can be achieved by the network with 100 hidden nodes. This demonstrates that the network trains to a better representation of the classification problem with the greater number of nodes, as would be expected given that in our problem considerable within-class variability is observed in the training data. However, near optimum results can be produced with fewer nodes (e.g., 50), with corresponding advantages for speed of training and execution.

The experiment with the addition of a second hidden layer of nodes (to form a 3-layer network) shows only slight variation in performance rates once a 24-node limit is passed, but again more nodes result in a better performance (Fig. 12). The decreased performance with 10 hidden nodes in the second layer is not unexpected given that 24-output classes are to be separated.

The tradeoff in both of these cases is an increase in the classification and training times. Training time increases linearly with an increase in the number of nodes, as with more nodes there are more error terms to calculate and more weights to alter on each training pass. When classifying, more weighted sums need to be performed.

The best classification performances of the 2- and 3-layer networks are presented in Table 3.

Experiment 3: Inclusion of centromeric index and length. Adjustments of the topology of the preclassifier for the "Denver" classification was performed in a similar manner to the selection of the topology of the main classifier except that only 2-layer networks were considered. Best classification performance was

Table 2
Respective "Denver" Group for Each Chromosome Class

"Denver" group	A1	A2	A3	B	C	D	E1	E2	F	G
Chromosomes in group	1	2	3	4, 5	6, 7, 8, 9, 10, 11, 12, 23(X)	13, 14, 15	16	17, 18	19, 20	21, 22, 24(Y)

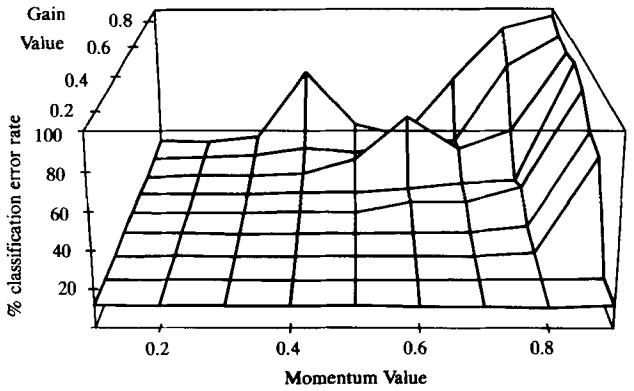


FIG. 10. Effect on classification error rates of variation of gain and momentum values.

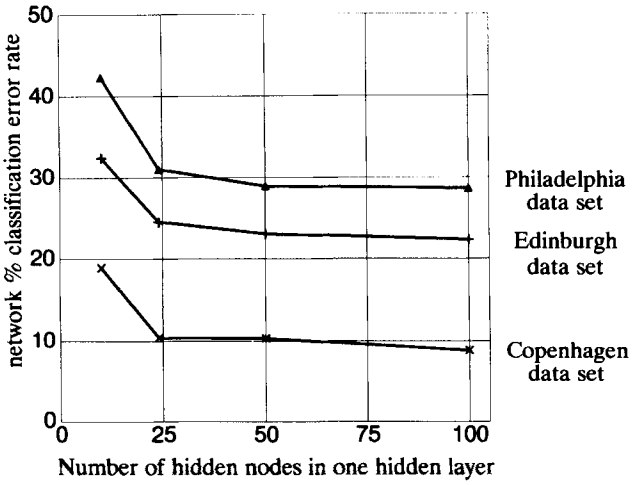


FIG. 11. Classification performance of networks with different numbers of hidden nodes in one hidden layer, tested with three data sets.

achieved with networks of 24 or 26 hidden nodes. The classification rates obtained are shown in Table 4, which also includes the performance of the network if the correct class lies within the top two or three highest ouputs. This is important since the values of probability that are high, but not maximum, may also affect the result of the main network, as all of the preclassifier's outputs, not just its highest output are presented as inputs to the main network.

The performance of the various network configurations using centromere and length features are presented in Table 5. These show that the separate inclusion of the centromere and length values both reduce error rates (method 1). Inclusion of the two features together (method 2) is more effective, but the approach using the preclassifier (method 3) is the most effective of all.

Experiment 4: Effect of the sampling rate of the banding profile on classification performance. The effect of the sample rate of the banding profiles for the three data sets is presented in Figure 13. As might be expected, very coarse sampling does not provide adequate information for classification. The best performances occur with 20–30 samples in each data set, although these are only slightly better than the performance obtained with 15 samples as used in other experiments. As greater numbers of inputs to a network result in longer classification times and significantly longer training times, we believed it was acceptable to use 15 inputs as a standard sampling rate for our purposes.

Experiment 5: Effect of the initial random weight settings in the network. Table 6 shows how the results of the 15-100-24 network, classifying the banding profiles only, change as the network is trained with different sets of initial random weights. The initial network conditions does not appear to affect performance significantly.

DISCUSSION

We have investigated the effects on classification performance of varying several of the parameters defining a multilayer perceptron network. Inappropriate choices of these parameters can result in classification performances considerably worse than optimal. However, the performance is not highly sensitive to any parameter. Figures 10–13 show that there are ranges of parameter values where overall classification performance rates are only slightly worse than the best obtained.

From experiment 1 we see that, within broad limits, gain and momentum values can be chosen to optimise training efficiency without compromising accuracy. Experiment 2 shows that discrimination is improved by adding more hidden nodes. However, for large numbers of nodes, the improvement is slow. Additional discriminating power is obtained by adding a second hidden layer.

The classification performance using the banding pattern alone is rather encouraging but, again unsurprisingly, improved by the addition of length and cen-

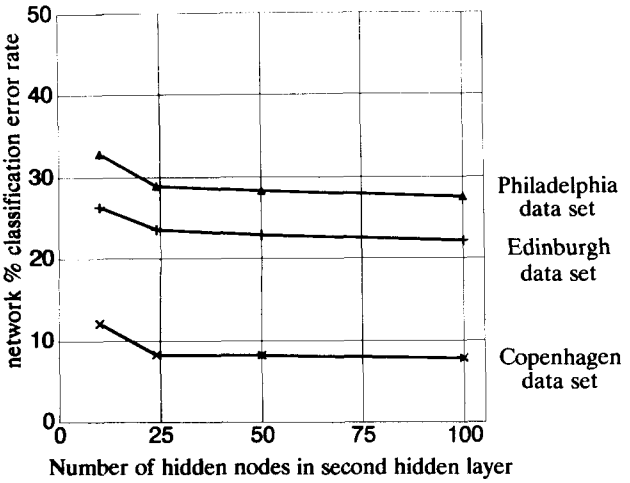


FIG. 12. Variation in classification performance of networks with different numbers of hidden nodes, in their second hidden layer. Experiment performed with 100 nodes in the first hidden layer and tested with three data sets.

Table 3
Classification Error Rates for Unseen Data for the Best Performing 2- and 3-Layer MLPs Using Banding Profiles as Sole Inputs (Experiment 2)

	Data set		
Network topology	Copenhagen	Edinburgh	Philadelphia
15-100-24	8.8%	22.3%	28.6%
15-100-100-24	7.8%	22.1%	27.5%

tromeric index features. Of these two, the centromeric index is the better additional discriminating feature. The improvement obtained by decomposing the net by factoring out the length and C.I. is interesting. These features are effectively being given a greater weight by this procedure. Presumably an "undecomposed" network with enough hidden nodes would be capable of achieving similar performance, but we achieved the performance gain by using knowledge of the problem to configure the network. It is tempting to assign an element of psychological plausibility to this configuration. Approximate classification on the basis of global features, refined using the local banding information, may reflect the process applied by human experts in classifying chromosomes.

Experiment 4 presents us with the result that the density profile representing the banding pattern need only be sampled fairly coarsely to obtain maximum discriminating power. This may be considered surprising given that the original profiles contained up to 140 samples and cytogeneticists examine fairly detailed features of the banding pattern. Two observations can be made here. First, few features may be necessary for classification of normal chromosomes, as distinct from identifying abnormalities. Even on the longer number 1 chromosomes of the high-quality Copenhagen data, there are only seven or eight discernible peaks (see Fig. 8). Second, the same resolution is used to identify long chromosomes (e.g., number 1) with several bands, as to identify the very short chromosomes (e.g., numbers 20, 21) with only one reliable band. We may be observing a "best compromise" between undersampling and long chromosomes and oversampling the short. Improved performance may be obtained by sampling the long chromosomes more finely than the short ones. This would necessitate having several networks, each corresponding to a different input resolution, with chromosomes assigned for classification on the basis of length, say. We have not investigated this extension, but it is a feasible route to improving results.

The overall classification performance varies with the three data sets as the difference in visual quality of the data might lead us to expect. It is of interest to compare the performance of our best network with that of existing classifiers based on conventional pattern recognition techniques. To our knowledge, the best classification performance over all classes that has been reported is that described by Piper and Granum (31). That classifier uses a carefully selected set of features based on weighted density distributions. A comparison with the results of (31) is likely to be particularly revealing since that study was conducted using the same data sets as we have used here. The final misclassification rates reported by those authors are obtained after application of the constraint which forces each class to contain at most two chromosomes. No such constraint dependent modifications have been applied in our case, although they could well be and would be expected to improve performance further. The context-free classification stage of Piper and Granum's work has been reported in (32,39), and we reproduce the results in Table 7 for comparison. Table 7 also shows the results obtained using another recently reported classifier (15), which uses a subset of the Copenhagen data set.

We are encouraged that we produced a small improvement in the mean classification performance for all three data sets. Application of a standard test for differences of proportions reveals that these improvements are not significant (significance levels of 22%, 25%, and 49% for the Copenhagen, Edinburgh, and Philadelphia data sets, respectively). Even this estimate of significance is likely to be optimistic given that identical samples were used in both cases. We can propose the following three advantages of the neural network classifier described here.

Robustness. The classification of several different data sets has been achieved using the same networks, notwithstanding the fact that the parameters were optimised on each set individually. Performance is not particularly sensitive to any parameter. It is worth noting that the network appears to be able to deal with polarity errors fairly naturally. This has not been investigated explicitly in this study, although it may be

Table 4
Classification Error Rates of "Denver" Classifiers Used for Preclassification

Data set	Best network topology	Error rate with only the highest output considered as correct	Error rate with the first and second highest outputs considered as correct	Error rate with the top three highest outputs considered as correct
Copenhagen	2-24-10	7.3%	2.9%	2.1%
Edinburgh	2-24-10	14.3%	4.4%	2.4%
Philadelphia	2-26-10	17.4%	7.7%	4.4%

Table 5
Classification Error Rates of Networks Using 15 Grey Level Banding Inputs With Different Representations of Centromere and Length Features

Features used	Data set		
	Copenhagen	Edinburgh	Philadelphia
Banding pattern alone	8.8%	22.3%	28.6%
Banding and normalised length	8.4%	19.4%	27.6%
Banding and centromeric index	7.7%	21.0%	26.5%
Banding, length and centromeric index	6.9%	18.6%	24.6%
Banding, and "Denver" groups	6.2%	17.8%	22.7%

noted that 37.4% of number 1 chromosomes in the Copenhagen data set were sampled with incorrect polarity. No attempt was made to take account of this, but the resulting classification error rate using banding information only for chromosome 1 was 4.5%.

Flexibility. The process of arriving at a good network configuration is a fairly mechanical one. In this case we have avoided the necessity for inspired selection of features. Neural networks have been shown to be highly adaptable classifiers in other fields (1,16, 26,36), and this may prove very useful as the material that cytogeneticists choose to study changes. The data sets used here were obtained from fairly short chromosomes with few bands. Today, longer chromosomes with greater numbers of bands are commonly used, and this trend is likely to continue. It can be expected that networks to deal with samples of this type will be trained fairly readily.

Parallel Hardware implementation. The networks presented here have all been implemented by software simulation. Although classification speed is not a problem, this may not be true of larger networks and we have noted the extra training burden that occurs as network size increases. The inherently parallel nature of the calculations makes implementation on parallel architectures fairly natural. In particular, there is a great deal of interest in hardware implementations (4). Such developments could lead to useful implementations in time critical applications such as classification in flow karyotyping.

The three data sets used in this study were chosen to enable a direct comparison with previously designed classifiers. Whereas the data sets are substantial, it may be that the number of chromosomes is insufficient for training either a network or a statistical classifier. A data set of 127,925 chromosome measurements is available (23), which could be used to investigate the effect of training set size on the network classification performance. This is a substantial study in itself and will be reported elsewhere.

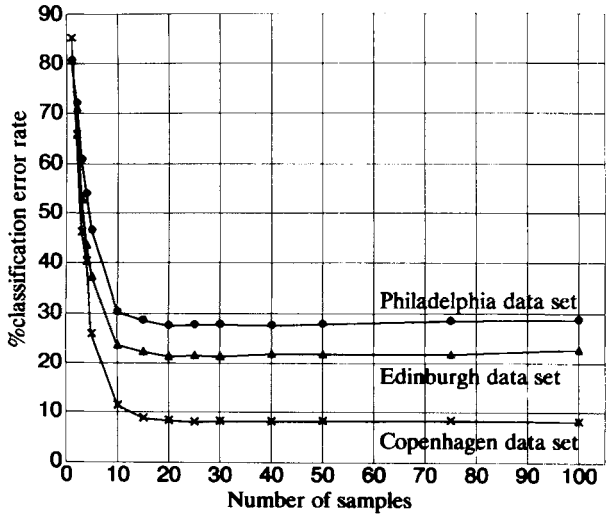


FIG. 13. Variation in network classification error rate with changes in sampling rate of density profiles. The network used contained 100 hidden nodes and 24 outputs.

ACKNOWLEDGMENTS

This work was greatly facilitated by the exchange of materials and ideas available within the Concerted Action of Automated Cytogenetics Groups supported by the European Community, Project No. II.1.1/13. We are grateful to Jim Piper of the MRC Human Genetics Unit, Edinburgh, for permission to reproduce some of his results.

Table 6
Performance Variation for the Same Network Trained on the Same Data With Five Different Sets of Initial Starting Weight Values

	Data set		
	Copenhagen	Edinburgh	Philadelphia
Average classification error rate	8.72%	22.16%	28.36%
Standard deviation in error rate	0.07%	0.17%	0.16%

Table 7
Comparison of Classification Error Rates for a Neural Network Classifier Using Both Banding and "Denver" Group Inputs with Two Recently Reported Statistical Classifiers (15,32)

	Classification error rate		
Data set	Network classifier	Parametric classifier using W.D.D. functions (32)	Parametric classifier using local band descriptors (15)
Copenhagen	6.2%	6.5%	11.5%
Edinburgh	17.8%	18.3%	N/A
Philadelphia	22.7%	22.8%	N/A

LITERATURE CITED

1. Anderson JA, Gately MT, Penz PA, Collins DR: Radar signal categorization using a neural network, Proceedings IEEE 78: 1646–1657, 1990.
2. Becker S, leCun Y: Improving the convergence of back propagation learning with second order methods. In: Proceedings Connectionist Models Summer School, 1988.
3. Baum EB: On the capabilities of multilayer perceptrons. Journal Complexity 4:193–215, 1988.
4. Boser BE, Sackinger E, Bromley J, leCun Y, Jackel LD: Hardware requirements for neural network pattern classifiers, IEEE Micro, 32–40, 1992.
5. Castleman KR, Wall RJ: Automated systems for chromosome identification. In: Nobel Symposium 23-Chromosome Identification, T. Caspersson (ed). Academic Press, New York, 1973.
6. Castleman KR, Melnyk J: An automated system for chromosome analysis—final report. Internal document No. 5040-30. Jet Propulsion Laboratory, Pasedena, TX, 1976.
7. Denver Conference: A proposed standard system of nomenclature of human mitotic chromosomes. Lancet 1:1063–1065, 1960.
8. Fahlman SE: Faster learning variations on back-propagation: An empirical study. In: Proceedings Connectionist Models Summer School, 1988, pp 38–51.
9. Graham J: Automation of routine clinical chromosome analysis I, Karyotyping by machine. Anal Quant Cytol Histol 9:383–390, 1987.
10. Graham J, Pycock D: Automation of routine clinical chromosome analysis II, Metaphase Finding. Anal Quant Cytol Histol 9:391–397, 1987.
11. Granlund GH: Identification of human chromosomes using integrated density profiles. IEEE Trans Biomed Eng BME 23:183–192, 1976.
12. Granlund GH: The use of distribution functions to describe integrated density profiles of human chromosomes. J Theor Biol 40: 573–589, 1973.
13. Granum E: Pattern recognition aspects of chromosome analysis: Computerized and visual interpretation of banded human chromosomes. PhD thesis, Technical University of Denmark, 1980.
14. Granum E: Application of statistical and syntactical methods of analysis to classification of chromosome data. In: Pattern Recognition Theory and Application, Kittler J, Fu KS, Pau LF (eds). NATO ASI (Oxford), Reidel, Dordreht, 1982, pp 373–398.
15. Groen FCA, Ten Kate TK, Smeulders AWM, Young IT: Human chromosome classification based on local band descriptors. Pattern Recognition Letters 9:211–222, 1989.
16. Haykin S, Deng C: Classification of radar clutter using neural networks. IEEE Transactions on Neural Networks 2:589–600, 1991.
17. Huang SC, Huang YF: Bounds on number of hidden neurons in multilayer perceptrons. IEEE Transactions on Neural Networks 2:47–55, 1991.
18. Jennings AM: Chromosome classification using neural nets. MSc thesis, University of Manchester, UK, 1990.
19. leCun Y, Denker J, Solla SA, Howard RE, Jackel ID: Optimal brain damage. In: Advances in Neural Information Processing systems, Vol. II. D. S. Touretzky (ed). Morgan Kaufman, San Mateo, CA, 1990.
20. Ledley RS, Ing PS, Lubs HA: Human chromosome classification using discriminant analysis and Bayesian probability. Comput Biol Med 10:209–218, 1980.
21. Lippmann RP: An introducton to computing with neural nets. IEEE ASSP 4:4–22, 1987.
22. Lundsteen C, Gerdes T, Granum E, Philip J: Automatic chromosome analysis II. Karyotyping of banded human chromosomes using band transition sequences. Clin Genet 19:26–36, 1981.
23. Lundsteen C, Gerdes T, Maahr J: Automatic classification of chromosomes as part of a routine system for clinical analysis. Cytometry 7:1–7, 1986.
24. Lundsteen C, Martin AO: On the Selection of Systems for Automated Cytogenetic Analysis. Am J Med Genet 32:72–80, 1989.
25. Lundsteen C, Phillip J, Granum E: Quantative analysis of 6985 digitised trypsin G-banded human metaphase chromosomes. Clin Genet 18:335–370, 1980.
26. McCulloch N, Ainsworth WA, Linggard R: Multi-layer perceptrons applied to speech technology. British Telecom Technol. Journal 6:131–139, 1988.
27. Neurath PW, Nablouzian B, Warms T, Serbagl R, Falek A: Human chromosome analysis by computer—an optical pattern recognition problem. Ann NY Acad Sci 128:1013–1028, 1966.
28. Neurath PW, Gallus G, Horton JB, Selles W: Automatic karyotyping: Progress, perspectives and economics. In: Automation of Cytogenetics, Proceedings of the Asilomar Workshop, Pacific Grove, CA, November 30–December 2, 1975, National Technical Information Services, Springfield, VA. 1975, pp 17–26.
29. Paris Conference (1971): Standardization in Human Cytogenetics. Original Article series, 8:7. National Foundation, New York, 1972.
30. Piper J, Nickolls P, McLaren W, Rutovitz D, Chisholm A, Johnstone I: The effect of digital image filtering on the perfor-

mance of an automatic chromosome classifier. Signal Processing 4:361–373, 1982.
31. Piper J, Granum E: On fully automatic feature measurement for banded chromosome classification. Cytometry 10:242–255, 1989.
32. Piper J: Aspects of chromosome class size classification constraint. CAACG Interlab meeting and Topical Workshop on High-level Classification and Karyotyping, Approaches and Tests, University of Aalborg, 13–14 March 1991.
33. Ruck DW, Roggers SK, Kabrisky M, Oxley ME, Suter BW: The multilayer perceptron as an approximation to the Bayes Optimal Discriminant Function. IEEE Transactions on Neural Networks 1:296–297, 1990.
34. Rumelhart DE, Hinton GE, Williams RJ: Learning internal representations by error propagation. In: Parallel Distributed Processing: Explorations in the Microstructures of Cognition. Rumelhart DE and McCelland JL (eds), Vol. 1. Foundations. MIT Press, Cambridge, 1986, pp 318–362.
35. Sartori MA, Antsaklis PJ: A simple method to derive bounds on the size and to train multilayer neural networks. IEEE Transactions on Neural Networks 2:467–471, 1991.
36. Sejnowski TJ, Rosenberg CR: Parallel networks that learn to pronounce English text. Complex Systems 1:145–168, 1987.
37. Thomason MG, Granum E: Dynamically programmed inference of Markov networks from finite sets of sample strings. IEEE Trans PAMI 8:491–501, 1986.
38. Tso MKS, Graham J: The transportation algorithm as an aid to chromosome classification. Pattern Recognition Letters 1:489–496, 1983.
39. Tso MKS, Klienschmidt P, Mitterrwiter I, Graham J: An efficient transportation algorithm for automatic chromosome karyotyping. Pattern Recognition Letters 12:117–126, 1991.
40. Wan EA: Neural network classification: A Bayesian interpretation. IEEE Trans Neural Networks 1:303–304, 1990.
41. Zimmerman SO, Johnston DA, Arrighi FE, Rupp ME: Automated homologue matching of human G-banded chromosomes. Comput Biol Med 16:223–233, 1986.

15. **Classification of chromosomes using a combination of neural networks.**
P. A. Errington and J. Graham, *Proceedings of the IEEE International Conference on Neural Networks, San Francisco, California, 1993, pp 1236-1241.*
doi:10.1109/ICNN.1993.298734

Classification of Chromosomes using a Combination of Neural Networks

Phil A. Errington and Jim Graham.
Department of Medical Biophysics,
University of Manchester, Oxford Road,
Manchester, M13 9PT, UK

Abstract- Visual analysis of microscope images containing chromosomes is an important clinical task in pre-natal diagnosis and cancer monitoring. In developing computer vision systems for analysing chromosomes images, a central task is the classification of the 46 chromosomes into 24 groups. We describe a combination of multi-layer-perceptrons for classifying isolated chromosomes and demonstrate that these perform as well as, or significantly better than a well developed statistical classifier. We suggest a method for using a competitive network to take advantage of constraints on the assignment of chromosomes to groups as a means of improving the classification rate.

I. Introduction.

The genetic material of all higher organisms is contained in a number of constituent parts of the organism's cell nuclei called chromosomes. At certain parts of the cell cycle these chromosomes exist as separate bodies which, appropriately stained, may be made visible under high resolution microscopy. Fig. 1 shows the appearance of a cell at the metaphase stage in which the chromosomes have been stained so that each exhibits a series of bands along its length (G-banding). The banding pattern, together with the chromosome length and centromere position (Fig. 2) can be used to assign the 46 chromosomes of a normal human cell into 24 groups (22 pairs of "autosomes" and two sex chromosomes: a pair of X chromosomes in the case of a female or an X and a Y chromosome in the case of a male) [15]. This classification by inspection (karyotyping) is a skilled and important task in pre-natal diagnosis of genetic abnormality and in diagnosis and monitoring of cancer. There has been considerable interest over many years in automating the analysis of chromosome images by computer vision [2], [5], [12], [13], [14]. A central issue in the development of automated systems is the specification of measurable features representing the banding pattern which cope with the considerable variability in banding appearance between cells. A range of features have been used for this purpose [2], [6], [7], [8], [12], [13], [16], [17], [21], usually derived intuitively and consequently lacking robustness to changes in preparation techniques.

An artificial neural network offers the possibility of an adaptable classifier for chromosomes [10]. Of particular interest are the feature extraction properties such models exhibit, which allow unrefined information to be presented to the classifier rather than specific intuitively defined features. This is reflected in our classification approach, as we use an artificial neural network to extract features from the raw grey level banding profile taken along the length of the chromosome. This profile is relatively easy to extract from chromosome images (Fig. 2). Additionally we use two other features representing the chromosome length and the position of its centromere (a characteristic constriction in the chromosome, see Fig. 2). This paper presents and compares the performance of an artificial neural network with a statistical classifier and discuses how the performance of the network classifier may be enhanced with the use of further neural networks.

Three extensive data sets of annotated measurements from G-banded chromosomes are used in our study, originating in Copenhagen, Edinburgh and Philadelphia. These have been used in previous classification studies using statistical methods [7], [16], [17], [21]. They cover a range of data quality, each set consisting of a large number

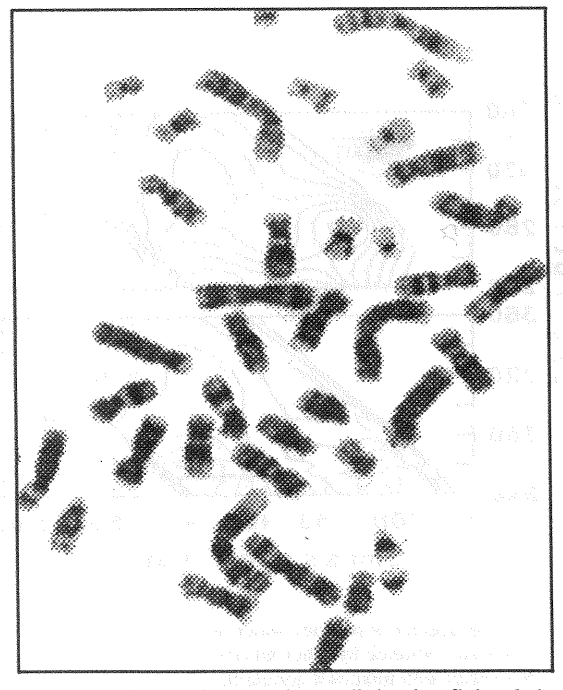


Fig. 1. An image of a metaphase cell showing G-banded chromosomes.

This work is supported by the UK Science and Engineering Research Council, Grant 90310105.

0-7803-0999-5/93/$03.00 ©1993 IEEE

TABLE I
DETAILS OF THE THREE CHROMOSOME DATA SETS USED.

Dataset	Tissue of Origin	Digitization method	No. in set	Data quality
Copenhagen	Peripheral blood	Densitometry from photographic negatives	8106	Good
Edinburgh	Peripheral blood	T. V. Camera	5469	Fair
Philadelphia	Chorionic villus	CCD line scanner	5817	Poor

of chromosome density profiles extracted from images of cells in the metaphase stage of cell division.

Of the three data sets the Copenhagen set is considered the highest visual quality, as its chromosomes were carefully measured by densitometry of photographic negatives from selected cells of high quality. The other two data sets were taken from routine material with no attempt to remove measurements errors arising from overlapped or bent chromosomes. The Philadelphia set is considered the poorer of these two, as the nature of the slide preparation method results in direct chorionic villus samples providing cells of significantly poorer visual quality than in the case of peripheral blood. Details of the three data sets appear in Table I. It should be noted that the chromosomes are all from normal human cells and not from those exhibiting abnormalities. Such cells are expected to contain 46 chromosomes of 24 classes. These 46 chromosomes consist of 22 pairs of classes 1 to 22, with either one X and one Y chromosome (in male cells) or a pair of X chromosomes (in female cells).

II. Classification Of Chromosomes

In our approach, classification takes place in two stages. The first involves classification of a chromosome independent of other chromosomes in a cell. For this task the Multi-Layer Perceptron (MLP) was selected. The bulk

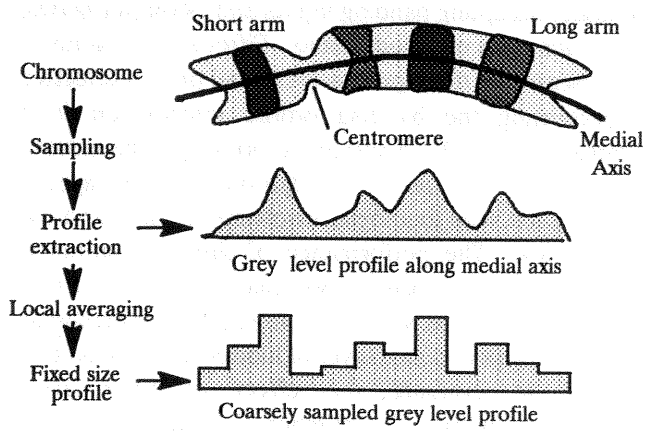


Fig. 2. An example of an extraction of a fixed length profile from a chromosome

of our work has been to modify and optimise classifiers built from this design of neural network.

During all classification experiments we use both a training and test set of data. Each of these sets is selected from approximately half of the data set under study. Two experiments are conducted, one with one half of the full data set as training data and then in a subsequent experiment as test data. Similarly the role of the other half of the data set is reversed. The classification rates we present are the mean classification rates over the two experiments.

III. First Stage Classification Using A MLP

A. Network Training Algorithm.

Preliminary work had shown the MLP to be a promising classifier for chromosome data [10] compared with other network topologies. For training our MLPs we chose a modification of the back-error propagation algorithm of Rummelhart, Hinton and Williams [19]. Our modification of the standard algorithm as described in [19] involves the use of a gradual reduction in gain (or learning rate). Initially the gain value in our network is set at a standard value (e.g. 0.1). As training proceeds two measures are monitored to select when a decrease in the gain term is required. These measures are the network classification error rate for the training data and the sum of the output node error signals for all of the training examples. The gain term is halved if the classification error rate does not decrease after 4 passes of the training data through the network. The gain is also halved if the sum of error signals increases by 10% over that observed on the previous pass of the training data set. This second measure (which is a scaled measure of the r.m.s. error between desired and actual outputs) prevents the network weights oscillating wildly with too high an original gain value, it is unlikely that such increases in the summed error signal will occur after the first few training passes.

The gain reduction mechanism permits larger values of gain to be initially used to allow considerable alteration in network weights, while allowing smaller more refined adjustments later in training for optimal classification performance. Fig. 3 shows a typical training curve for network error and classification performance using Copenhagen data.

Other algorithms (reviewed in [4]) have not yet been investigated for our MLP training, if training time becomes an issue it may be necessary to adopt one of these.

B. Use of the MLP for Chromosome Classification.

The number of samples in the banding profile of individual chromosomes in the data sets varies considerably. Profiles with up to 140 samples are present,

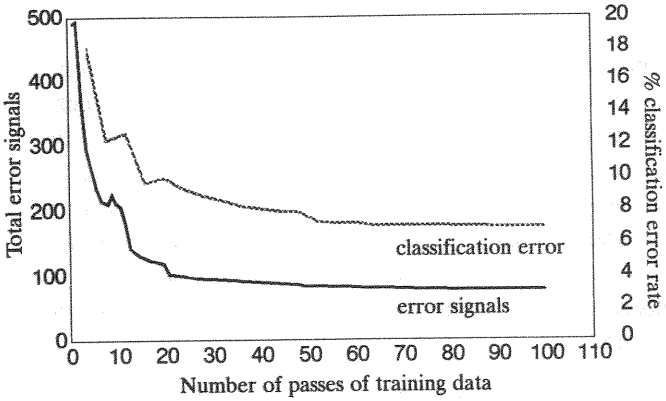


Fig. 3. The reduction in total error signals and classification error rate with training

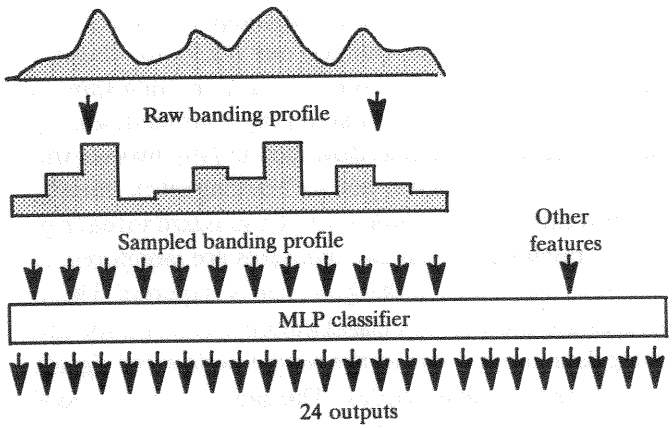


Fig. 4. Presentation of classification features to the MLP

although most profiles have approximately 90 samples. To maintain consistent inputs to the MLP the chromosome profiles are scaled to a constant length and local averaging used to produce a fixed number of averaged samples along the chromosome length (Fig. 4). These averaged inputs are presented to the MLP input nodes. If extra features are used these are presented alongside the banding profile at extra input nodes.

The network is trained so that the highest output denotes the category of the input pattern. As there are 24 classes of chromosome, 24 output nodes are required. A variable number of hidden nodes are used in one or two layers (see below).

C. Optimisation of the MLP Classifiers.

We have conducted a number of experiments to optimise our MLPs, using banding samples as the only inputs. The first stage of the optimisation involved testing the sensitivity of a particular network topology to changes in the value of the gain and momentum parameters. After varying the values of these parameters between 0.1 and 0.9 (involving 81 separate experiments in two halves) it was discovered that

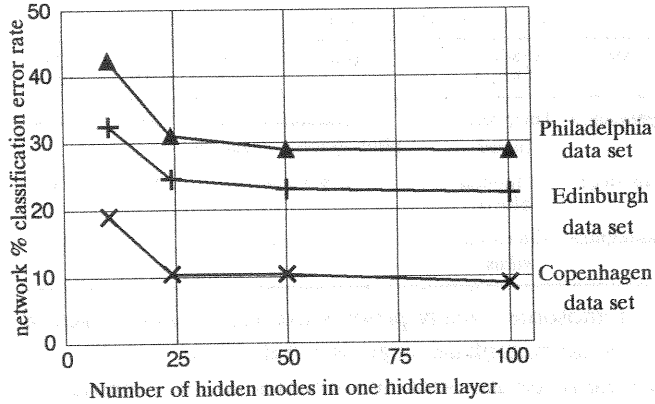


Fig. 5. The variation in classification performance of networks with different numbers of hidden nodes in one hidden layer

medium and high values of gain (greater than 0.6) and high momentum (e.g. 0.8, 0.9) resulted in unstable classifiers, while if the gain value was initially low (0.1) near optimal classification performance could be achieved with the entire range of momentum values. The result of the experiment was to select the best combination, in terms of training time efficiency, of gain and momentum which produces optimal classification performance. This combination was found to be an initial gain of 0.1 and a constant momentum value of 0.7.

Selection of an optimal topology for our problem was the next task. Although there are theoretical guide-lines to the number of hidden nodes required for a classification problem [1], [9], [20], these involve knowing something about the expected variability of the input data. As chromosome classification requires the network to cope with highly variable data, we selected the optimal topology for the MLP by experimentation.

Topology testing was performed with a fixed number of input nodes accepting banding inputs and 24 output nodes, one for each class of chromosome. Fifteen input samples were presented to the inputs as these had proved effective at representing the banding profile information in a preliminary study [10]. A variety of topology combinations of hidden nodes were tried. Initially a single hidden layer of nodes was used, with topologies involving 10, 24, 50 and 100 hidden nodes. The performance of these classifiers is shown in Fig. 5, which shows that the classification performance increases with increasing network complexity. Experiments with a second layer of hidden nodes were also conducted to evaluate the effect of their extra discriminating ability. The number of nodes in the first layer was set at 100 to reflect the best performing single hidden layer network. The results of trying 10, 24, 50 and 100 nodes in a second hidden layer is shown in Fig. 6. This

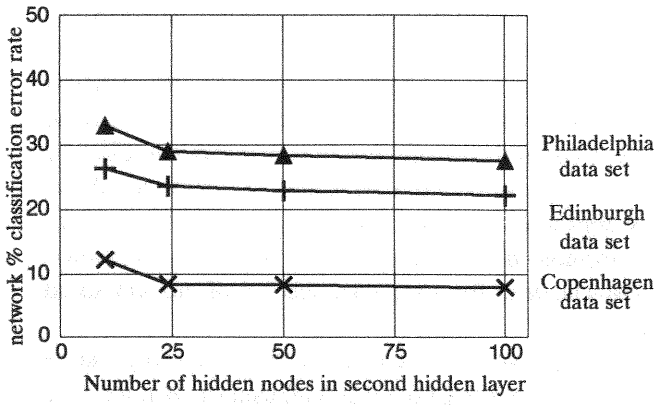


Fig. 6. The variation in classification performance of networks with a first hidden layer of 100 nodes and different numbers of nodes in a second hidden layer..

shows that there is very little variation in performance with increasing numbers of second hidden layer nodes. This is interesting as the training and classification times of the larger nets are far greater than those with fewer hidden nodes.

D. Classification Performance.

Once a good choice of topology and network parameters were made, the main advance in the performance of the classifier was achieved by including two extra features representing the length and centromere position. The centromere divides the chromosome into a long 'arm' and a short 'arm'. The ratio of the length of the short arm to that of the whole chromosome is called the centromeric index, and can be used as a representation of the centromere position, which varies depending on chromosome class. Length values are normalised to remove the effects of considerable inter-cell variation.

Three methods of including the centromeric index and length features were tried. The first involved each feature as an extra input along with banding inputs in a large MLP. The second used both features along with the banding information, but by far the most effective method was the use of an MLP pre-classifier (see Fig. 7).

Using the centromere position and chromosome length alone it is possible to classify the chromosomes into 7 broader groups, corresponding to the 'Denver' classification [3]. The pre-classifier was built to perform this broader classification, accepting the two features as inputs and producing likelihoods of membership of the 7 broader groups as outputs. This 7 group information was passed, together with the banding inputs, to a second MLP trained to produce the 24 class classification.

The optimisation of the MLP pre-classifier was performed in a similar manner to that discussed above; a number of topologies involving 2 inputs and 7 outputs were tried. The best performing of these topologies and their performance at classifying the 24 chromosome classes into the corresponding 7 Denver groups is shown in Table II.

The performances of the three inclusion methods for the centromere position and length features are shown in Table III, which indicates that a succession of two MLPs, the first performing a broad classification, later refined by a second using extra data, can out perform a single large MLP working on all the data.

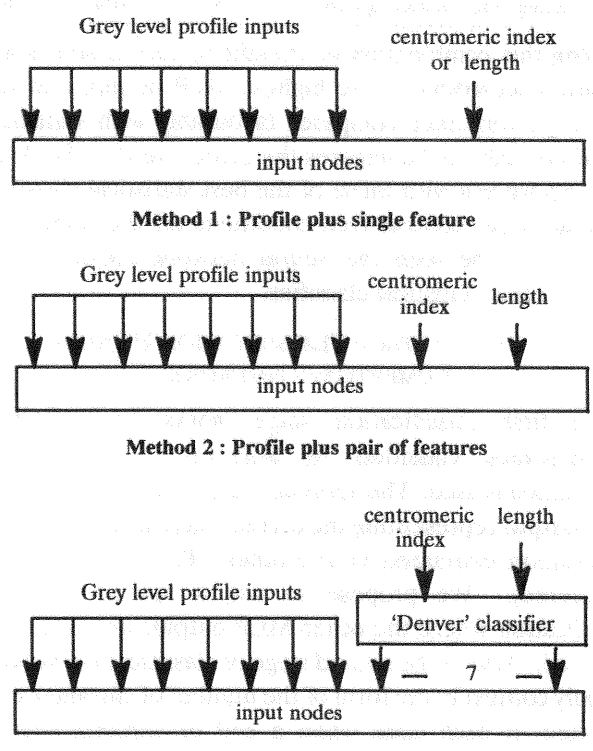


Fig. 7. Different methods of inclusion of the centromeric index and length features.

TABLE II
CLASSIFICATION ERROR RATES FOR THE BEST DENVER CLASSIFIERS USED FOR PRE-CLASSIFICATION

Data set	Network Topology	Error rate for classifying into 7 groups
Copenhagen	2–14–7	5.4%
Edinburgh	2–14–7	10.1%
Philadelphia	2–14–7	14.6%

TABLE III
CLASSIFICATION ERROR RATES OF NETWORKS USING 15 GREY LEVEL BANDING INPUTS WITH DIFFERENT REPRESENTATIONS OF CENTROMERE AND LENGTH FEATURES.

	Data set		
Features Used	Cop.	Edi.	Phi.
Banding pattern alone	8.8%	22.3%	28.6%
Banding and normalised length	8.4%	19.4%	27.6%
Banding and centromeric index	7.7%	21.0%	26.5%
Banding, length and centromeric index	6.9%	18.6%	24.6%
Banding, and 'Denver' groups	5.8%	17.0%	22.5%

Using this combination of classifiers, chromosomes are classified according to the highest MLP output, and the overall performance compares favourably with statistical classifiers. Table IV compares the error rates for the three data sets we use with those of the best statistical classifier working under context free conditions on the same data [17]. As can be seen the neural network classifier outperforms this statistical classifier.

IV. SECOND STAGE CLASSIFICATION USING A COMPETITIVE NETWORK

The first classification stage works on individual chromosomes classified in isolation; no contextual information is used. The approach also relies on the highest MLP output representing the correct class of chromosome; information contained in the other MLP outputs is not considered. We propose to use a second stage of classification where the other MLP outputs are examined and used. Also in the second stage of classification we wish to apply context in the form of the number of chromosomes expected in each class when a cell of chromosomes is classified. We are investigating the application of a competitive network to both of these tasks.

From an MLP's output vector it is possible to select not only the most likely class, but secondary and less likely classes, using the second highest output, the third highest etc [18]. By considering all the MLP outputs, it may be possible to correctly classify chromosomes mis-classified on the basis of highest MLP output alone. To test the feasibility of this approach we have trained a competitive network using chromosomes mis-classified by the MLP and used it as a post-classifier for a separate test set of mis-classified chromosomes.

The competitive network we chose to use is a single layer topology of competitive nodes trained using a 'Winner Take All' algorithm. Each node receives the same input vector and compares this to its pattern of weights (which initially are random). The node with a pattern of weights closest to the input pattern is designated the winner. During training, winning nodes alter their weight values so that they are closer to their inputs. After a period of training each node has specialised to represent a class of similar input vectors (those for which it 'won'), and can be labelled with the class(es) of these vectors. The nodes can then be used to classify input vectors, the class of each vector being decided by the label of the winning node. The vectors we use are those produced at the 24 output nodes from the MLP first stage classifier. At present no lateral inhibition or Kohonen neighbourhoods [11] are used, each winning node updating only its own weights. It may be necessary to introduce some form of refinement near class boundaries as it becomes clear how the subclasses lie in weight space.

The results of classifying mis-classified chromosomes from the Copenhagen, Edinburgh and Philadelphia data sets, using a competitive network are presented in Table V.

The performance of this classifier is encouraging. It shows that, even when the highest value in the MLP output vector does not correspond to the true class, the entire vector contains information which allows classification to be made. However the classifier is not attempting to classify all chromosomes, only those mis-classified by the MLP. It is possible to train other competitive nodes to classify

TABLE IV
COMPARISON OF CLASSIFICATION ERROR RATES FOR A NEURAL NETWORK CLASSIFIER USING BOTH BANDING AND DENVER GROUP INPUTS WITH A HIGHLY OPTIMISED PARAMETRIC CLASSIFIER.

	Classification error rate.		Significance of
Data set	Network	Parametric [17]	Improvement
Copenhagen	5.8%	6.5%	2% level
Edinburgh	17.0%	18.3%	5% level
Philadelphia	22.5%	22.8%	Non significant

TABLE V
PERCENTAGE OF CHROMOSOMES MIS-CLASSIFIED ON THE BASIS OF MLP OUTPUT WHICH ARE CORRECTLY CLASSIFIED BY A COMPETITIVE NETWORK TRAINED ON MIS-CLASSIFIED CHROMOSOMES.

	Percentage of chromosomes correctly classified	
Data set	Training data	Unseen data
Copenhagen	48.0%	23.1%
Edinburgh	42.1%	30.4%
Philadelphia	43.4%	31.7%

chromosomes correctly classified by the MLP. The nodes classifying these may then be included with nodes classifying mis-classified chromosomes. Combining 'correct' trained and 'error' trained competitive nodes in this manner, we have so far only managed to achieve a classification performance equivalent to selecting the highest MLP output as the correct class.

Our experiments involving the application of context to the classification of chromosomes also make use of a competitive network. The contextual constraint is that a cell of 46 chromosomes will possess 2 chromosomes each of classes 1 to 22, with either one X and one Y chromosome or a pair of X chromosomes. Application of this constraint has been shown to effect an improvement in the performance of statistical classifiers, [22].

We are currently investigating methods of applying this constraint using a competitive network. One method currently under consideration is to to classify all the chromosomes in a cell using a competitive network pre-trained to recognise MLP output vectors. A mechanism of penalising and rewarding competitive nodes according to how well they match the contextual constraints is applied in the winner take all competition. Nodes winning too few chromosomes in classification should therefore receive more, while those winning too many chromosomes should receive less.

V. Conclusions

Overall the application of trainable neural networks for chromosome classification has proved effective. The first stage of classification involving 2 MLPs out-performs a highly optimised statistical classifier working with the same data and splitting mechanisms [17]. We have begun investigations into the use of competitive networks in a second classification stage, with emphasis on applying contextual constraints for classifying all the chromosomes in a cell. Results so far are equivocal.

Acknowledgments

This work was greatly facilitated by the exchange of materials and ideas available within the Concerted Action of Automated Cytogenetics Groups supported by the European Community, Project No. II.1.1/13. We are grateful to Jim Piper of the MRC Human Genetics Unit, Edinburgh for permission to reproduce some of his results.

References

[1] E. B. Baum. "On the capabilities of multilayer perceptrons." *Journal Complexity* Vol 4 pp 193–215, 1988.

[2] K. R. Castleman and J. Melnyk. "An automated system for chromosome analysis– final report." *Internal document* No. 5040–30. Jet Propulsion Laboratory, Pasedena, Texas 1976.

[3] "Denver Conference. A proposed standard system of nomenclature of human mitotic chromosomes." *Lancet* Vol 1 pp. 1063–1065, 1960.

[4] S. E. Fahlman. "Faster learning variations on back-propagation: an empirical study." in *Proceedings Connectionist Models Summer School*, 1988, pp 38–51.

[5] J. Graham. "Automation of routine clinical chromosome analysis I, Karyotyping by machine." *Analytical and Quantitative Cytology and Histology*, Vol. 9 pp. 383–390, 1987.

[6] G. H. Granlund. "Identification of human chromosomes using integrated density profiles." *IEEE Trans. Biomed. Eng.* Vol. 23 pp. 183–192 1976.

[7] E. Granum. "Application of statistical and syntactical methods of analysis to classification of chromosome data." in *Pattern Recognition Theory and Application*, Kittler J. Fu KS, Pau LF, eds. NATO ASI (Oxford), Reidel, Dordreht, 1982, pp 373–398.

[8] F. C. A. Groen, T. K. Ten Kate, A. W. M. Smeulders and I. T. Young. "Human chromosome classification based on local band descriptors." *Pattern Recognition Letters* Vol. 9. pp. 211–222, 1989.

[9] S.C. Huang and Y. F. Huang. "Bounds on number of hidden neurons in multilayer perceptrons." *IEEE Trans on Neural Networks*, Vol. 2 pp. 47–55. 1991.

[10] A. M. Jennings. "Chromosome Classification Using Neural Nets," *MSc Thesis*, University of Manchester, U.K. 1990.

[11] T. Kohonen. "Self-Organisation and Associative Memory," *Series in Information Sciences*, Vol 8. Springer-Verlag, Berlin-New York-Tokyo, 1984. 2nd ed. 1988.

[12] R. S. Ledley, P. S. Ing and H. A. Lubs. "Human chromosome classification using discriminant analysis and Bayesian probability." *Comput. Biol. Med.* 10:209–218, 1980.

[13] C. Lundsteen, T. Gredes, E. Granum and J. Philip. "Automatic chromosome analysis II. Karyotyping of banded human chromosomes using band transition sequences." *Clin. Genet.* Vol. 19 pp. 26–36 1981.

[14] C. Lundsteen, T. Gerdes and J. Maahr. "Automatic classification of chromosomes as part of a routine system for clinical analysis." *Cytometry* Vol. 7 pp. 1–7, 1986.

[15] Paris Conference (1971), "Standardization in Human Cytogenetics. " *Original Article series*, 8:7. The National Foundation, New York 1972.

[16] J. Piper and E. Granum. "On fully automatic feature measurement for banded chromosome classification." *Cytometry* 10:242–255, 1989.

[17] J. Piper. "Aspects of chromosome class size classification constraint, " *CAACG Interlab meeting and Topical Workshop on High-level Classification and Karyotyping, Approaches and Tests*, University of Aalborg, 13–14 March 1991.

[18] D. W. Ruck, S. K. Roggers, M. Kabrisky, M. E. Oxley and B. W. Suter. "The Multilayer perceptron as an approximation to a Bayes Optimal Discriminant Function," *IEEE Trans Neural Networks*, Vol. 1 pp. 296–297. 1990.

[19] D. E. Rummelhart, G. E. Hinton and R. J. Williams. "Learning Internal Representations by Error Propagation." in *Parallel Distributed Processing: Explorations in the Microstructures of Cognition.* Rummelhart DE and McCelland JL (eds.), Vol. 1 Foundations, MIT Press, Cambridge, MA, 1986, pp. 318–362.

[20] M. A. Sartori and P. J. Antsaklis. "A Simple Method to Derive Bounds on the Size and to Train Multilayer Neural Networks." *IEEE Trans on Neural Networks*, Vol. 2 pp. 467–471, 1991.

[21] M. G. Thomason and E. Granum. "Dynamically programmed inference of Markov networks from finite sets of sample strings." *IEEE Trans. PAMI* Vol 8 pp. 491–501, 1986.

[22] M. K. S. Tso, P. Kleinschmidt, I. Mitterreiter and J. Graham. "An efficient transportation algorithm for automatic chromosome karyotyping. " *Pattern Recognition Letters* Vol. 12 pp. 117–126, 1991.

16. **Classification of Chromosomes: A comparative study of neural network and statistical approaches.** J Graham and P.A. Errington in *"Artificial Neural Networks in Biomedicine", P. Lisboa, E. Ifeachor and P. Szczepaniak (eds), Springer, London, pp 259-268, 2000*. doi:10.1007/978-1-4471-0487-2_19

Chapter 18

Classification of Chromosomes: A Comparative Study of Neural Network and Statistical Approaches

Jim Graham [1] and Phil A. Errington

18.1 Introduction

18.1.1 Chromosome Analysis and its Applications

In a normal human cell there are 46 chromosomes which, at appropriate stages of cell division (prophase and metaphase) can be observed as separate objects using high resolution light microscopy. Figure 18.1 shows chromosomes in a metaphase cell. The chromosomes have been stained to exhibit a series of bands along their lengths. Each chromosome also has a characteristic constriction called the centromere (indicated for some chromosomes in Figure 18.1a). Analysis of the appearance of chromosomes is routinely undertaken in hospital laboratories, for example, for diagnosis of inherited, or acquired, genetic abnormality or the monitoring of cancer treatment.

The time-consuming nature of chromosomes analysis has resulted in considerable interest in the development of automated systems based on machine vision. There is a long and noble history of work on automated chromosome analysis going back to the 1960s, when it was among a small number of problems that stimulated pattern recognition research [1]. Since the early 1980s commercially available systems have come into widespread use (e.g. [2], for a review see [3]).

[1] University of Manchester, Imaging Science and Biomedical Engineering, Stopford Building, Oxford Road, Manchester M13 9PT, UK. Email: Jim.Graham@man.ac.uk.

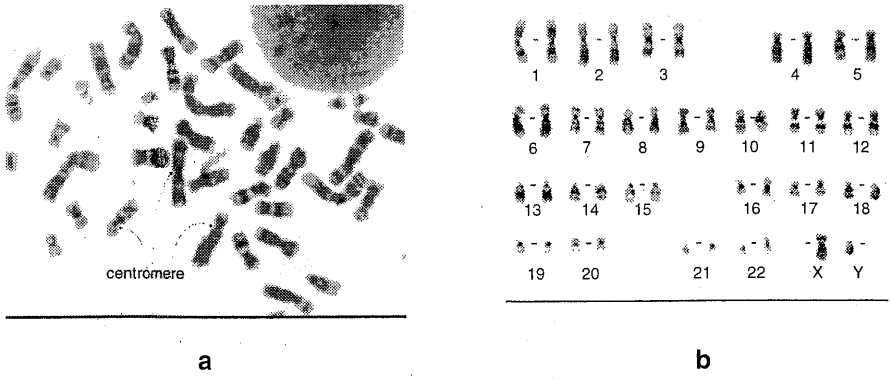


Figure 18.1: Chromosome images. (a) Metaphase chromosomes as they appear on the microscope slide. The position of the centromere (see text) is indicated for some of them. (b) A karyotype with chromosomes grouped into homologous pairs. This is a male cell, having one copy each of the X and Y sex chromosomes.

18.1.2 Chromosome Classification

The visual analysis of chromosome images, known as karyotyping, involves counting the chromosomes and examining them for structural abnormalities. To determine the significance of both numerical and structural abnormality it is necessary to classify the chromosomes into 24 classes, or groups, on the basis of their relative sizes, the patterns of bands and the centromere positions. The result of the classification is a karyotype, often displayed as a tabular array of chromosomes arranged in their groups (Figure 18.1b). Twenty-two of these groups normally contain two homologous (structurally identical) chromosomes. The other two groups contain the sex chromosomes X and Y. In the case of a normal male cell, the X and Y groups contain one chromosome each; in a female cell there is a homologous pair of X chromosomes and the Y group is empty. In the absence of banding information it is possible to assign chromosomes to seven groups (the 'Denver' groups) on the basis of size and centromere position alone. The lay-out of the karyotype in Figure 18.1b reflects the Denver grouping.

Automated systems analyse the microscope images by first isolating (segmenting) the chromosomes, measuring the chromosomes' sizes, shapes and banding patterns and using these measurements in a classifier to assign the chromosomes to appropriate groups (see [4]). The chromosome classification performance of these systems depends on the type of material used, but is typically in the range 6%-18% misclassification [5] which compares poorly with visual classification by cytotechnician [6]. All automated systems in clinical use operate interactively, allowing an expert user to correct machine errors in image segmentation, feature extraction and classification, resulting in useful performance [4].

The aim of the study described here was to investigate the possibility of using a neural network as a chromosome classifier. The form of the data provides a natural input for neural networks. The existence of a number of expertly annotated data-sets (see below) and the large body of work on classical statistical and syntactic classifiers using these data provided an opportunity to evaluate the performance of network classifiers in a well-studied domain. In particular, we could attempt to answer questions such as: Does a network classifier offer any performance advantages? Do the 'feature extraction' properties of neural networks remove the need for intuitively defined classification features? What is the effect of training set size on the performance of network classifiers?

18.1.3 Experimental Data

Table 18.1: Summary of the data-sets of chromosome measurements

Data-set	Tissue of origin	Data acquisition method	No. of chromosomes	'Quality' of chromosome images
Copenhagen	Peripheral blood	Densitometry	8106	High
Edinburgh	Peripheral blood	Analogue TV camera	5469	Medium
Philadelphia	Chorionic villus	Linear CCD array	5817	Low
CPR	Amniotic fluid	Analogue TV camera	110636	Medium (routine)

The network was trained and tested using several data-sets of annotated measurements from banded chromosomes. The characteristics of these data-sets are summarised in Table 18.1. The data in the Copenhagen set were obtained by densitometry of photographic negatives of selected cells of good appearance. Chromosomes involved in touches or overlaps were rejected from the data-set, so the visual 'quality' of the chromosomes was high. The Edinburgh and Philadelphia sets were digitised directly from microscope images of routine material. The preparation techniques used in chorionic villus sampling resulted in the poor visual quality of the chromosome images in the Philadelphia set. Furthermore, the cells in the Edinburgh set were selected to have very few overlapping chromosomes. These three data-sets gave a reasonably large quantity of data for network training and testing covering a range of quality representative of that found in a real implementation. The 'CPR' data-set also originated from Copenhagen and consisted of a very large set of metaphase chromosomes from amniotic fluid cells, which occurred in the routine clinical workload, analysed

using a commercially available chromosome analysis system. The visual quality of chromosomes in amniotic fluid preparations was intermediate between that of blood and chorionic villus samples. No attempt was made to edit this data-set for 'quality', and it included chromosomes which overlapped in the original image resulting in obscuration of part of the banding pattern. Given the routine nature of the acquisition material, it is inevitable that there will be a proportion of misclassified chromosomes.

18.2 The Neural Network Classifier

18.2.1 Representation of Chromosome Features

The features which can be used for chromosome classification are the size, centromere position and banding pattern. An important issue for automatic classification is the representation of the banding pattern. Several different classifiers have been reported using statistical or syntactic approaches (e.g. [7], [8], [9], [10]. Each of these involves the extraction of a number of intuitively defined features to represent the banding pattern. These features are usually associated with the density profile: a one-dimensional pattern obtained by projecting the chromosome's density onto its centre line, which reflects the largely linear organisation of the chromosome structure (Figure 18.2). There is a risk of losing information in parameterising the banding profile. This risk may be reduced by using the density profile itself as the banding representation. This type of one-dimensional pattern is a natural form of input for an artificial neural network, and we chose to use the profile samples themselves as banding features, rather than defining a set based on intuition.

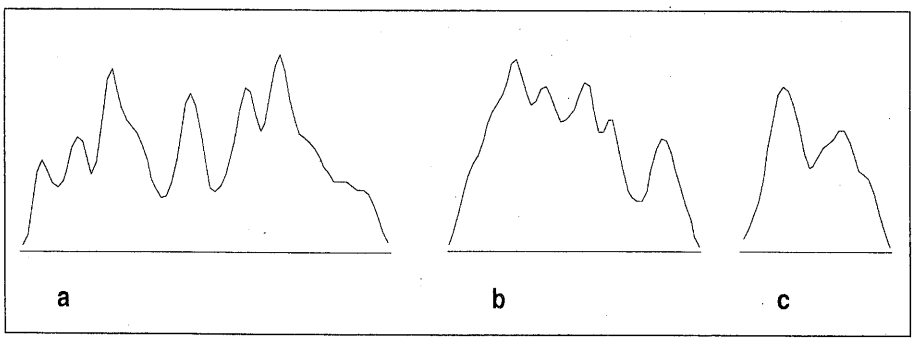


Figure 18.2: Chromosome density profiles. These are typical one-dimensional patterns obtained by projecting the chromosome's density onto its centre line. The profiles illustrated here correspond to (a) Chromosome 1, (b) chromosome 6 and (c) chromosome 17.

The inputs to our network were therefore:

Chromosome Size: this may be measured either as the length of the chromosome or its area; the two measures are very highly correlated. In the data-sets used in this study, the length was used.

Centromeric index: the centromere divides the chromosome into long and short 'arms' (Figure 18.1). The centromeric index (CI) is the ratio of the length of the short arm to that of the whole chromosome, and gives a measure of shape.

Banding profile: the number of samples representing the banding profile can vary between 10 and 140 depending on the class of the chromosome and the state of contraction of the cell in which it occurred. The classification module requires a consistent input vector and all banding patterns must therefore be represented by the same number of samples. We found that a constant number of samples could be used to represent the profile, irrespective of the original chromosome length, and that this number could be quite small (as low as 15 samples for all profiles) with very little loss of classification accuracy [11], [12]. The use of a uniform number of samples meant that the profiles of long chromosomes had to be subsampled by local averaging, and the short chromosomes oversampled by interpolation.

18.2.2 Network Topology and Training

In this application we have a classification problem using continuous-valued inputs, where the classes are well defined and expert classification of the training data is available. It is a clear case for a Multi-Layer Perceptron (MLP). The training algorithm employed was the classical backpropagation method [13], using a strategy of progressive reduction in gain (learning rate) during the training. Two measures were used to monitor performance: total network error and classification accuracy on the training data. These measures are not identical (due to the fact that the classification result is determined only by the highest output) and both are useful measures of performance. During training, the gain was halved if the total net error had increased by more than 10%, or the classification performance had not improved over the previous presentation of the training data. Training was halted when the value of gain dropped below 10^{-4}. The gain reduction strategy proved extremely valuable in this application. Table 18.2 shows the mis-classification rates on training data after convergence of networks trained to classify banding features alone using a subset of the Copenhagen data [11]. There is a clear advantage in using gain reduction and in using two performance characteristics to monitor the network.

Network parameters were set empirically. The number of classes fixes the number of output nodes at 24. The number of input nodes was the lowest that could be used without loss of accuracy in classifying the banding profile 15. The size

of the single hidden layer was increased until classification improvements became insignificant at 100 nodes. The best combination of initial gain and momentum values was 0.1 and 0.7.

Table 18.2: The effect on classification performance of gain reduction during training, monitored using total network error and accuracy of classification of the training data

Training Strategy	No gain reduction	Gain reduction (net error only)	Gain reduction (net error and classification accuracy)
Misclassification Rate	53(%)	12(%)	4(%)

18.2.3 Incorporating Non-Banding Features

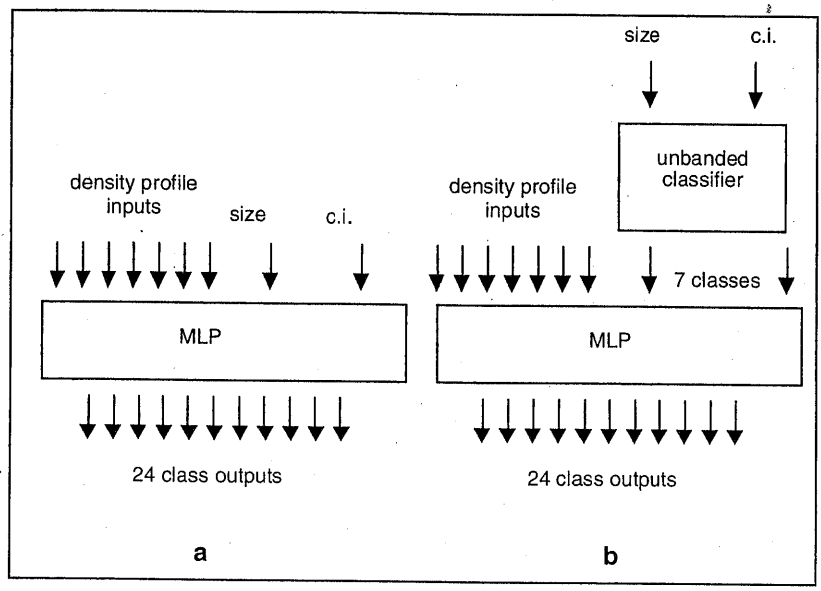


Figure 18.3: Two schemes for incorporating size and centromeric index (c.i.) along with the banding profile features. (a) The two morphological features are simply added as additional MLP inputs, scaled to be of similar magnitude to the profile values. (b) Size and c.i. are used as inputs to a pre-classifier: a MLP trained to assign chromosomes to the seven groups of the 'Denver' convention for unbanded classification (see text).

In principle, it is possible to classify chromosomes on the basis of the banding pattern alone. However, the size and centromeric index are extremely powerful classification features, and must be included for most accurate results. These

features are quite a different source of data from the banding pattern, and we are here performing a kind of data fusion.

The simplest way to add the new features would be to use them as inputs to the network in addition to the banding features as shown in Figure 18.3a. However, we have already noted that size and centromeric index can be used in isolation to partially classify the chromosomes into seven Denver classes. We therefore investigated an alternative form of input, in which these two features were processed by a pre-classifier, also an MLP, trained to produce a Denver classification. This network has, by definition, two inputs and seven outputs. The number of nodes in the hidden layer was determined empirically to be 14. The outputs of the pre-classifier were then used along with the banding features as inputs to the main classifier (Figure 18.3b). Classification results on the three sets of chromosome data are discussed in the next section.

18.3 Classification Performance

18.3.1 Classification Experiments

The relative sizes and overall densities of chromosomes in a cell are fairly consistent; however absolute lengths and densities can vary between cells. Length and density measures were therefore normalised to a constant value for each cell before classification. Classifiers using the networks described in the previous section were evaluated using the Copenhagen, Edinburgh and Philadelphia data-sets. 'Hold-out' cross validation was used by splitting each data-set in half. Table 18.3 shows the results of classification using banding data alone, banding data supplemented with size and centromeric index features and banding data with Denver classes. That is to say the networks consisted of 15+X inputs, 100 hidden nodes and 24 outputs, where X=0, 2 or 7. Clearly size and centromeric index are powerful features, and they are used to greater advantage in the pre-classifier.

Table 18.3: Misclassification rates for the network classifier using three different forms of input

Input Features	**Data-set**		
	Copenhagen	Edinburgh	Philadelphia
Profile only (15-100-24)	8.8%	22.3%	28.6%
Profile +size and CI (17-100-24)	6.9%	18.6%	24.6%
Profile +Denver classes (22-100-24)	5.8%	17.0%	22.5%

18.3.2 Comparison with Statistical Classifiers

A feature of developing a neural network classifier for chromosome analysis is the possibility of comparing a network solution to classical statistical methods. There have been a number of approaches to chromosome classification, but the most successful prior to this study was that of Granum [8], subsequently greatly refined [5]. This method extracts banding features using 'weighted density distributions'; essentially the banding profile is multiplied by a number of intuitively defined weighting functions, approximating a set of basis functions for the banding pattern. The features extracted from the density profiles in this way are combined with length and CI features, and classified using a maximum likelihood classifier. Table 18.4 compares the best network performance with the statistical method of Piper and Granum [5] in performing context-free classification of individual chromosomes. The network performance is significantly better for the Copenhagen and Edinburgh data-sets and identical for the Philadelphia data-set.

Table 18.4: Classification performance of two MLP configurations compared with that of a parametric statistical classifier [5].

Classifier	Data-set		
	Copenhagen	Edinburgh	Philadelphia
Network classifier	5.8%	17.0%	22.5%
Parametric classifier	6.5%	18.3%	22.8%
Significance of MLP improvement	2% level	5% level	not significant

18.3.3 The Influence of Training-Set Size

The Copenhagen, Edinburgh and Philadelphia sets are quite large by the standards of classification studies. However, when we consider that we need to distinguish 24 classes, the numbers of training data are still far from enormous. The very large CPR data-set allows us to investigate the effects of training set size on network performance. The following experiment was conducted. A 'profile-only' classifier was constructed (15-100-24 nodes) and trained on subsets of the CPR data with a range of sizes: 3450, 6900, 13800, 27646 and 55337 chromosomes. At the smallest size, 32 data-sets were available for training and at the largest there were two. Each trained classifier was used to classify the same data on which it was trained (often referred to as the 'apparent' error rate) and also an independent evaluation set (a 'hold-out' evaluation). In each case the training and evaluation sets were approximately equal in size.

The 'apparent' error rate will clearly be biased optimistically. The independent-set evaluation will produce a result which is unbiased with respect to the classifier resulting from a particular set of training data. However, due to the limited size of the training set, the hold-out classification will be biased pessimistically with respect to a classifier trained on a much larger, more representative, training set. As the training set size increases, the classification estimates of the two evaluations should converge. In the limit, when the training set is large enough to be fully representative, the 'apparent' error rate will cease to be optimistic and the independent-set evaluation set will cease to be pessimistic.

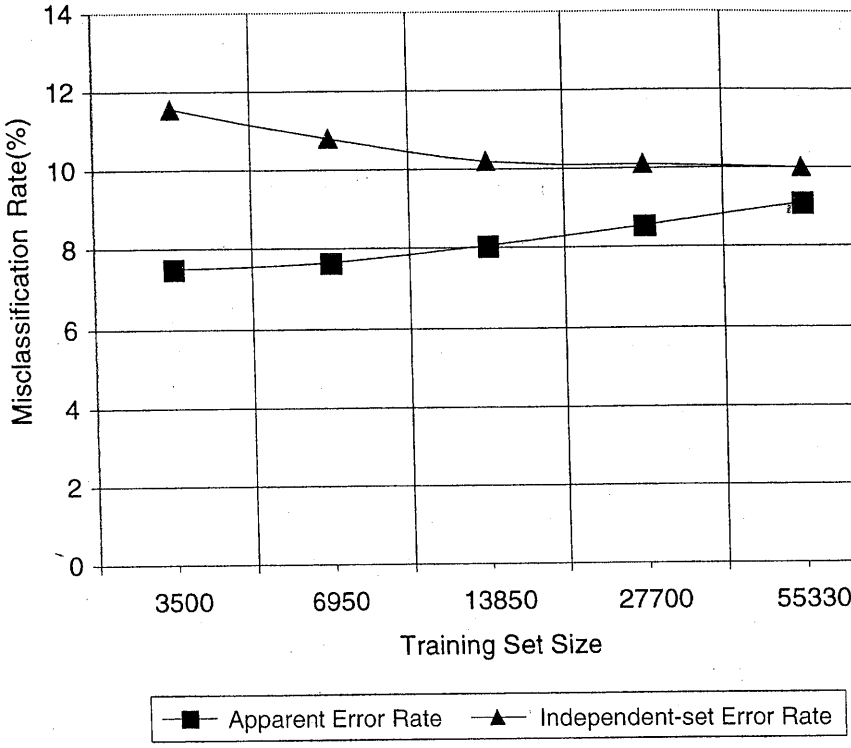


Figure 18.4: Classification performance as a function of the training set size. Each point is the average misclassification rate for a network using banding profile features of the CPR data-set.

Figure 18.4 shows the results of the classification experiments. The lower curve represents the 'apparent' error estimates and the upper curve, the independent-set estimates. As expected the difference between the results diminishes as the data-sets get larger. By extrapolation we can estimate that the curves would converge at a training set size of about 100,000 examples.

Piper [14] conducted a similar study to this one on the CPR data-set using the weighted density distribution features in a maximum likelihood classifier [5]. He derived curves similar to Figure 18.4, in which 'apparent' and independent-set error estimates converged as the training set size increased. In particular, he considered the question of how large a training set must be to be considered 'representative', and concluded that a useful rule of thumb is that the number of training examples should be about ten times the number of free parameters to be set. (This heuristic value is much larger than earlier proposals, e.g. that of ten times the number of classes, in James [15]). Taking the number of free parameters in our neural network classifier to be the number of weights to be set, we would agree in general terms with Piper's conclusion, but we would increase his heuristic multiplier by a factor of two or three.

18.4 The Use of Context in Classification

18.4.1 The Karyotyping Constraint

It is possible to effect significant improvement on the classification of individual chromosomes by application of the *karyotyping constraint*, namely that there are exactly two chromosomes in (almost) all classes. This is an example of a global constraint, affecting the classification of all the chromosomes in a cell. It must be applied as a reassignment of chromosomes following context-free classification. There have been many approaches to applying this constraint from 'cascade' rearrangement [16] through to Genetic Algorithms [17]. Tso and Graham [18] showed that the constraint can be cast in the form of the Transportation problem, well known in Operations Research. The general context of this problem is that of transporting goods from sources to destinations along different routes of known cost. The requirement is to achieve the most economical transportation subject to the requirements at each destination. The problem can be solved by a linear programming algorithm. The chromosome classification problem is a special case of the Transportation problem, in that the destinations (the individual chromosomes) all have a demand of unity on the sources (the chromosome classes). This leads to a particularly efficient solution [19]. The importance of the Transportation Algorithm in this context is that, since the constraints are linear (two chromosomes in each class), the algorithm is guaranteed to generate the globally maximum *a posteriori* likelihood of assignment subject to the constraint, given the input classification probabilities of the individual chromosomes. Transportation rearrangement achieves misclassification rates which have not been bettered by any other approaches. Genetic search [17] has produced comparable results, albeit at significant computational cost.

18.4.2 Applying the Constraint by a Network

We wished to apply a network solution to the problem of applying contextual constraints. Our approach to this was via a competitive network, the inputs to which were the outputs of the MLP classifier. In addition to developing an algorithm for applying context we wished to explore the possibility of recognising the misclassified chromosomes and reassigning them. The rationale was that, since the MLP assignment was based on the highest output node, there might be further information available in the pattern of output nodes. We trained a competitive network by a winner-take-all algorithm [20], using as inputs the MLP outputs corresponding to incorrectly classified chromosomes. The network did recognise clusters of input vectors in this data holding out the possibility of recognising these cases [21]. The misclassified chromosomes are, of course, unknown, and it is necessary to train the network with all output nodes. The network resulting from this training proved unable to recognise the misclassified cases. Classification by competitive network using the full set of MLP outputs gave no advantage over classification by highest output only.

While this result was disappointing, we were able to modify the application of a competitively trained network to apply the karyotyping constraint. The modified algorithm was as follows:

1. Train the competitive network in a winner-take-all competition on the MLP outputs and label the resulting nodes. Each node is assigned an *influence value* α – a weight factor applied to the distance metric – initially set to 1.0.
2. Present the MLP outputs from an unseen cell. Each vector is labelled according to the nearest node in the trained network.
3. Apply the constraint rule. The number of chromosomes observed in each class n_o is compared with the desired number of chromosomes n_d. The influence value is altered accordingly:
   if $n_o = n_d$ then $\alpha = \alpha$
   if $n_o > n_d$ then $\alpha = \alpha - \varepsilon$
   if $n_o < n_d$ then $\alpha = \alpha + \varepsilon$
   This allows chromosomes in the borders between clusters of nodes of different classes to be 'released' by one and 'captured' by the other (see Figure 18.5).
4. Normalise all the influence values and repeat step 3 a fixed number of times.
5. From the rearrangements observed in all these iterations choose (a) the one with the largest number of classes in which $n_o = n_d$, and of these (b) the one with the highest joint likelihood of class membership (as defined by the MLP output values).

Empirically, a value of ε of 0.01 was found to give best results, requiring 100-200 passes of the reassignment algorithm.

18.4.3 Results of Applying the Context Network

Table 18.5 shows the effect of reassigning classifications using the competitive network compared with application of the Transportation Algorithm to the MLP outputs.

The performance of the competitive network rearrangement is very close to that of the Transportation Algorithm. This is rather pleasing as the Transportation Algorithm is guaranteed to produce a globally maximum likelihood assignment subject to the constraints. (This is not the same thing as guaranteeing the lowest error rates, due to errors in the calculation of individual likelihoods. See for example the results of Edinburgh data shown in Table 18.4.)

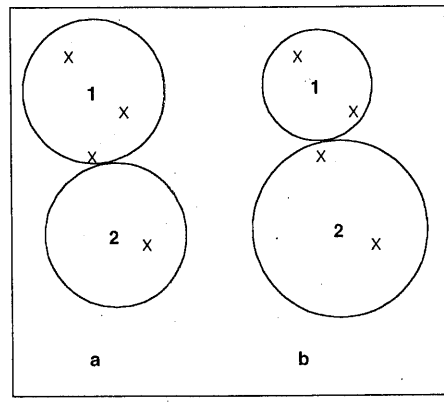


Figure 18.5: Alteration of influence values of competitively trained nodes. (a) Neighbouring nodes have been trained to assign their closest objects to classes 1 and 2. Initially, both nodes have equivalent influence values. Three objects are assigned to class 1 and one to class 2. (b) Application of the context rule results in the influence of node 1 decreasing, while that of node 2 increases, resulting in a chromosome being moved from class 1 to class 2. In this case the karyotyping constraint is satisfied.

Table 18.5: Results of applying context using a competitively trained network and the Transportation Algorithm.

Context Application	Data-set		
	Copenhagen	Edinburgh	Philadelphia
None (highest MLP output)	5.8%	17.0%	22.5%
Competitive Network	4.4%	14.2%	19.4%
Transportation Algorithm	4.2%	14.4%	18.9%

18.5 Conclusion and Discussion

Chromosome classification is an important element in automated cytogenetic analysis. The classification problem in this case is far from trivial; most classification problems have considerably fewer than 24 classes. We have constructed a chromosome classifier using a MLP network whose performance equals or betters that of a well-developed classifier using traditional statistical methods. The form of the network is standard, with the exception that known properties of the classification features allowed the network to be 'factored' into two steps to achieve optimum classification performance.

18.5.1 Comparison with Statistical Classifiers

The results show that a network classifier can give higher classification accuracy than a classical parametric method in context–free classification. While the improvement was statistically significant, however, it was still small in absolute terms. We did better, but not by so much that it makes a real difference. The classification performance of both types of classifier is good for the Copenhagen data, probably acceptable for data of routine quality, such as is found in the Edinburgh set, and inadequate in the case of the poor quality Philadelphia data. The development costs of the network classifier were arguably appreciably smaller, since the time from proposing the concept to arriving at a final configuration was considerably shorter and involved less manpower than was the case for the conventional classifier. From an implementation point of view, the network classifier is likely to be more adaptable. Our experience is that the best network parameters (topology, gain, and momentum) are stable in the face of wide variation in the quality of data. However, it is unlikely that a single 'hard-wired' network would be adequate for any implementation. This is because there is a tendency, in clinical investigations, to use chromosomes at less contracted phases for routine analysis. The longer chromosomes exhibit more bands, and therefore greater resolution of clinical information. The effect on our classifier would be that the appropriate rate of resampling of the density profile would need to be reinvestigated. This, however, is a fairly mechanical process, and it would be a straightforward matter to arrive at an appropriate sampling resolution (and hence number of input nodes) for a particular application.

The classification of chromosomes has been widely studied, and most sensibly designed classifiers exhibit similar performance. Lerner [22] has also trained a MLP classifier, and tested it using the Edinburgh data-set. (He also used another data-set, but used that only to classify five of the twenty four classes.) He paid rather more attention to selecting the density profile features to be used than we did in our sub-sampling approach but, interestingly, found that best results were obtained with 15 samples. His classification used a two stage classifier identical in

organisation to the one described above, size and centromeric index being used by a MLP to provide seven 'Denver' classes, which were then combined with the 15 density features as input into a second stage classification. The reported classification error rate on the Edinburgh data-set was 16.4%, the slight improvement over the result in Table 18.3 possibly being due to the more carefully chosen density features.

One advantage of a network classifier is that, in building one, we make no assumptions about the underlying distributions of the classification features. This can allow a network to out-perform a statistical classifier if the latter makes assumptions (for example about normality) that do not hold. Conversely, if the assumptions are realistic, then a statistical classifier should do well. This appears to be the case for chromosome classification. Kleinschmidt et al [23] have conducted a study in which they 'pulled out all the stops' available to a parametric classifier, using every available feature, regularising the covariance matrices to produce maximum discrimination and rearranging using the Transportation Algorithm. They report misclassification rates of 2.0%, 11.2% and 14.7% for the Copenhagen, Edinburgh and Philadelphia data-sets respectively. This is probably the last word in classification studies on banded chromosomes. (It is worth noting that the misclassification rates for context-free classification reported in this study were 5.3%, 16.9% and 22.4% for the three standard data-sets, which are still barely distinguishable from the results in Table 18.4.) A number of companies market computer systems for automated karyotyping. They do not tend to state explicitly which classification algorithms they use. Experience would suggest that so long as they make accurate feature measurements, the nature of the classifier does not matter much. The algorithms implemented will be determined largely be the experience and inclinations of the engineers concerned.

18.5.2 Training Set Size and Application of Context

The existence of a large annotated set of chromosome features allowed us to investigate the effects of training set size on classification performance. We concluded that the number of training samples required is about 10-30 times the number of weights to be trained. This conclusion is in broad agreement with Piper [14], who conducted a similar study using a maximum likelihood classifier taking the class means and covariances as the free parameters to be set. While this conclusion may provide a useful rule of thumb, it should be borne in mind that both studies were conducted with the same data-set. Extrapolation to other circumstances should be carried out with care. Again, the parametric classifier probably benefits from the assumptions about the distributions being fairly realistic in this case.

We have been able to apply global contextual rules in a network classifier using competitive learning. The performance of the rule assisted classification compares

favourably with the Transportation Algorithm, which provides the optimum solution. While the competitive network offers no advantage in this case, we note that we are, to some extent, fortunate in this problem in that the global constraints are linear (expected class sizes), allowing the Transportation Algorithm to be used. Nonlinear constraints (such as 'two chromosomes in class one unless there are three in class four, in which case class one is empty') could not be applied in this way. The competitive network is not restricted to linear constraints. As long as a rule can be stated and tested, it provides a mechanism for post-processing context-free classifications which comes close to optimum performance.

18.5.3 Biological Context

In recent years the practice of cytogenetics has been transformed by the adoption of new techniques from molecular genetics. These techniques make it possible to hybridise chromosomes with very specific probes (*in-situ* hybridisation). If these probes carry fluorescent labels (fluorescence *in-situ* hybridisation, or FISH), then it is very easy to visualise any region of the chromosome and to identify abnormalities which may be too small to be resolved by normal microscopy [24]. Extensions of these techniques, such as Comparative Genomic Hybridisation (CGH) can allow highly sensitive analysis of over or under-representation of genetic material which may be associated with cancers or other clinical conditions [25].

These developments in themselves do not make a great impact on the need for banded analysis systems and classifiers. They extend the capabilities of the cytogenetic laboratory, rather than replacing existing methods. More significantly, the recent development of FISH using multi-colour fluorochromes (m-FISH) [26], results in a different approach to karyotyping, and manufacturers of automated systems are offering colour classifiers. For the time being, the 'bread and butter' work of cytogenetics laboratories still comprises banded karyotyping. The companies that sell systems for capturing and analysing colour fluorescence images need to be able to provide karyotyping systems, with classifiers, as well. Given the pace of development of molecular techniques, it is difficult to know how long that will continue to be the case.

Acknowledgement

The data were made available within the Concerted Action of Automated Cytogenetics Groups supported by the European Community (Project No II.1.1.13). We are grateful for the co-operation of colleagues within this project.

References

[1] Ledley, R. S. High speed automatic analysis of biomedical pictures. *Science*, 146:216-223, 1964.

[2] Graham, J. Automation of routine clinical chromosome analysis I. Karyotyping by machine. *Analytical and Quantitative Cytology and Histology*, 9:383-390, 1987.

[3] Lundsteen, C., and Martin, A. O. On the selection of systems for automated cytogenetic analysis. *American Journal of Medical Genetics*, 32:72-80, 1989.

[4] Graham, J., and Piper, J. Automatic Karyotype Analysis. *Chromosome Analysis Protocols. Vol. 29.* J.R. Gosden (Ed.) Humana Press, 1994, pp. 141-185.

[5] Piper, J., and Granum, E. On fully automated feature measurement for banded chromosome classification. *Cytometry*, 10:242-255, 1989.

[6] Lundsteen, C., and Granum, E. Visual classification of banded chromosomes, I. Karyotyping compared with classification of isolated chromosomes. *Annals of Human Genetics*, 40:87-97, 1976.

[7] Granlund, G. H. Identification of human chromosomes by integrated density profiles. *IEEE Transactions on Biomedical Engineering*, 23:182-192, 1976.

[8] Granum, E. Application of statistical and syntactic methods of analysis to classification of chromosome data. *Pattern Recognition Theory and Applications*. J. Kittler, K.S. Fu, and L.S. Pau (Eds.) D. Reidel, 1982, pp. 373-397.

[9] Groen, F. C. A., ten Kate, T. K., Smeulders, A. W. M., and Young I. T. Human chromosome classification based on local band descriptors. *Pattern Recognition Letters*, 9:211-222, 1989.

[10] Gregor, J., and Thomason, G. Hybrid pattern recognition using Markov networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:651-656, 1993.

[11] Jennings, A. M., and Graham, J. A neural network approach to automatic chromosome classification. *Physics in Medicine and Biology*, 38:959-970, 1993.

[12] Errington, P. A., and Graham, J. Application of artificial neural networks to chromosome classification. *Cytometry*, 14:627-639, 1993.

[13] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning internal representations by error propagation. *Parallel Distributed Processing: Explorations in the Microstructures of Cognition. Vol. 1.* D.E. Rumelhart and J.L. McClelland (Eds.) MIT Press, 1986, pp. 318-362.

[14] Piper, J. Variability and bias in experimentally measured classifier error rates. *Pattern Recognition Letters*, 13:685-692, 1991.

[15] James, M. *Classification Algorithms*. Collins, London, 1985.

[16] Piper, J. Classification of chromosomes constrained by expected class size. *Pattern Recognition Letters*, 4:391-395, 1986.

[17] Piper, J. Genetic algorithm for applying constraints in chromosome classification. *Pattern Recognition Letters*, 16:857-864, 1992.

[18] Tso, M. K. S., and Graham, J. The Transportation Algorithm as an aid to chromosome classification. *Pattern Recognition Letters*, 1:489-496, 1983.

[19] Tso, M. K. S., Kleinscmidt, P., Mitterreiter, I., and Graham, J. An efficient Transportation Algorithm for automatic chromosome karyotyping. *Pattern Recognition Letters*, 12:117-126, 1991.

[20] Rumelhart, D. E., and Zipser, D. Feature discovery by competitive learning. *Parallel Distributed Processing: Explorations in the Microstructures of Cognition. Vol. 1.* D.E. Rumelhart and J.L. McClelland (Eds.) MIT Press, 1986, pp. 151-193.

[21] Errington, P. A., and Graham, J. Classification of chromosomes using a combination of neural networks. In *IEEE International Conference on Neural Networks* (1993), San Francisco, CA. pp. 1236-1241.

[22] Lerner, B. Toward a completely automatic neural-network-based human chromosome analysis. *IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetics*, 28:544-552, 1998.

[23] Kleinschmidt, P., Mitterreiter, I., and Piper, J. Improved chromosome classification using monotonic functions of mahalanobis distance and the transportation method. *ZOR - Mathematical Methods of Operations Research*, 40:305-323, 1994.

[24] Trask, B. J. Fluorescence in situ hybridisation: applications in cytogenetics and gene mapping. *Trends in Genetics*, 7:149-154, 1991.

[25] Kallioniemi, A., Kallioniemi, O.-P., Sudar, D., *et al.* Comparative genomic hybridisation for molecular cytogenetic analysis of solid tumours. *Science*, 258:818-821, 1992.

[26] Speicher, M. R., Ballard, S. G., and Ward, D. C. Karyotyping human chromosomes by combinatorial multi-fluor FISH. *Nature Genetics*, 12:368-375, 1996.

17. **A Neural Network Classifier for Chromosome Analysis**. J. Graham, in *"Handbook of Neural Computation" Chapter G4.3. E. Fiesler and R. Beale (eds.) Oxford University Press and IOP Publishing, 1996.*

G4.3 A neural network classifier for chromosome analysis

Jim Graham

Abstract

Analysis of chromosomes is an important and time-consuming task in the diagnosis of inherited or acquired genetic abnormality. Machine vision systems can contribute to the visual inspection of microscope images and the assignment of chromosomes to 24 classes is a critical stage in this analysis. A multilayer perceptron classifier has been developed for use in an automated chromosome analysis system. The inputs to the classifier are chromosome size, centromere position and a representation of the banding pattern measured from microscope images of dividing cells. The outputs are likelihoods of class membership. Optimum performance was obtained by factoring the classifier into two networks, one using size and centromere position alone to provide a first assignment into seven groups, followed by a second step in which the banding information was incorporated to give a final classification. The network is trained by backpropagation and considerable advantage is obtained by using a strategy of gain reduction using both total error and classification accuracy as network monitoring parameters. Classifier performance was tested on fairly large sets of chromosome measurements covering a representative range of data quality. Overall classification accuracy was found to equal or exceed that of a well developed statistical classifier applied to the same data.

G4.3.1 Introduction

In a normal human cell there are 46 chromosomes which, at an appropriate stage of cell division (metaphase), can be observed as separate objects using high-resolution light microscopy. Appropriately stained they show a series of bands along their length and a characteristic constriction called the centromere. Figure G4.3.1(a) shows a typical metaphase cell, stained to produce the most commonly used banding appearance (G-banding). Chromosome analysis, which involves visual examination of these cells, is routinely undertaken in hospital laboratories, for example, for diagnosis of inherited or acquired genetic abnormality or monitoring of cancer treatment.

This visual analysis, known as karyotyping, involves counting the chromosomes and examining them for structural abnormalities. To determine the significance of both numerical and structural abnormality it is necessary to classify the chromosomes into 24 groups on the basis of their relative size, the pattern of bands and the centromere position (see figure G4.3.1). Twenty-two of these groups normally contain two homologous (structurally identical) chromosomes. The other two groups contain the sex chromosomes X and Y. In the case of a normal male cell, the X and Y groups contain one chromosome each; in a female cell there is a homologous pair of X chromosomes and the Y group is empty.

The time-consuming nature of chromosome analysis has resulted in considerable interest in the development of automated systems based on machine vision. A number of such systems are now in routine use in many hospitals (e.g. Graham 1987, Graham and Pycock 1987, for a review see Lundsteen and Martin 1989). The processing stages in analyzing the microscope images are illustrated in figure G4.3.2. Chromosomes are isolated from the images, measurements are made of chromosome size, shape and

© 1997 IOP Publishing Ltd and Oxford University Press

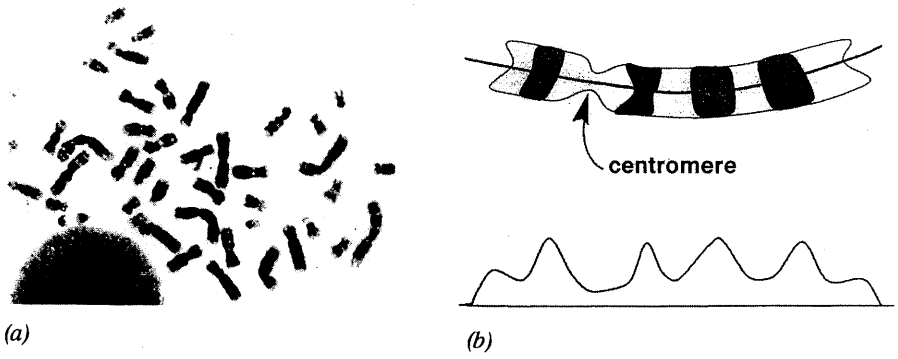


Figure G4.3.1. Chromosomes and chromosome features. (*a*) A cell at metaphase. The individual chromosomes show the banding pattern (G banding) produced by staining. (*b*) Schematic drawing of a chromosome showing the position of the centromere. The density profile (below) is formed by projecting the density onto the curved centerline.

banding pattern, these measurements are used in a classifier to assign the chromosome to appropriate groups and the information is displayed to the user, usually in the form of a karyogram in which the chromosomes are arranged in a tabular array of their classes (see Graham and Piper 1994). The chromosome classification performance of these systems depends on the type of material used, but at best the misclassification rate is 6–18% (Piper and Granum 1989) which compares poorly with visual classification by a cytotechnician (Lundsteen and Granum 1976). All automated systems in clinical use operate interactively, allowing an expert operator to correct machine errors in image segmentation, feature extraction and classification, resulting in useful performance (Graham and Piper 1994). However, there is clear scope for improvement in automatic classification. The objective of this study was to investigate the use of a neural network in the classification module.

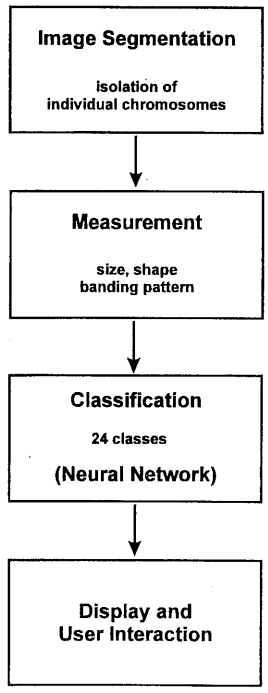


Figure G4.3.2. Block diagram of an automated chromosome analysis system. Classification of chromosomes follows segmentation and measurement modules, and is implemented in this study as a neural network. The display and interaction module permits correction of errors in machine analysis and diagnostic decision making.

© 1997 IOP Publishing Ltd and Oxford University Press

G4.3.2 Design process

G4.3.2.1 Design constraints

An important issue for automatic classification is the representation of the banding pattern. Several different classifiers have been reported using statistical or syntactic approaches (e.g. Granlund 1976, Granum 1982, Groen *et al* 1989, Thomason and Granum 1986). Each of these involves the extraction of a number of intuitively defined features, usually associated with the chromosome's density profile. The density profile is a one-dimensional pattern obtained by projecting the chromosome's density onto its center line (figure G4.3.1(*b*)), and reflects the largely linear organization of the chromosome structure. The processing involved in extracting features from the profiles involves the risk of losing information, a risk which may be eliminated by using the density profile itself as the banding representation. This type of one-dimensional pattern is a natural form of input for an artificial neural network. The potential advantage of neural network classifiers lies in their flexibility; they can be readily retrained for classification of new types of data. This property is likely to be useful for chromosome classification as specimen preparation techniques in routine use evolve very rapidly, resulting in changes in chromosome appearance. In particular, there is an increasing clinical requirement to use higher-resolution banding for diagnostic purposes, resulting in routine examination of longer (prometaphase) chromosomes. This will result in the need for greater adaptability in automated karyotyping systems.

Figure G4.3.2 indicates that the classification module is easily isolated from the rest of the system. The outputs of the classifier are the probabilities of membership of each of the 24 classes corresponding to the inputs for each chromosome. The inputs are the chromosome size, the centromeric index and the banding profile.

Size. This may be measured either as the length of the chromosome or its area; the two measures are very highly correlated. In the datasets used in this study, the length was used.

Centromeric index. The centromere divides the chromosome into long and short 'arms' (figure G4.3.1(*b*)). The centromeric index (CI) is the ratio of the length of the short arm to that of the whole chromosome, and gives a measure of shape.

Banding profile. The number of samples representing the banding profile can vary between 10 and 140 depending on the class of the chromosome and the state of contraction of the cell in which it occurred. The classification module requires a consistent input vector and all banding patterns must therefore be represented by the same number of samples. Considerable experimentation (Jennings and Graham 1993, Errington and Graham 1993) gave the result that a constant number of samples could be used to represent the profile, irrespective of the original chromosome length, and that this number could be quite small (as low as 15 samples for all profiles) with very little loss of classification accuracy. The use of a uniform number of samples meant that the profiles of long chromosomes had to be subsampled by local averaging, and the short chromosomes oversampled by interpolation.

The principal requirement of the classifier module is classification accuracy. The overall system performance is closely dependent on presenting the clinical user with a classification of the chromosomes in a cell which requires minimal interactive correction. Statistical classifiers give (barely) acceptable performance and it would be desirable to improve on this using a neural network classifier, although similar performance would be acceptable in view of the potential benefits in adaptability.

G4.3.2.2 Network topology

In this application we have a classification problem using continuous-valued inputs, where the classes are well defined and expert classification of the training data is available. It is a clear case for a *multilayer perceptron* (MLP). A preliminary study (Jennings and Graham 1993) compared the suitability of the MLP topology with the *Kohonen self-organizing map*, and confirmed the expected result that significantly better classification was obtained using the supervised training regime of the MLP. Optimum network parameters (starting gain, momentum, number of hidden nodes) were determined empirically (Errington and Graham 1993). C1.2 C2.1.1

In principle, it is possible to classify chromosomes on the basis of the banding pattern alone. However, the size and centromeric index are extremely powerful classification features, and must be included for the most accurate results. These features might be used as inputs to the network in addition to the banding features as shown in figure G4.3.3(*a*). It is known, however, that size and centromeric index can classify

© 1997 IOP Publishing Ltd and Oxford University Press

chromosomes into seven groups in the absence of banding information (the 'Denver' classification, Denver Conference 1960). An alternative form of input was therefore investigated, in which these two features were processed by a preclassifier, also an MLP, and trained to produce outputs corresponding to the 'Denver' classes. The seven outputs of the preclassifier were then used along with the banding features as inputs to the main classifier (figure G4.3.3(*b*)). The main classifier consisted of a network with 15 input nodes for banding features, plus the nodes necessary for the size and centromeric index features, 100 hidden nodes and 24 output nodes (one for each class), as illustrated in figure G4.3.3. The classification results in the three sets of chromosome data (see below) are given in table G4.3.3. It is clear that preprocessing the centromeric index and size features gave a considerable advantage.

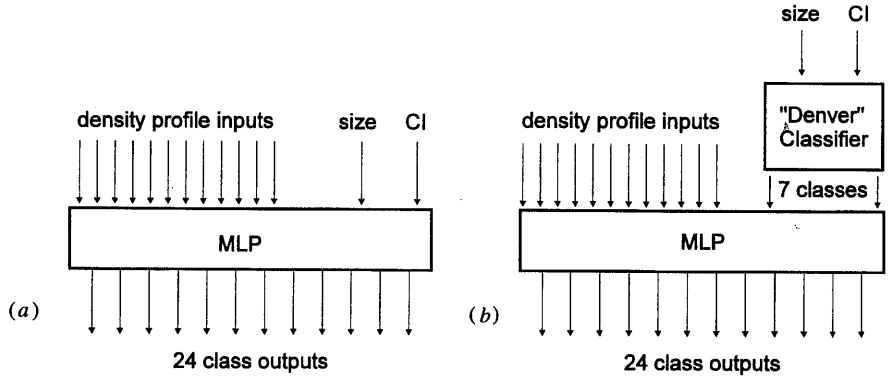


Figure G4.3.3. Two possible configurations for including size and centromeric index features in the input vector. (*a*) The two features are simply additional features along with the banding profile samples. (*b*) The features are processed to produce seven values corresponding to the probability of membership of the 'Denver' groups. The banding profile then provides information to refine the classification to 24 classes. In either case there are 24 outputs corresponding to the membership likelihoods of each of the classes.

G4.3.3 Training methods

The network was trained and tested using three data sets of annotated measurements from G-banded chromosomes. The characteristics of these data sets are summarized in table G4.3.1. The data in the Copenhagen set were obtained by densitometry of photographic negatives of selected cells of good appearance. The other two data sets were digitized directly from microscope images of routine material. The preparation techniques in chorionic villus sampling results in poor visual quality of the chromosome images in the Philadelphia set. The three data sets give a reasonably large number of data for network training and testing covering a range of quality representative of that found in a real implementation.

Table G4.3.1. Summary of the data sets of chromosome measurements.

Data set	Tissue of origin	Data acquisition method	Number of chromosomes	'Quality' of chromosome images
Copenhagen	Peripheral blood	Densitometry	8106	High
Edinburgh	Peripheral blood	TV camera	5469	Medium
Philadelphia	Chorionic villus	Linear CCD array	5817	Low

C1.2.3 The training algorithm employed was the classical *backpropagation* method (Rummelhart *et al* 1986), using a strategy of progressive reduction in gain (learning rate) during the training. Two measures were used to monitor performance: total network error and classification accuracy on the training data. These measures are not identical due to the fact that the classification result is determined only by the highest output, but they are both useful measures of performance. During training, the gain was halved if the total network error had increased by more than 10%, or the classification performance had not improved over

 © 1997 IOP Publishing Ltd and Oxford University Press

the previous presentation of the training data. Training was halted when the value of gain dropped below 10^{-4}. The gain reduction strategy proved extremely valuable in this application. Table G4.3.2 shows the misclassification rates on training data after convergence of networks trained to classify banding features alone in the preliminary study (Jennings and Graham 1993). There is a clear advantage in using gain reduction and in using two performance characteristics to monitor the network.

Table G4.3.2. The effect on classification performance of gain reduction during training, monitored using total network error and accuracy of classification of the training data.

Training strategy	No gain reduction	Gain reduction (network error only)	Gain reduction (network error and classification accuracy)
Misclassification rate	53 (%)	12 (%)	4 (%)

In the classification experiments, the network was trained using approximately half of each data set, the remainder being used for 'unseen' testing. The roles of the training and test sets were then reversed, and the classification rate obtained as the average of the two unseen tests. In all classification experiments the initial gain value used was 0.1 and the momentum value 0.7.

G4.3.4 Preprocessing

As noted above, the banding profiles were represented by 15 sample values, obtained by averaging or interpolation from the 'raw' profiles. The relative sizes and overall densities of chromosomes in a cell are fairly consistent; however, absolute lengths and densities can vary between cells. Length and density measures were therefore normalized to a constant value for each cell before classification.

The size and CI features were preprocessed using an MLP with two inputs, seven outputs and a hidden layer of 14 nodes (see figure G4.3.3(*b*)).

G4.3.5 Output interpretation

The network output is a vector of 24 class assignment values for each chromosome, approximating the Bayesian probabilities of the chromosome belonging to each class. The class to which the chromosome is assigned is that with the highest output. Classification results are shown in table G4.3.3. It is worth noting here that the classification of chromosomes is constrained by the fact that (in a normal cell) each class contains exactly two chromosomes (or one in the case of the sex chromosomes in a male cell). Application of this constraint can significantly improve the classification accuracy over 'context-free' classification of individual chromosomes (Tso *et al* 1991). Network approaches can give good results in applying constraints (Errington 1994), but consideration of these methods is beyond the scope of this chapter which is restricted to considering the classification of isolated chromosomes.

Table G4.3.3. Classification performance of two MLP configurations compared with that of a parametric statistical classifier (Piper and Granum 1989).

	Data set		
Classifier	Copenhagen	Edinburgh	Philadelphia
MLP, banding, length and centromeric index	6.9%	18.6%	24.6%
MLP, 'Denver' preclassifier	5.8%	17.0%	22.5%
Parametric classifier	6.5%	18.3%	22.8%
Significance of MLP improvement	2% level	5% level	not significant

© 1997 IOP Publishing Ltd and Oxford University Press

G4.3.6 Development

As we were required to carry out a number of experimental investigations using the network, and to arrive at a configuration which could be incorporated with other software modules, we implemented our own network simulators. They were programmed in Pascal and ran on UNIX workstations.

G4.3.7 Comparison with traditional methods

A feature of developing a neural network classifier for chromosome analysis is the possibility of comparing a network solution to classical statistical methods. There have been a number of approaches to chromosome classification, but the most successful prior to this study was that of Granum (1982), subsequently greatly refined by Piper (Piper and Granum 1989). This method extracts banding features using 'weighted density distributions'; essentially, the banding profile is multiplied by a number of intuitively defined weighting functions, approximating a set of basis functions for the banding pattern. The features extracted from the density profiles in this way are combined with length and CI features, and classified using a parametric classifier. Table G4.3.3 compares the best network performance with the statistical method of Piper and Granum (1989) in performing context-free classification of individual chromosomes. The network performance is significantly better for the Copenhagen and Edinburgh data sets and identical for the Philadelphia data set.

The results show that a network classifier can give higher classification accuracy than a classical technique. While the improvement is statistically significant, however, it is not overwhelming. The classification performance of both types of classifier is good for Copenhagen data, probably acceptable for data of routine quality, such as is found in the Edinburgh set, and inadequate in the case of the poor-quality Philadelphia data. The development costs of the network classifier are arguably appreciably smaller, since the time from proposing the concept to arriving at a final configuration was considerably shorter and involved less manpower than was the case for the conventional classifier. From an implementation point of view, the network classifier is likely to be more adaptable. Our experience is that the best network parameters (topology, gain, momentum) are stable in the face of wide variation in the quality of data. It seems likely then that a single 'hard-wired' network would be adequate for any implementation, requiring only a mechanical training process to adapt to the properties of the chromosome data in a new installation. Training 'on the fly' could be applied to account for slow changes in chromosome appearance arising from changes in the nature of the preparation techniques, etc.

G4.3.8 Conclusions

Chromosome classification is an important element in automated cytogenetic analysis. The classification problem in this case is far from trivial; there are few applications where there is a requirement to assign objects to as many as 24 classes. We have constructed a chromosome classifier using a multilayer perceptron network whose performance equals or betters that of a well developed classifier using traditional statistical methods. The form of the network is standard, with the exception that known properties of the classification features allowed the network to be 'factored' into two steps to achieve optimum classification performance. Equivalent performance can be obtained with a single network composed of many more nodes (Errington 1994).

In this study we have had the luxury, not afforded to many network implementations, that data sets have been available with fairly large quantities of expertly classified real-world examples. The data were made available within the Concerted Action of Automated Cytogenetics Groups supported by the European Community (project no II.1.1.13). An interesting feature of this application is that we have been able to make a direct comparison with a statistical classifier applied to the same data.

References

Denver Conference 1960 A proposed standard system of nomenclature of human mitotic chromosomes *Lancet* **1** 1063–5

Errington P A 1994 Application of neural network models to chromosome classification *PhD Thesis* University of Manchester

Errington P A and Graham J 1993 Application of artificial neural networks to chromosome classification *Cytometry* **14** 627–39

 © 1997 IOP Publishing Ltd and Oxford University Press

Graham J 1987 Automation of routine clinical chromosome analysis I, Karyotyping by machine *Anal. Quantit. Cyt. Hist.* **9** 383–90

Graham J and Piper J 1994 Automatic karyotype analysis *Chromosome Analysis Protocols* ed J R Gosden (Totowa, NJ: Humana) pp 141–85

Graham J and Pycock D 1987 Automation of routine clinical chromosome analysis II, Metaphase finding *Anal. Quantit. Cyt. Hist.* **9** 391–7

Granlund G H 1976 Identification of human chromosomes using integrated density profiles *IEEE Trans. Biomed. Eng.* **23** 183–92

Granum E 1982 Application of statistical and syntactical methods of analysis to classification of chromosome data *Pattern Recognition Theory and Application* ed J Kittler, K S Fu and L F Pau, NATO ASI (Dordrecht: Reidel) pp 373–98

Groen F C A, tenKate T K, Smeulders A W M and Young I T 1989 Human chromosome classification based on local band descriptors *Patt. Recog. Lett.* **9** 211–22

Jennings A M and Graham J 1993 A neural network approach to automatic chromosome classification *Phys. Med. Biol.* **38** 959–70

Lundsteen C and Granum E 1976 Visual classification of banded human chromosomes I, Karyotyping compared with classification of isolated chromosomes *Am. J. Human Genet.* **40** 87–97

Lundsteen C and Martin A O 1989 On the selection of systems for automated cytogenetic analysis *Am. J. Med. Genet.* **32** 72–80

Piper J and Granum E 1989 On fully automatic measurement for banded chromosome classification *Cytometry* **10** 242–55

Rummelhart D E, Hinton G E and Williams R J 1986 Learning internal representations by error propagation *Parallel Distributed Processing: Explorations in the Microstructures of Cognition* vol 1 *Foundations* ed D E Rummelhart and J L McCelland (Cambridge, MA: MIT Press) pp 318–62

Thomason M G and Granum E 1986 Dynamically programmed inference of Markov networks from finite sets of sample strings *IEEE Trans.* **8** 491–501

Tso M K S, Kleinschmidt P, Mitterreiter I and Graham J 1991 An efficient transportation algorithm for automatic chromosome karyotyping *Patt. Recog. Lett.* **12** 117–26

© 1997 IOP Publishing Ltd and Oxford University Press

18. T**rainable Grey-Level Models for Disentangling Overlapping Chromosomes.** G.C. Charters and J. Graham, *Pattern Recognition, 32: 1335-1349, 1999.* doi:10.1016/S0031-3203(98)00171-X

PERGAMON

Pattern Recognition 32 (1999) 1335–1349

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

Trainable grey-level models for disentangling overlapping chromosomes

Graham C. Charters, Jim Graham*

University of Manchester, Wolfson Image Analysis Unit, Department of Medical Biophysics, Stopford Building, Oxford Road, Manchester M13 9PT, UK

Received 11 September 1997; received in revised form 29 September 1998

Abstract

We propose and evaluate a mechanism for resolving the segmentation of overlapping chromosomes using trainable models of the expected banding appearance. The models consist of templates of sub-chromosome length band profiles. Candidate chromosome segments are classified according to their responses to the entire set of templates, and matched on the basis of the classifications. Evaluation of the models using a set of annotated banding profiles yields correct classification rates of 90.8% for isolated chromosomes, and 55.4% for chromosome fragments; 70.6% of overlapping chromosome pairs, simulated using the profile data set, are correctly resolved. © 1999 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.

Keywords: Chromosome analysis; Trainable models; Template matching; Overlapping chromosomes; Chromosome banding patterns; Classification; Segmentation

1. Introduction

Cytogenetics is the study of the genetic constitution of individuals at a scale which is revealed by light microscopy. At this level, the genetic material of a cell can be seen as a number of distinct bodies – the chromosomes. Analysis of the appearance of chromosomes can provide information on inherited or acquired syndromes, exposure to genotoxic agents or the presence of cancers. There are 46 chromosomes in normal human cells. Appropriately stained, they can be made to exhibit a sequence of bands which, together with the chromosome size and the position of a characteristic constriction called the centromere (see Fig. 2), can be used to assign chromosomes visually into one of 24 classes (labelled 1-22, X and Y). The last two classes are the sex chromosomes, of which females have two in class X and males one X and one Y. All other classes contain two identical (*homologous*) chromosomes in normal individuals. It is often helpful to display chromosomes arranged in a *karyotype* – a tabular array in which the chromosomes are aligned in pairs (see Fig. 1).

The appearance of chromosomes depends on the stage of the cell division cycle at which they are viewed. For much of the cell cycle (*interphase*), individual chromosomes cannot be distinguished. They only appear as distinct bodies towards the end of the cycle, at *prophase*, when they are long string-like objects, contracting and separating at *metaphase*, just before cell division takes place. Although not biologically significant, it is common to refer to an intermediate stage of contraction between

* Corresponding author. +44 161 275 5150; Fax: +44 161 275 5145; E-mail: Jim.graham@man.ac.uk

0031-3203/99/$20.00 © 1999 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.
PII: S0031-3203(98)00171-X

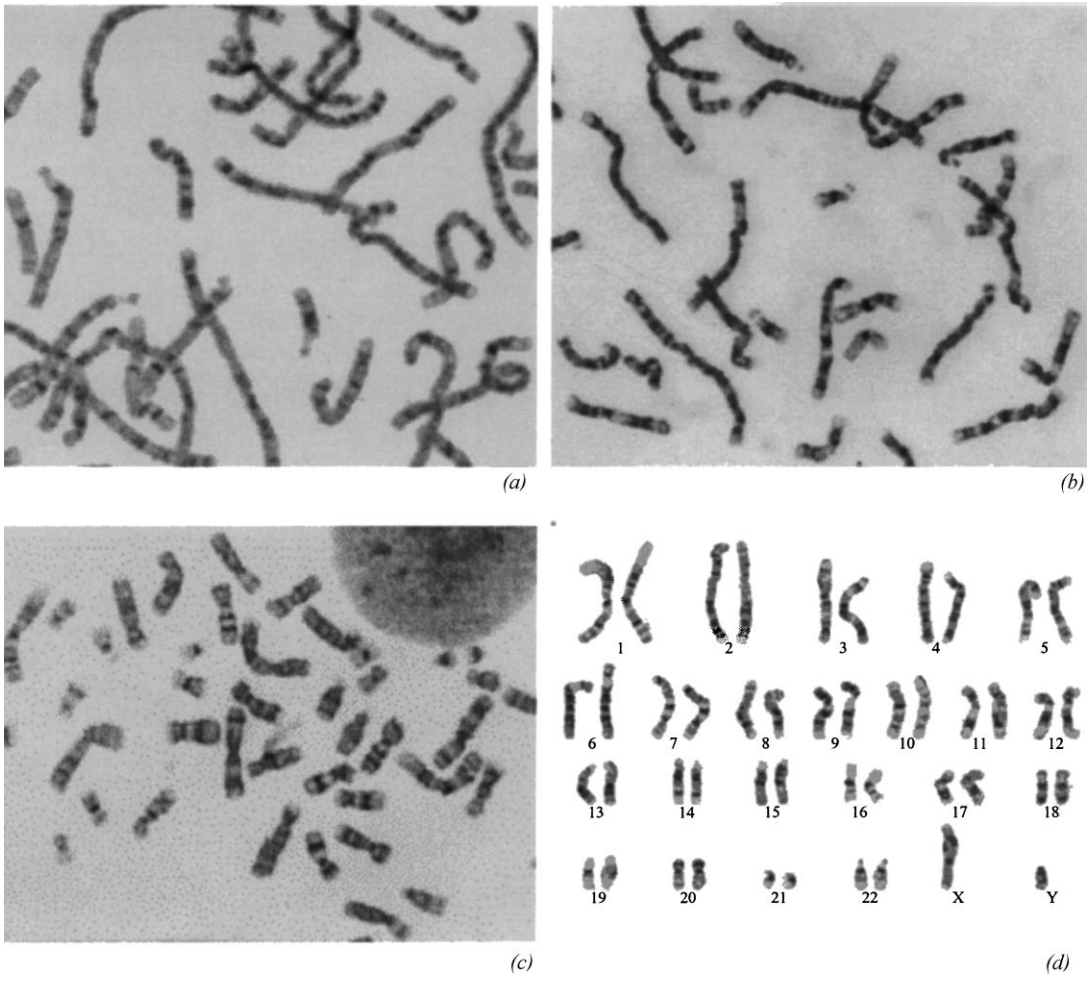


Fig. 1. Chromosomes at different stages of contraction. The images show parts of cells at (a) prophase. (b) prometaphase and (c) metaphase. The increasing contraction and decreasing resolution of the banding information on the chromosomes is clear. In the more elongated stages, the interpretation is complicated by the large number of overlaps. A karyotype of a prometaphase cell is shown in (d).

prophase and metaphase as *prometaphase*. Fig. 1 shows examples of prophase, prometaphase and metaphase chromosomes stained by a commonly used method (G-banding) together with a metaphase karyotype. Operationally, these stages in the division cycle are defined by the number of bands visible in the cell. The more elongated chromosomes exhibit more bands than shorter ones. Metaphase cells have around 450 bands and prometaphases are defined to have 550 bands and above. A cell is not considered to be in prophase unless at least 850 bands are visible. The greater number of bands at the more elongated stages provides a higher resolution description of the chromosome structure, which is advantageous for analysis. The analysis is, however, much more difficult due to the greater complexity of the banding patterns and the fact that longer chromosomes touch and overlap each other much more frequently than shorter ones (see Fig. 1).

The automation of chromosome analysis was first proposed in the 1960s [1]. Many years of effort have resulted in the development of commercial cytogenetics systems for analysis of banded chromosome preparations [2]. A great deal of attention in cytogenetics has recently been focused on molecular techniques and the application of image analysis to interpretation of fluorescence microscope images, including the location of specific hybridisation sites at interphase [3,4]. However, analysis of banded chromosomes remains of great importance particularly at the prometaphase or prophase stages of contraction. Most studies in automation, on the other hand, have concentrated on metaphase chromosomes, avoiding the segmentation difficulties arising from touches and

overlaps in the prophase and prometaphase cells. Graham and Piper [5] provide a review of methods used in automated chromosome analysis. Segmentation and classification of chromosomes into groups are important stages in the analysis and are generally taken to be separate tasks. Segmentation is usually performed by one or other of a number of thresholding methods [6–9]. Classification schemes use a representation of the banding pattern, generally derived from the integrated density profiles of the segmented object, together with size and centromere position [10–13] (see Fig. 2).

The straightforward segment-measure-classify strategy is inadequate for the analysis of images of chromosomes in their less contracted phases as it provides no mechanism for resolution of overlaps. Some attempts have been made to deal with clusters of touching (but not overlapping) chromosomes [6–9] where combinations of geometric and densitometric evidence have been used to resolve segmentation ambiguities.

Clearly, automatic separation of overlapping chromosomes is important for the analysis of prophase and prometaphase images, but has received relatively little attention compared to other aspects of the chromosome analysis problem, such as classification. Ji [7,14,15] has proposed methods for automatically segmenting both touching and overlapping clusters. His approach to segmentation of overlaps was to decompose a thresholded object into individual components using *geometric evidence*, i.e. by reasoning about shapes. Agam and Dinstein [16] have similarly applied reasoning about boundary curvature to separating touching or slightly overlapping chromosomes. In this paper we propose an alternative, additional source of information for disentangling overlaps, namely the banding pattern. We suggest a strategy for using the banding pattern, which uses the same chromosome evidence for reasoning about segmentation solutions and for classifying segmented chromosomes. The method consists of identifying consistent pairings of short sections of banding pattern to make believable chromosomes. This approach conforms to the approach of human experts, who often piece together chromosomes in overlaps by matching short banding sequences to a mental model of the chromosome classes.

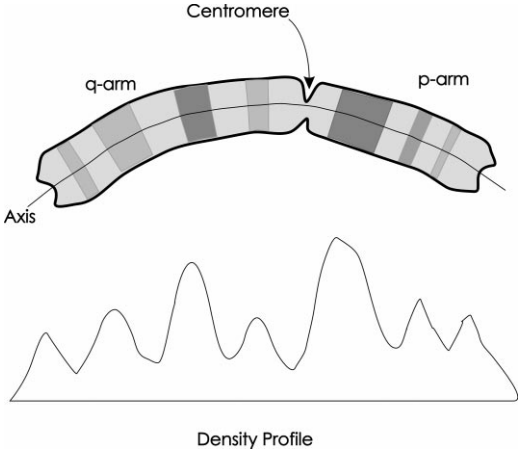


Fig. 2. Schematic representation of features used in chromosome classification. The bands are arranged linearly along the chromosome. Integrating density normal to the medial axis gives a density profile which is characteristic of the chromosome. The overall length of the chromosome and the position of the centromere are also important features. The centromere divides the chromosome into two "arms", conventionally labelled *p* and *q*.

2. Outline of the method

We summarise our method here using, as an illustration, the schematic pair of overlapping chromosomes in Fig. 3a. The banding pattern in the overlapping region is obscured, but four short sections of banding pattern are visible. The uncertainty in segmentation arises because each of these segments could be matched to any of the other three to generate complete chromosomes (Fig. 3c, d and e). If the classes of the partial chromosome segments can be identified by their local banding pattern, then the segments belonging to the same class can be matched, and the segmentation uncertainty resolved.

The classification of the segments is performed by matching the unobscured sections of banding pattern to a set of templates previously selected to match preferentially to particular sub-chromosome sequences of the banding pattern. These are called Partial Chromosome Models (PCMs). Fig. 3b shows PCMs which match preferentially to the end sections of the schematic chromosomes. The match is determined using a measure of fit described in Section 4.2, and used to classify the visible chromosome segments.

This approach takes inspiration from the work of Lockwood et al. [17–19], who observed that certain sections of the banding pattern are highly characteristic and used in visual classification of prophase chromosomes. They described a method of classifying prophase chromosomes on the basis of a set of short sub-chromosome-length banding segments. By conducting an exhaustive band-by-band search of digitised prophase ideograms (stereotyped banding patterns) they identified a set of 94 "unique band sequences" [19].

The PCMs we describe here are derived by training. Sub-chromosome-length segments of banding profiles are extracted from a training set of profiles in such a way as to give a large number of templates with a range of lengths (from half to three times the length of the shortest chromosomes), located at positions covering the banding profiles of all classes. For each of the templates, corresponding segments of banding profile are taken from the homologous chromosomes in all cells in the training set

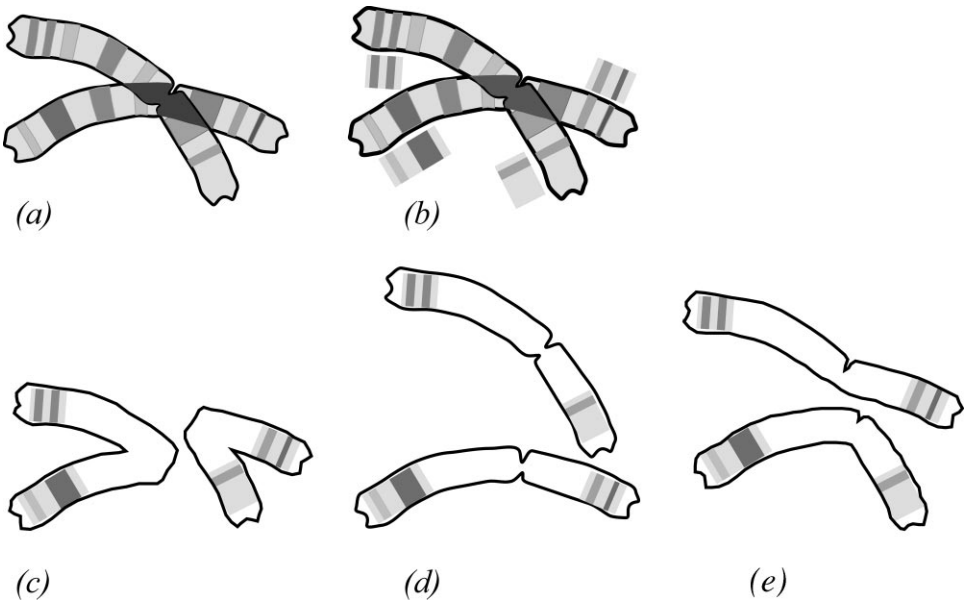


Fig. 3. Schematic illustration of the use of PCM templates in resolving a single overlap. (a) In the overlap region the banding pattern is obscured, but short lengths of banding pattern are visible which may be sufficient to identify the chromosomes. (b) Partial chromosome models (banding templates) are matched to each location on the overlapped object to classify the individual segments. The positions of best fit of four templates are shown. (c–e). There are three possible ways in which the segments can be paired to create whole chromosomes. The segment classifications obtained at (b) are used to resolve the uncertainty.

to provide a statistical description of the segment, which forms the PCM. Homologous chromosomes from different cells may have significantly different lengths, introducing a plastic deformation into the profile. This is accounted for by resampling each training profile to have the same number of samples, effectively length-normalising the segments (see Section 4.1 and Fig. 4). A collection of PCMs is created, derived from different segments of different classes, and each is matched at every position on the unobscured sections of banding pattern in an overlap to be resolved (the "target" segments). A measure of fit is calculated between each PCM and each target segment (the *response* of the target to the PCM).

We need a set of templates which is highly *discriminating*. That is, the collection of responses from the set of PCMs should provide accurate assignment of the target segment to the correct chromosome class. We assume that this requires the PCMs to be *specific* for their "own" chromosomes. That is to say the response of a PCM should be higher when matched to a target segment corresponding to the one from which it was trained than when matched elsewhere. Depending on how characteristic and consistent the banding pattern is at different locations, the specificity of the PCMs will vary. We select from the very large set of possible PCMs, a smaller "specific" set. If the specificity of each of the selected PCMs were very high, it would be sufficient to classify each target segment on the basis of the best matching PCM. It turns out that even the selected PCMs are not that specific (Fig. 6), and we classify the target segments on the basis of the *response vector* – the vector of maximum responses from each PCM.

The reason for classifying segments is to provide a mechanism for deciding which segments should be matched in resolving the overlap. The response vector is used as a feature vector in calculating probabilities of the segment being part of a chromosome from each of the classes. By treating the class assignments of each of the segments as independent events, we calculate the probability of assigning each *pair* of segments to the chromosome classes. The pairings that give the maximum *combined* probability for classification of *both* resulting chromosomes provide the resolution of the segmentation uncertainty and the final classification.

Assuming that the class probabilities assigned to each segment are independent, we apply Bayes' formula. The use of Bayes formula in contextual assignments is well covered in standard sources (e.g. [20]). Here we summarise the steps in arriving at a resolution of the overlap.

The probability, $P(u_{ik})$, of each segment, x_i, being a member of class k from the 24 classes is given by

$$P(u_{ik}) = P(g_k \mid x_i) = \frac{P(g_k)\,P(x_i \mid g_k)}{\sum_{j=1}^{24} P(g_j)\,P(x_i \mid g_j)} \quad k = 1, 2, \ldots 24. \tag{1}$$

$P(g_k)$ is the prior probability of obtaining class g_k (equal for all classes except the sex chromosomes). $P(x_i \mid g_k)$ is the probability of obtaining the features x_i, given the class g_k. This probability is given by application of a suitable classification function to the feature vector. We use a linear classifier (Section 5.4).

The probabilities that candidates should be matched to form a single (identified) chromosome are calculated according to

$$P(c_{ij}) = \max_k (P(u_{ik}, u_{jk})) = \max_k (P(u_{ik})P(u_{jk})),$$
$$k = 1, 2, \ldots, 24. \quad (2)$$

$P(c_{ij})$ is the maximum probability that segments i and j belong to a single chromosome from any class.

The resolution of the overlap is the configuration which gives the maximum combined probability for both chromosomes

$$P(s_{ijlm}) = P(c_{ij}, c_{lm}) = P(c_{ij})\,P(c_{lm}), \quad (3)$$

where $P(s_{ijlm})$ is the probability that segments i and j form a single chromosome and segments l and m form another.

3. Chromosome data and experimental approach

For our study we have used a digitised set of prometaphase chromosomes which have been used in a number of previous classification studies [21–24]. This data set (known as the 600-band data set) consists of 6177 chromosome banding profiles from 136 G-banded blood cells, fixed and stained at the 600-band stage. Each chromosome is represented by its density profile (Fig. 2) together with its centromere position and class specified by a trained cytogeneticist. Isolated images of each chromosome are also available, although not used in this study. The profiles in the data set carry with them an identifier for their cell of origin. In some circumstances (such as the simulation experiments of Section 6) it is useful to consider the chromosomes in the context of complete cells, as would occur in "live" analysis.

Approximately 34% of the chromosomes in the data set were touching neighbouring chromosomes in the original images. For the purposes of the earlier studies (for which the data were collected), cells were selected which contained very few overlaps; no information is retained concerning which of the chromosomes were involved in these.

In the following section, we describe some details of PCM representation, how the fit value is calculated and the method of assigning segments to classes. The method of selecting a "specific" set of PCMs from an initially large set of candidates is described in Section 5. Cross-validation experiments are used to evaluate the specificity of individual PCMs. In Section 6 we evaluate the selected PCMs for their ability to classify whole chromosomes and chromosome segments, and to resolve overlaps simulated from the 600-band data. Paradoxically, the "clean" nature of the data is an advantage for this study, as we can use simulation to generate a large set of overlaps with known correct resolution.

4. Trainable template matching

As we use the templates to generate probabilities of banding segments belonging to each chromosome class, the classes need to be defined in terms of observations made on the image features. While the banding patterns of chromosomes are characteristic of individual classes, there is considerable within-class variability. Furthermore, variations in the protocols for chromosome sample preparation lead to differences in the appearance of chromosomes imaged in different cytogenetic laboratories. It is important that templates represent the range of banding patterns that are likely to be observed in chromosome images, both in terms of an average banding pattern and the allowed variability. We capture the appearance and variability of the chromosomes by defining the PCMs with respect to a training set of banding profiles. PCMs are *trainable* templates. We require a representation of the banding sequence which can be determined by training and which includes a description of variability. We also need a method of assessing how closely the template fits to the density profile of a target chromosome segment. Having established a set of templates we require a method for using these templates for assigning the segments to chromosome classes.

4.1. Template representation

There are many possible ways of representing a banding profile, and several of these have been used in constructing chromosome classifiers [11,12]. Errington and Graham [13] have noted that the sequence of profile samples itself is as effective a representation as any. The banding profile illustrated in Fig. 2 is sampled at single pixel intervals along the chromosome axis. For classification purposes, the sampling interval can be much larger than a single pixel, and Errington and Graham used coarsely sampled profiles in their neural network classifier. This representation has also been used by Nivall [24] in classifying chromosome profiles, including the 600-band data set, and we adopt it for this study (see Fig. 4). The most appropriate density of profile sampling for template matching is a parameter to be determined empirically (Section 5.2).

The PCM template is a mean profile constructed from the corresponding profile segments in each cell in the training set, together with the covariance matrix. The different "raw" profiles in the training set corresponding

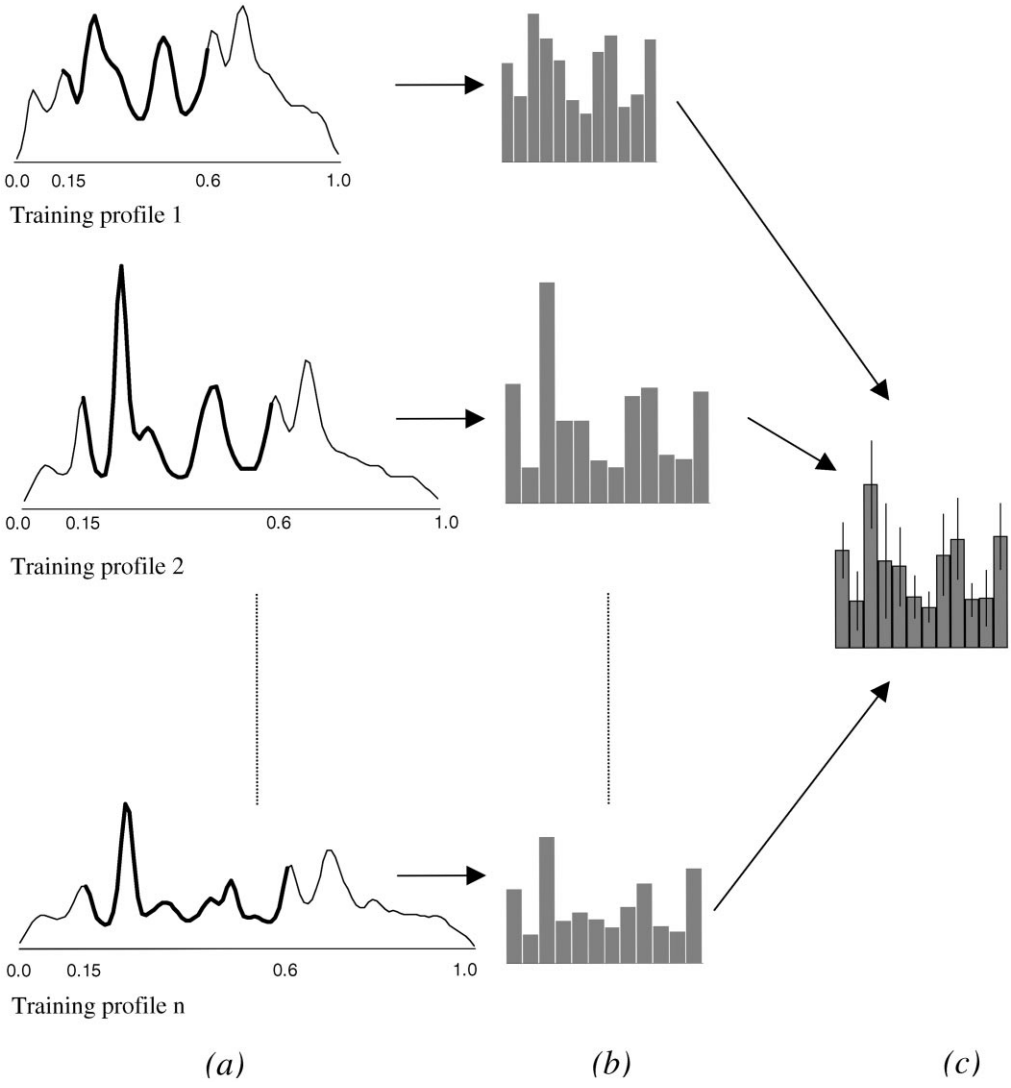


Fig. 4. The stages of training a PCM. (a) A training set of n homologous chromosome profiles is selected. The "raw" profiles are sampled at intervals of a single pixel along the chromosome axis (Fig. 2). Homologous chromosomes from different cells vary in the detailed form of the profile as well as in length, and hence in their number of samples. A particular PCM corresponds to a section of the profile, indicated by the heavier line. The start and end positions are specified in fractions of the chromosome length (in this case 0.15–0.6). (b) The appropriate sections are extracted from each training example and resampled more coarsely to give a fixed number of samples in each. (c) The variation in corresponding sample values is used to calculate a mean profile segment and covariance matrix. Templates of this kind are generated for a number of partial profile segments in each class, of different lengths and positions.

to the same chromosome class are represented by a different number of samples, due to differing degrees of contraction of different cells and small variations in segmentation parameters. These differences can be quite marked – up to 50% in profile length. If the samples are to be used as classification features, a constant length is required. In generating the coarsely sampled partial profiles, a fixed number of samples is used for each template, the start and end positions of the segments being expressed as a fraction of the chromosome length. This is illustrated in Fig. 4. A large set of PCMs is created in this way, each specified by differing start and end positions on chromosomes of all classes (see Section 5.1).

4.2. Matching method

As chromosome profiles are highly variable, there are several possible matching methods which might be

used. We describe elsewhere [25] the evaluation of four methods using the classification of whole chromosome profiles as the assessment criterion. The methods investigated were rigid template matching (cross-correlation), flexible template matching by dynamic programming and linear and quadratic classification. The details of our method of testing are not relevant to this paper, but the results showed template matching using a quadratic classification function to be clearly the method of choice. Here (and later in Section 5.4) we use linear and quadratic classification in the conventional sense, to mean classification functions which assume the features to have multivariate normal distributions with respectively pooled and unequal covariance matrices (see e.g. [20]). When used for template matching, the PCMs define the classes, and the target profile values provide the features to be classified. The value of the classification function provides the measure of fit. As an implementation detail, we noted that best matching results were obtained by normalising each template to a standard density and including the integrated density of the template as an additional element in the template vector.

4.3. Chromosome discrimination

As we shall see in Section 5.4, 182 "specific" PCMs are defined with a range of lengths and starting points distributed along the profiles corresponding to all chromosome classes. For chromosome segmentation and identification, we seek a response from each PCM to each segment of chromosome as illustrated in Fig. 3. The unobscured segments of chromosome are easily identified, and each PCM is matched to all available locations on each segment. The response for each PCM is the maximum value of the measure of fit generated over all locations on a segment. The responses for all PCMs form the *response vector* for the segment. Any appropriate classifier may be used for assigning the segments to chromosome classes using the response vector. Because of constraints on the size of our training set we use a linear classifier.

5. Identifying the set of partial chromosome models

There is a very large number of candidate sub-chromosome profile segments of different lengths at different locations, each of which could form a PCM. Not all of these will provide specific matches against target profiles. We wish to identify a set of templates that collectively gives best discrimination between chromosome fragments. We assume that this set will be contained within the set of templates that match with high specificity to their "own" chromosome segments. We determine this specific set empirically using cross-validation experiments. The data are split into two sets, A and B. PCMs trained using set A are used to identify chromosomes in set B and vice versa. In splitting the data set (here and in the evaluation experiments of Section 6), assignment to subsets is made on the basis of cells, i.e. chromosomes from the same cell are always placed in the same data set.

For each trained PCM, a successful identification in the evaluation set is counted if the correct chromosome appears in one of its top two scores after matching to all possible locations. (There are two potential correct fits to each template. In the case of chromosomes with only one example in a cell, such as the sex chromosomes, a success is scored only when the top fit is correct.) Each candidate PCM is evaluated according to its number of successes (or *recognition rate*).

5.1. Generation of candidate sequences

In principle, we need to generate all possible candidate templates for evaluation (all lengths of template with starting points all along the chromosome density profiles derived from each class). This is a dauntingly large task. The size of the task was reduced by limiting the range of template lengths tested and evaluating the matches at points separated by more than a single sample.

The experiments of Lockwood et al. [18] showed that, for prophase chromosomes, their unique band sequences ranged from slightly less than the length of the shortest chromosome class to approximately twice the length of the shortest class. Taking this result into account, we chose to test six sizes of sequences ranging from one half to three times the length of the shortest class.

Working along the profiles, we used a separation of three samples between candidate template positions, reducing the number of tests required by approximately one-third. It will become clear later that this was a reasonable separation due to the fact that adjacent sequences produced similar results (Fig. 6). This resulted in 1308 candidate sequences for evaluation.

5.2. Optimisation of sample density

Errington and Graham [13] have noted that, for classification, the optimum number of samples used to describe the profiles is rather less than the number in the "raw" profiles obtained by single pixel sampling along the axis. The use of more coarsely sampled profiles not only reduces the computation required, but also increases classification rates. We can therefore maximise the specificity of the PCMs by selecting the correct sampling density for the templates. We select the sampling density empirically by measuring the recognition rate (as defined above) at different sampling densities. Profiles are resampled from the "raw" profiles by local averaging. The "optimum sampling density" is the one that gives the highest recognition rate. Fig. 5 shows the results for classes 1 and 2. The graphs show the optimum number of

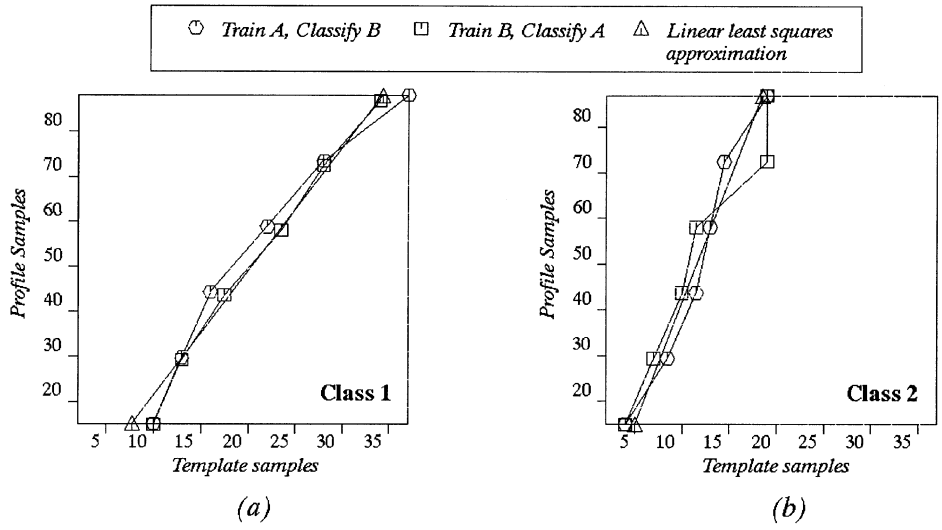


Fig. 5. Selecting the number of samples used to represent the template. Sample densities giving optimal performance for class 1 chromosomes (a) and class 2 chromosomes (b). Consistent results are obtained for training on the two halves of the data split, and the relationship between the optimum number of samples and the measured length of the profile is approximately linear.

samples corresponding to different lengths of template (measured in units of "raw" profile samples). The curves are approximately linear, showing that the fractional sampling rate is, to a first approximation, independent of PCM length. Results for each of the independent cross-validation experiments are shown to indicate that the results are in broad agreement. Similar results are obtained for all chromosomes [25], although the fraction by which the sampling density can be reduced is different for each class. The difference for classes 1 and 2 is clear from Fig. 5. This means that the sampling density appropriate for a PCM needs to be selected according to the class from which the template is derived. Calibration curves similar to Fig. 5 have been calculated for each class.

5.3. Candidate template evaluation

The 1308 candidate templates generated as described in Section 4.1 recognise their "own" banding sequences with different specificities. In this section we describe the process of selecting the set of the most *specific* PCMs. Fig. 6 shows the variation in specificity for PCMs of different lengths derived from different locations on chromosome 1. Specificity is assessed by the recognition rate. Results from the separate evaluation experiments are shown, indicating that the independent training sets are broadly comparable. The average density profile for class 1 chromosomes is shown in Fig. 6a. The rest of the figure shows the results of matching the six different lengths of PCM (b–g). In each of the graphs, the horizontal axis represents the location of the PCM along the profile. (The structure of the profile at relevant points can be determined from the density curve at (a).) The vertical axis of each curve is the recognition rate achieved for a PCM trained at that location. The short numbered bars indicate the positions and lengths of the PCMs finally selected, and give an indication of the length of PCMs evaluated at each level. The differences in the recognition rates of templates of different lengths generated from different locations are clear, and correspond to intuition. For example, the long lightly stained region at the right-hand end of Fig. 6a is a readily recognisable feature of chromosome 1, and the templates derived from this region achieve high recognition rates at all template lengths. Conversely, the region in the middle of chromosome 1 is not very characteristic and gives poor recognition rates at all lengths. Generally, PCMs of different lengths at the same location give similar responses, although sometimes a high level of specificity is obtained at one particular length which captures a locally characteristic banding appearance. An example of this is the section labelled 5 in Fig. 6c, where a locally high response occurs, in contrast to the responses of longer and shorter PCMs at that location. Notice that the recognition rate for any individual template is never very high (about 75% at best); it is their use in combination which gives specificity. Notice also that specificity varies slowly along the chromosome, justifying the strategy of generating templates centred on every third sample.

Similar specificity diagrams were generated for all classes, and templates were selected from the complete set of 1308 according to the following criteria.

1. The most specific templates of each class were chosen (as in Fig. 6).

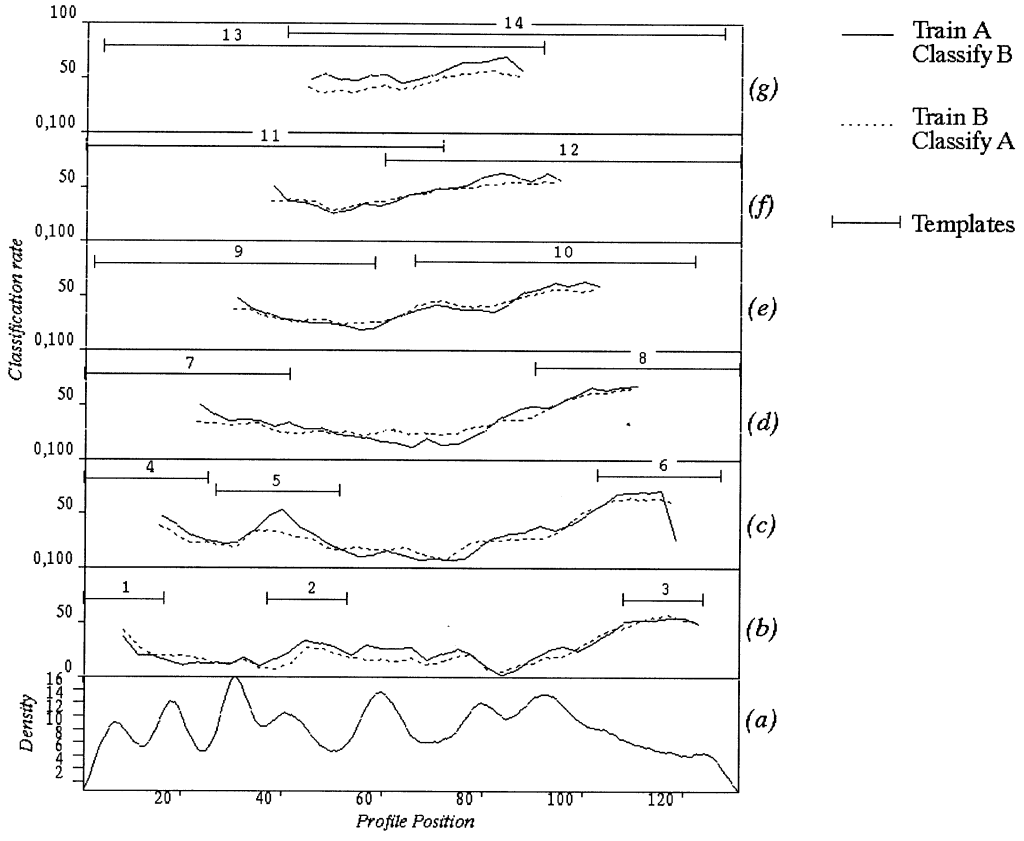


Fig. 6. Template matching performance on candidate PCMs for class 1 chromosomes. The mean density profile for chromosome 1 is shown at (a). The remainder of the figure shows the recognition rates for templates derived at each location along the profile. The recognition rates measure the specificity of each PCM in recognising its "own" banding sequence. Results for six different lengths of PCM templates are shown, going from shorter templates (b) to longer ones (c–g). The dotted and solid curves represent different training/testing splits of the data, and show consistent performance. Each numbered horizontal bar corresponds to a sequence template selected for further investigation.

2. Where several templates of the same length shared substantial sections of profile, only the most specific of them was used. Short templates overlapping with longer ones were retained, even though they may be less specific, on the grounds that they may be useful in resolving overlaps where sections of longer templates might be obscured.

Application of these selection criteria reduced the set to 182 templates, of which the fourteen selected for chromosome 1 are shown in Fig. 6.

5.4. Using PCMs for segment classification

This set of PCMs, selected for their individual specificity in matching to their "own" chromosomes, is used to generate a feature vector for classifying target segments. In this way, the response of each PCM to a target contributes to the segment's classification. We attempted to reduce the dimension of this feature vector using Forward and Backward stepwise selection [26,27]. However, removal of features consistently resulted in reduced classification performance, so the full set of 182 PCMs was retained.

Any suitable classifier might be used for classifying segments. The dimension of the feature vector is quite large (182). Although the data set is substantial, there is also a large number of classes, so that the quantity of training data for each class is limited. As a consequence we use a linear classifier. We refer to this classifier as the *linear PCM classifier*.

6. Evaluation of PCMs for resolving overlaps

In this section we evaluate the performance of the linear PCM classifier in classifying whole chromosomes,

classifying chromosome segments and resolving overlaps.

6.1. Cross-validation strategy

We wish to reduce bias in the estimation of our classifier performance by cross-validation. This is often achieved by splitting an annotated data set into two, using each half in turn to act as a training set for classifying the other, as in Section 5. In our evaluation experiments we require to train two sets of models: the PCM templates themselves, and the linear PCM classifier. The PCM templates used are those selected as described in Section 5.3 They are trained by gathering statistics from the chromosome fragments as described in Section 4.1. The linear PCM classifier based on these templates is itself trained using the response vector of identified profile fragments. To reduce bias, these different models should be trained on separate data, with yet further data being used for evaluation. We therefore split the data into thirds, using each subset of the data in turn for (i) training templates (the PCM training set), (ii) training the linear classifier (the classifier training set) and (iii) evaluation, requiring six complete classification experiments to use all the data for validation.

6.2. Simulation of overlaps

The chromosomes in the 600-band data set have been selected so that occlusion by overlapping was kept to a minimum. We use this set of clean profiles to conduct experiments on resolving overlaps by simulating the density profiles from overlapping chromosomes. The advantage of this approach over the use of genuine overlaps is that any number of overlapping configurations may be created, each with a known true resolution, for both training and evaluation. The disadvantage is that the appearance of the density profile at the overlap may not be totally realistic. This is not particularly problematic as overlapping regions in real chromosome images can be identified by a number of straightforward criteria [14].

We simulate overlaps by obscuring short sections of profile at randomly selected positions on pairs of chromosomes. The number of overlaps in any cell was selected at random from a range determined from the observed numbers in prophase images (about 26% of chromosomes contain at least one overlap). The positions of obscured sections of profile were selected randomly along the lengths of the profiles; the widths of the obscured sections were generated from the observed distributions of chromosome widths. Profile densities in the obscured region were set to a value darker than the normal maximum density. We use the information on the cell of origin of the chromosomes to simulate the analysis of a complete cell at a time.

6.3. Experiments

We performed the three following experiments using simulated overlaps. In each case the PCM templates were trained as described in Section 4.1 using the PCM training sets. The experiments differ in the evaluation sets used and the corresponding classifier training sets.

Among the conditions to be varied in the experiments, the training data and the evaluation data may consist of "clean" profiles (containing no overlaps) or "representatively overlapped" profiles (simulated overlaps occurring in about 26% of chromosomes). In the latter case, to obtain sufficient numbers of overlaps for training and evaluation, the overlap simulation procedure was applied to the data in three passes, generating around 11 000 segments. To evaluate the effect of training set size, a set of about twice that number was generated from six passes. The larger numbers of overlaps were generated by multiple passes, rather than a single pass with a higher overlap rate, so that the distribution of sizes of unoverlapped segments would remain representative. We will refer to the three-pass or six-pass training or evaluation sets.

Experiment 1 (*Classification of Whole Chromosomes*). In this experiment we sought to obtain a measure of classification performance assuming all overlaps have been correctly resolved. Three different evaluations were carried out using different regimes of profile simulation.

(i) Evaluation of "clean" profiles: no simulated overlaps introduced.
(ii) Evaluation of "representative overlaps": each cell contained a number of overlaps as described in Section 6.2.
(iii) Evaluation on "wholly overlapped" profiles: isolated chromosomes which were not involved in simulated overlaps were excluded from the evaluation.

To obtain sufficient evaluation examples, experiments (ii) and (iii) were conducted using a three-pass evaluation set.

Linear PCM classifiers were trained on each of the three types of simulated data, and each classifier used in turn to classify evaluation data of each type: nine classification experiments in all.

Experiment 2 (*Classification of Chromosome Segments*). In this experiment we tested the PCM classification performance when applied to the classification of chromosome segments (uncorrupted sub-chromosome length sections of profile extracted from overlapping chromosomes in the three-pass evaluation set). From this we obtained a measure of the ability to classify chromosome fragments, distinct from the results on overlap resolution (below).

Lengths of segments varied from under 10% of chromosome length to complete chromosomes. (The shortest segments were 15 profile samples long, corresponding to the shortest templates generated.) Classification was tested using linear PCM classifiers trained in three different ways to compare different training regimes.

(i) Trained on whole chromosomes.
(ii) Trained on segments from the three-pass training set.
(iii) Trained on segments from the six-pass training set.

Experiment 3 (*Resolution of Overlaps*). One hundred and thirty-six-overlapping pairs of chromosomes (one pair from each cell in the data set) were simulated. Each overlapping pair consisted of four segments. We performed overlap resolution experiments using linear PCM classifiers, trained using three-pass and six-pass training sets respectively. Overlap resolution was conducted as described in Section 2 (Eqs. (1)–(3)).

6.4. Results

Experiment 1. Table 1 shows the results for classification of whole chromosomes using PCMs. Each column corresponds to one of the three forms of data used to train the classifier, and each row corresponds to the data classified.

Experiment 2. Table 2 shows the results for the classification of chromosome segments according to the training data used to generate the classifier. Training on chromosome segments is clearly superior to training on whole chromosomes, and some advantage is gained from the larger training set.

Experiment 3. Table 3 shows the results of using the PCM templates to resolve overlapping pairs of chromosomes. The rows correspond respectively to the smaller and larger training set for the linear PCM classifier. Each row shows the percentage of overlaps correctly resolved and the percentage of the correctly resolved overlaps which were correctly classified.

Table 2
Classification of chromosome segments. Correct classification rates of isolated chromosome segments using the linear PCM classifer trained on whole chromosomes (without simulated overlaps) and on fragments extracted from the training set. Training on fragments is clearly superior, and the larger number of fragments in the six-pass training set gives advantage (see text)

Classifer training	Percentage correct classifications
Whole (clean) chromosomes	32.8
Chromosome segments (3 passes)	51.7
Chromosome segments (6 passes)	55.4

Table 3
Results for resolving overlapping chromosome pairs. Correct overlap resolution rates for linear PCM classifers trained on segments derived from three passes and six passes of segment generation from the training set. The rightmost column shows the number of correctly identified chromosomes which were also correctly classified

Segment training set	Correctly resolved overlaps	Correctly classified chromosomes
Three-pass training set	66.5%	79.6%
Six-pass training set	70.6%	82.6%

7. Discussion and conclusions

Experiment 1 confirms the expected result that classifying "clean" chromosomes gives better results than classifying chromosomes with overlaps. The absolute classification rate for complete clean chromosomes is encouraging, and compares well with previously published classification methods. Table 4 shows a comparison with the results of previous classification studies using the 600-band data. The first column shows the

Table 1
Whole chromosome classification using linear PCM classifiers. Correct classification rats are shown for training and evaluation on clean, representative and wholly overlapped data (see text). Best classification is obtained when the appropriate data are used for classifer training

		Training data		
		Clean	Representative	Overlapped
Classification data	Clean	90.8%	88.8%	74.8%
	Representative	77.5%	83.7%	71.5%
	Overlapped	52.6%	64.4%	67.7%

Table 4
Performance of template matching by quadratic discrimination. This table compares the template matching scheme used here as a "whole chromosome" classifier with the best previously published classification rates for the "600-band" data set. Column 1 shows the percentage of correct classifications obtained using banding profile and density features only in a quadratic classifier. In column 2 the method has been applied including normalised size and centromere position as additional features. Column 3 shows the rate achieved by Kleinschmidt et al. [23] using a maximum likelihood classifier on a different representation of the profile together with size and centromere position. Column 4 shows the rate achieved by Nivall [24] using a similar representation to that used here

Template matching by quadratic classifer (banding data only)	Template matching by quadratic classifier (including length and centromere position)	Kleinschmidt et al.	Nivall
90.2%	92.4%	91.3%	91.6%

result of using trainable templates to classify clean chromosomes with whole (rather than partial) chromosome templates and a quadratic (i.e. multivariate gaussian), instead of a linear, classifier. The classification performance is almost identical with that shown in Table 1, indicating that the PCMs provide as complete a description of the banding pattern as the full profiles. If anything, they do slightly better. The remainder of Table 4 sets this performance in context by comparing the template matching classifier of column 1 with the results of Kleinschmidt et al. [23] and Nivall [24], who have previously achieved the best classification performance on the 600-band data. Both of these studies used, in addition to the banding pattern, the powerful features of chromosome length and centromere position (Fig. 2), which are not available to the PCM classifier. The second column of Table 4 shows the result of the template matching classifier of column 1 when these additional features are used. Template matching achieves an improvement in classification over both Kleinschmidt and Nivall. Although these improvements are significant (at the 1.4% and 5% levels respectively), they are small. The object of the comparison is not to achieve a better classifier, but to demonstrate that the form of the template and the matching method are capable of creditable results in recognising banding patterns on prometaphase chromosomes, and that PCMs adequately represent the banding information.

Tables 1 and 2 show that the performance of the linear PCM classifier is best when trained with data of the same type as is being classified. It is not surprising that the classification of "clean" chromosomes is best done using "clean" templates, but it is more difficult to see why the converse should be true. At the moment, we have no explanation for this observation. However, using the figures of Table 1, we propose that if we have no idea whether a chromosome to be classified is isolated or overlapped then the best results for classifying representative overlapped data is 83.7%. In analysing a chromosome image, it is usually possible to know which chromosomes are isolated and which are involved in overlaps. In which case an appropriately trained classifier could be used for each chromosome. Given our observation that 26% of chromosomes are typically involved in an overlap, then the best result we can obtain is approximately 84.8% $(0.74 \times 90.8 + 0.26 \times 67.7)$. We could improve this rate to 86.0% if we were to use the quadratic discriminant template classifier to classify the clean chromosomes (Table 4).

The classification rates presented here will be to some extent underestimated. The assumption that the chromosomes in the 600-band data set contain no overlaps is not entirely true. There is a small (unknown) level of residual overlap in the data, providing an element of noise in the measurements, which will result in a slight depression of the classification performance. This does not affect our conclusions, as the same assumption has been made in all studies making use of this data set.

Experiment 2 indicates that chromosome fragments can also be classified fairly well. The fragments range in length from about 10% to about 90% of the chromosome length. It is unsurprising that the correct classification rate is much lower than for isolated chromosomes.

Experiment 3 shows that about 70% of simulated overlaps can be correctly resolved *using banding information alone*. Of those correctly resolved 82.6% are also correctly classified. For overlap resolution we can tolerate the modest classification performance for segments observed in Experiment 2, and it is possible to resolve the overlap without correctly identifying each of the chromosomes. If two chromosomes overlap, it is sufficient to have good evidence for identifying one of them, provided there is no strong evidence for an alternative erroneous interpretation.

The approach adopted in this study expands on the "unique band sequences" of Lockwood et al. [17–19]. They identified a set of 94 such sequences [18], which were used as templates to be matched to candidate chromosomes. We have extended this idea in two important ways to provide the banding evidence for chromosome segmentation and classification. Firstly, we incorporate knowledge of *profile variability* into the choice and use of

sequences, by the use of trainable models. Secondly, our sequence selection is determined by measured specificity of the templates, rather than by an intuitive assessment of "uniqueness". Lockwood et al. did not evaluate their approach fully for classification, concentrating on evaluating matching specificity. Using cross-correlation to match templates to about 20% of the possible profile positions on 850-band chromosomes they correctly identified 88% of template sequences [18]. Our results demonstrate that this approach, applied to somewhat more condensed chromosome material, can achieve results comparable to classification on whole chromosome profiles.

The principal motivation for this study was the resolution of overlapping chromosomes, which we have considered as quite separate from the case of clusters of touching chromosomes. Previous work [14,16] has demonstrated that geometric evidence concerning local boundary shape can lead to fairly accurate extraction of individual chromosomes from touching or slightly overlapping configurations. In principle, banding information could also be brought to bear to resolve touching clusters, but in that case, the classification of complete hypothesised chromosomes could be used. We have sought here to concentrate on the case of total overlaps, where complete banding information is not available. The most successful previous study with this aim is that of Ji [14], who used purely geometric reasoning. Geometric cues are often powerful for resolving overlaps (see Fig. 1), and Ji achieves a correct resolution rate of 94.6% in resolving 46 overlaps. He subsequently showed how his overlap resolution method can be combined with splitting touching chromosomes for successful automatic segmentation of unbanded chromosomes [15] (stained to be uniformly dark – used in counting chromosome aberrations for environmental monitoring).

On the face of it, Ji's approach gives significantly better overlap resolution. However two points can be made. Firstly, Ji's study [14] illustrates a difficulty in carrying out this type of investigation: his methods were tested on relatively small numbers of overlaps. This arises from the difficulty of identifying a sufficient number of configurations of chromosomes which are suitable for the analysis, and for which "correct" solutions are known for both training and evaluation. The approach taken here overcomes this difficulty by simulating the appearance of overlaps from "clean" data. We can therefore generate as many partial chromosomes, with known classifications, as we wish. There is a potential criticism of such an approach, in that it requires the appearance of the simulated overlaps to be realistic. Since we use only the banding profile information away from the obscured regions, we feel safe that nothing of significance is lost in pretending that these sections came from genuinely overlapping chromosomes.

Secondly, the method we describe here uses the banding pattern as the only source of evidence for segmentation. Rather than being seen as an alternative, the banding pattern provides additional evidence which can to be combined with geometry to provide a more informed basis for the assessment of segmentation hypotheses. We have demonstrated that this source of evidence alone can provide a useful contribution to resolving the segmentation uncertainty. The issue of *trainability* is important here: using trainable models means that the methods are not tied specifically to the properties of a given type of material, nor are they critically dependent on the setting of arbitrary heuristic parameters. Furthermore, basing features on a training set results in measures of compatibility between segments which approximate to true probabilities and which could, in principle, be used in combination with geometric cues to improve the performance of both approaches. Problems of relative scaling between disparate sources of evidence can be overcome using Bayesian methods if all evidence is expressed as probabilities. We have investigated [25] how PCMs may be combined with a trainable geometric method, using the images of isolated chromosomes available with the 600-band data, and will describe this in a later publication.

The disadvantage of using trainable models is in the necessity for a large training set. The 600-band data set used in this study was adequate in terms of numbers of samples and length of chromosomes to demonstrate feasibility. The methods would be used to best effect in the segmentation and classification of prophase (850 band) chromosomes. A set of 850-band data is currently being assembled [24], but the task is labour-intensive and time consuming and an insufficient number of chromosomes has so far been collected to allow the study described here to be repeated at this stage.

8. Summary

Disentangling overlaps is an important task in the analysis of images containing chromosomes at the prophase or prometaphase stages of contraction. Previous investigations into segmenting overlapping chromosomes have relied on reasoning about chromosome shapes to resolve the ambiguities in interpretation. Information contained in the chromosome banding patterns can also be used for this purpose. We propose and evaluate a mechanism of using the banding information based on trainable grey level models. The models, referred to as *Partial Chromosome Models*, consist of a set of templates corresponding to banding sequences of subchromosome length, selected so that they provide good discrimination between chromosome classes. Candidate profiles are matched to templates using a quadratic classification function. Chromosome segments are assigned to chromosome classes on the basis of their responses to the entire set of templates, by using the matching scores

as features in a linear classifier. The classifications of the segments are then used to propose matches to identify complete chromosomes within a composite object. We evaluate the method using a set of chromosome banding profiles derived from prometaphase chromosomes, whose classes have been expertly identified. The form of the model and the matching method are shown to be capable of high specificity, achieving correct classification results on whole chromosomes, using whole chromosome models, of 92.4% which improves on previously published classification results on this set of data. Using Partial Chromosome Models, a correct classification rate of 90.8% is obtained for isolated whole chromosomes and 55.4% for chromosome fragments, some of which represent less than 10% of the chromosome length. We test the ability of the models to resolve overlaps by simulating overlapping pairs of chromosomes using the profile data set. Despite the rather low rate of correct classification for chromosome fragments, 70.6% of simulated overlaps are correctly resolved. We discuss the possibility of combining the use of grey-level cues with geometric cues for untangling overlapping chromosomes.

Acknowledgements

The data used in this study were made available as part of the Concerted Action of Automated Cytogenetics Groups, Project No. II.1.1/13 and the Concerted Action on Automated Molecular Cytogenetic Analysis, Project No. BMH1-CT92-1307, supported by the European Community. We are grateful for the co-operation of colleagues within these projects.

References

[1] R.S. Ledley, High speed automatic analysis of biomedical pictures, Science 146 (1964) 216–223.
[2] C. Lundsteen, A.O. Martin, On the selection of systems for automated cytogenetic analysis, Amer. J. Med. Genetics 32 (1989) 72–80.
[3] J. Piper et al., Computer image-analysis of comparative genomic hybridization, Cytometry 19 (1995) 10–26.
[4] H. Netten, L. van Vliet, H. Vrolijk, W.C.R. Sloos, H.J. Tanke, I. T. Young, Fluorescent dot counting in interphase nuclei, Bioimaging 4 (1996) 93–106.
[5] J. Graham, J. Piper, Automatic karyotype analysis, in: J.R. Gosden (Ed.), Chromosome Analysis Protocols, vol. 29, Humana Press, Totowa, NJ, 1994, pp. 141–185.
[6] J. Graham, Resolution of composites in interactive karyotyping, in: C. Lundsteen, J. Piper (Eds.), Automation of Cytogenetics, Springer, Berlin, 1989, pp. 191–203.
[7] L. Ji, Intelligent splitting in the chromosome domain, Pattern Recognition 22 (1989) 519–532.
[8] A.M. Vossepoel, Separation of touching chromosomes, in: C. Lundsteen, J. Piper (Eds.), Automation of Cytogenetics, Springer, Berlin, 1989, pp. 205–216.
[9] Q. Wu, J. Snellings, L. Amory, P. Suetens, A. Oosterlink, Model-based contour analysis in a chromosome segmentation system, in: C. Lundsteen, J. Piper (Eds.), Automation of Cytogenetics, Springer, 1989, pp. 205–216.
[10] R.S. Ledley, P.S. Ing, H.A. Lubs, Human chromosome classification using discriminant analysis and Bayesian probability, Comput. Biol. Med. 10 (1980) 209–219.
[11] F.C.A. Groen, T.K. ten Kate, A.W.M. Smeulders, I.T. Young, Human chromosome classification based on local band descriptors, Pattern Recognition Lett. 9 (1989) 211–222.
[12] J. Piper, E. Granum, On fully automated feature measurement for banded chromosome classification, Cytometry 10 (1989) 242–255.
[13] P.A. Errington, J. Graham, Application of artificial neural networks to chromosome classification, Cytometry 14 (1993) 627–639.
[14] L. Ji, Decomposition of overlapping chromosomes, in: C. Lundsteen, J. Piper (Eds.), Automation of Cytogenetics, Springer, Berlin, 1989, pp. 177–190.
[15] L. Ji, Fully automatic chromosome segmentation, Cytometry 17 (1994) 196–208.
[16] G. Agam, I. Dinstein, Geometric separation of partially overlapping nonrigid objects applied to automatic chromosome classification, IEEE Trans. Pattern Anal. Mach. Intell. 19 (1997) 1212–1222.
[17] D.H. Lockwood, V.M. Riccardi, D. Johnston, Unique Band Sequences (UBS) of the ISCN prophase chromosome ideogram: a starting place for accurate computer-assisted chromosome analysis (CACA), Amer. J. Hum. Genetics 35 (1983) p. A140.
[18] D.H. Lockwood, V.M. Riccardi, S.O. Zimmerman, D.A. Johnston, Prophase chromosome Unique Band Sequences: definition and utilisation, Cytogenetics Cell Genetics 42 (1986) 141–153.
[19] D.H. Lockwood, D.A. Johnston, V.M. Riccardi, S.O. Zimmerman, The Use of subchromosome-length Unique Band Sequences in the analysis of prophase chromosomes, Amer. J. Hum. Genetics 43 (1998) 934–947.
[20] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1973.
[21] P.A. Errington, Ph.D. Thesis, University of Manchester, 1994.
[22] J. Piper, Stein's paradox and improved quadratic discrimination of real and simulated data by covariance weighting, 12th Int. Conf. on Pattern Recognition, Jerusalem, Israel, IEEE Computer Society Press, Silver spring, MD, 1994, pp. B529–B532.
[23] P. Kleinschmidt, I. Mitterreiter, J. Piper, Improved Chromosome Classification using Monotonic Functions of Mahalanobis Distance and the Transportation Method, ZOR – Math. Methods Oper. Res. 40 (1994) 305–323.
[24] S. Nivall, Ph.D. Thesis, Chalmers University of Technology, 1995.
[25] G.C. Charters, Ph.D. Thesis, University of Manchester, 1994.
[26] W.R. Klecka, Discriminant Analysis, Sage Publications, Beverley Hills, CA, 1984.
[27] R.J. McKay, N.A. Campbell, Variable selection techniques in discriminant analysis I. description, Br. J. Math. Stat. Psychol. 35 (1982) 1–29.

About the Author—GRAHAM C. CHARTERS received his B.Sc. in Computer Science in 1990, his M.Sc. in Numerical Analysis and Computing in 1991 and his Ph.D. in Machine Vision in 1995, all from the University of Manchester, England. He is currently with the IBM U.K. Scientific Centre in Winchester, England, where his research interests are Medical Imaging and Pattern Recognition.

About the Author—JIM GRAHAM received the B.Sc. degree in Physics from the University of Edinburgh, Scotland in 1974 and his Ph.D. in Structural Biology from the University of Cambridge, England in 1978. He joined the Wolfson Image Analysis Unit at the University of Manchester in 1978, where he is currently Senior Lecturer in Medical Biophysics and Honorary Lecturer in Computer Science. His research interests are biomedical and industrial applications of Machine Vision and Pattern Recognition.

19. **Disentangling Chromosome Overlaps by Combining Trainable Shape Models with Classification Evidence.** G.C. Charters and J. Graham. *IEEE Trans. Signal Processing 50: 2080-2085, 2002.* doi: 10.1109/TSP.2002.800421

Correspondence

Disentangling Chromosome Overlaps by Combining Trainable Shape Models With Classification Evidence

Graham C. Charters and Jim Graham

***Abstract*—Resolving chromosome overlaps is an unsolved problem in automated chromosome analysis. We propose a method that combines evidence from classification and shape, based on trainable shape models. In evaluation using synthesized overlaps, certain cases are resolvable using shape evidence alone, but where this is misleading, classification evidence improves performance.**

***Index Terms*—Biological cells, evidence combination, image segmentation, occlusion, shape modeling.**

I. Introduction

The automation of chromosome analysis, involving segmentation of chromosomes and classification into 24 groups, was one of the earliest applications of pattern recognition research. Many years of effort have resulted in the development of commercial cytogenetics systems for automated analysis of banded chromosome preparations (see Fig. 1). Analysis is complicated by the occurrence of clusters of unseparated chromosomes. Mostly, these consist of "touches," which are chromosomes that do not overlap but lie so close together that thresholding has failed to separate them (see A of Fig. 1). Commercial systems usually include methods that will separate a proportion of touches. The separation of true overlaps, such as those marked B in Fig. 1, invariably requires operator interaction.

The problem of extracting individual objects from overlapping configurations is one that occurs in a number of applications in machine vision. It is a difficult problem in general and can be thought of in two parts: 1) identifying overlapping objects in the first place and 2) identifying the individual components. In the case of chromosomes, it is usually fairly easy to identify an overlap on straightforward shape criteria; we discuss this further in Section VI. In this paper, we concentrate on the issue of identifying the individual objects in the cluster. The approach adopted here falls under the generic heading of "hypothesize and test." Candidate objects are hypothesized based on some data-driven process and the hypothesized objects evaluated against a model of their expected appearance. The model of the objects in this case is that they are long and flexible with a recognizable banding pattern. We propose a method for matching hypothesized candidates against that model.

In a previous publication [1], we have described a method for resolving overlaps, which makes use of classification evidence from the chromosomes' banding pattern without taking into account the shape of the hypothesized object. In this paper, we describe a method for modeling shape, which can provide geometric evidence for resolving overlaps. The models are trainable and readily provide a measure of probability that a hypothesized shape is "good." Previous approaches to analysis of cluster geometry [2], [3] have been heuristic, using simple shape models that are unsuitable for providing probability estimates. We further describe a strategy for combining shape and classification evidence to improve the performance of either approach by treating them as independent estimates of the probability that a hypothesized chromosome is genuine.

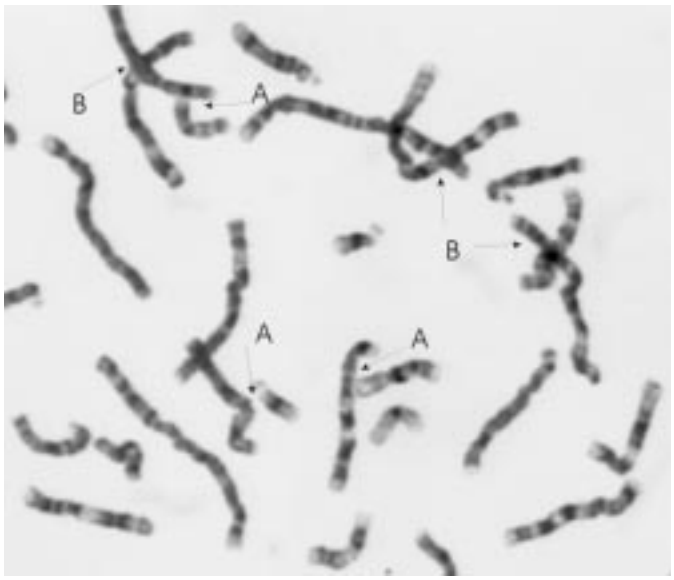


Fig. 1. Part of a dividing cell showing chromosomes stained to display a pattern of bands. The banding pattern can be used to classify the chromosomes into 24 groups. Isolated chromosomes can generally be segmented by thresholding, but chromosomes often touch each other (indicated by A) or overlap (B), requiring additional analysis for complete segmentation.

II. Sources of Evidence

We summarize our method using as an illustration the schematic pair of overlapping chromosomes in Fig. 2(a). The banding pattern in the overlapping region is obscured, but four short segments of chromosomes are visible. We assume that these segments can be detected by some suitable data-driven cueing process, such as skeletonization and identification of significant branches (see Section VI). Each of these segments could be matched to any of the other three to hypothesize six possible complete chromosomes [Fig. 2(b)–(d)]. We can evaluate the hypotheses by assessing the shape of the candidate complete chromosomes. Some of the shapes in Fig. 2(b)–(d) are much more likely to be real chromosomes than others. Also the banding patterns of some of the hypothesized chromosomes may correspond more closely with a genuine banding pattern than others.

Classification Evidence: In a normal cell, there are two chromosomes in each class (with the exception of the male sex chromosomes). Each member of a homologous pair has an identical banding pattern that is distinct from the banding patterns that are characteristic of all the other classes. Fragments of chromosomes, as in Fig. 2, can be matched

Manuscript received June 15, 1999; revised April 12, 2002. This work was supported by the Engineering and Physical Sciences Research Council, U.K. The data used in this study were made available as part of the Concerted Action of Automated Cytogenetics Groups, Project II.1.1/13, and the Concerted Action on Automated Molecular Cytogenetic Analysis, Project BMH1-CT92-1307, supported by the European Community. The associate editor coordinating the review of this paper and approving it for publication was Dr. Maria Joao Rendas.

G. C. Charters was with Imaging Science and Biomedical Engineering, the University of Manchester, Manchester U.K. He is now with the IBM U.K. Laboratories, Winchester, U.K. (e-mail: charters@uk.ibm.com).

J. Graham is with Imaging Science and Biomedical Engineering, the University of Manchester, Manchester, U.K. (e-mail: Jim.Graham@man.ac.uk).

Publisher Item Identifier 10.1109/TSP.2002.800421.

1053-587X/02$17.00 © 2002 IEEE

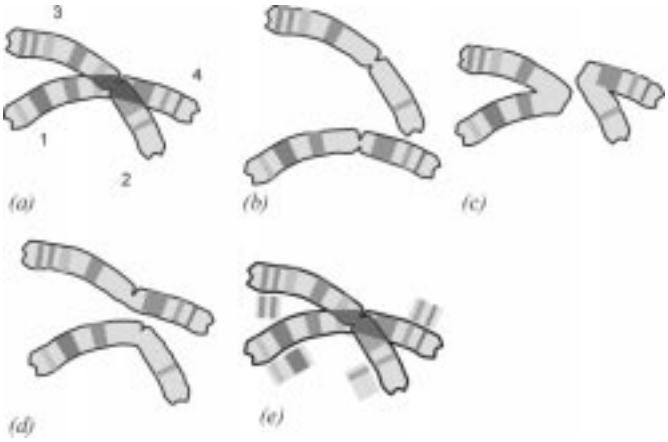


Fig. 2. Resolving a two-chromosome overlap. (a) Four isolated segments can be labeled. These can be combined to form six hypothesized chromosomes in three different solutions: (b) 1+4; 2+3. (c) 1+3; 2+4. (d) 1+2; 3+4. The probability of these hypotheses corresponding to genuine chromosomes can be assessed on the basis of shape. The chromosomes in (b) look most likely; those in (d) seem reasonable; those in (c) appear to be the least likely solution although not impossible as chromosomes are rather flexible (see Fig. 1). The banding pattern can also be used to assess the hypotheses by matching templates of short lengths of banding pattern to the visible segments (e). Evidence for a hypothesized chromosome will be strengthened if the matched segments can be paired with a high probability of belonging to a true chromosome.

by identifying that their banding patterns form part of the characteristic banding pattern of a particular class. The more a pair of fragmentary banding patterns appear to belong to the same class, the more likely they are to belong to the same chromosome. We have described in detail previously [1] how we obtain classification evidence for resolving overlaps. The method only uses short sequences of band pattern that are not obscured by the overlapped region. Each segment is matched with a large set of trained templates [partial chromosome models (PCMs), illustrated in Fig. 2(e)] to yield a feature vector for a linear classifier. The output of this classifier yields a probability $\Pr(C_k|\mathbf{u}_i)$ that segment i, represented by feature vector $\mathbf{u}_i$, should be assigned to class k, which is one of the 24 chromosome classes. (The possibility of a chromosome being observed in either of its possible orientations is dealt with at that stage.) We seek to match fragments to form a single chromosome by maximizing the probability $\Pr(C|\mathbf{u}_i,\ \mathbf{u}_j)$ that an object, which is formed by combining segments i and j, is a chromosome of *any* class. (C is the union of all chromosome classes.) Since $\mathbf{u}_i$ and $\mathbf{u}_j$ represent disjoint fragments of the banding pattern of a complete chromosome, we assume that $\Pr(C_k|\mathbf{u}_i)$ and $\Pr(C_k|\mathbf{u}_j)$ are independent estimates of the chromosome's class. We take the prior probabilities of all chromosome classes $\Pr(C_k)$ to be equal, allowing us to write $\Pr(C|\mathbf{u}_i,\ \mathbf{u}_j)$ in terms of the class-conditional probabilities:

$$\Pr(\mathbf{u}_i,\ \mathbf{u}_j|C) = \max_k \Pr(\mathbf{u}_i|C_k)\Pr(\mathbf{u}_j|C_k). \tag{1}$$

The assumption of equal priors is very closely true for almost all chromosome classes. A small bias is introduced concerning the sex chromosomes by using this assumption.

We showed that using combinations of short banding sequences, classification performance is comparable with that obtained using the entire banding pattern on unobscured complete chromosomes. Good classification performance was also obtained on chromosomes partially obscured by overlap.

Shape Evidence: As we wish to avoid heuristic approaches to evidence combination, we need a method of describing shape that yields a probability of being a valid chromosome shape. This will allow us to conduct the combination of shape and classification evidence in a principled, probabilistic framework. In Section III, we describe a method for representing shape using trainable models that allows us to calculate the probability that any given shape belongs to the distribution defined by the training set. In Section IV, we will describe the method of combining shape and classification probabilities and examine the assumptions necessary to make the problem tractable.

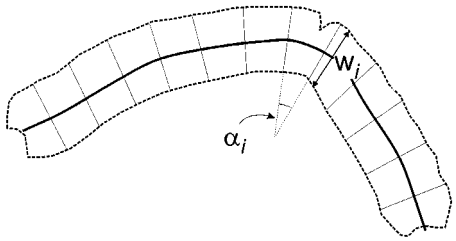


Fig. 3. Parameterization of the chord distribution model. A chromosome shape is described by a set of chords equally spaced along its axis. Each chord is parameterized by the angle it makes with the preceding chord (α_i) and the chromosome width (chord length) w_i. The length of the axis is also included in the shape description.

III. Trainable Shape Models

We base our shape models on the point distribution models (PDMs) of Cootes *et al.* [4], which describe shape using the statistics of a set of landmark boundary points. Among other desirable features, PDMs deal compactly with natural variability and are highly robust and specific, provided the shape variations are approximately linear (e.g., changes of scale or local deformations). Large nonlinear deformations, such as bending, reduce the specificity of the model both for image search and object description. Since bending is a universal feature of chromosome shapes, we modify the method for use in this case.

We call our shape parameterization the *chord distribution model* (CDM), and it is illustrated in Fig. 3. Each shape is described by a set of equally spaced chords lying perpendicular to the curved central axis. There are two parameters for chord i: 1) the angle change α_i between chords $i-1$ and i, and 2) the length of the chord (i.e., the width of the chromosome) w_i; each is scaled by the standard deviations of the parameters. The length l of the chromosome along the central axis is included, as this is an important parameter, varying dramatically from one chromosome to another. The vector $\mathbf{x} = (\alpha_0,\ w_0,\ \alpha_1,\ w_1,\ \ldots,\ \alpha_{n-1},\ w_{n-1},\ l)^T$ specifies the shape of a chromosome, which is represented by n chords. To maintain a consistent shape description, n is a constant across all shapes. In the case of a hypothesized chromosome extracted from an overlapping configuration, as in Fig. 2, it is a straightforward matter to locate n chords along the axis of the object formed by joining the segments.

Let $\mathbf{x}_i$ be the vector describing the ith example from a training set of N shapes. We can calculate the mean shape ($\overline{\mathbf{x}} = 1/N \sum_{i=1}^{N} \mathbf{x}_i$), the deviation of each shape from the mean ($d\mathbf{x}_i = \mathbf{x}_i - \overline{\mathbf{x}}$) and the covariance matrix ($\mathbf{S} = 1/(N-1) \sum_{i=1}^{N} d\mathbf{x}_i d\mathbf{x}_i^T$) for the training set. In general, considerable variability will be observed in the individual parameters. However, these variations are usually highly correlated and correspond to a set of *modes of variation* of the entire shape. The modes of variation may be found by calculating the unit eigenvectors of $\mathbf{S}$, $\mathbf{p}_i,\ i = 1 \cdots n$ ($\mathbf{S}\mathbf{p}_i = \lambda_i \mathbf{p}_i$), where λ_i is the ith eigenvalue, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_i \geq \lambda_{i+1} \geq \cdots \geq \lambda_n \geq 0$, and $\mathbf{p}_i^T \mathbf{p}_i = 1$. Any shape from within the range of training shapes can be recreated using the weighted sum of the eigenvectors (2)

$$\mathbf{x} = \overline{\mathbf{x}} + \mathbf{P}\mathbf{b}. \tag{2}$$

$\mathbf{P} = (\mathbf{p}_1,\ \mathbf{p}_2,\ \ldots,\ \mathbf{p}_n)$ is the matrix of eigenvectors, and $\mathbf{b} = (b_1,\ b_2,\ \ldots,\ b_n)$ is a vector of weights applied to each eigen-

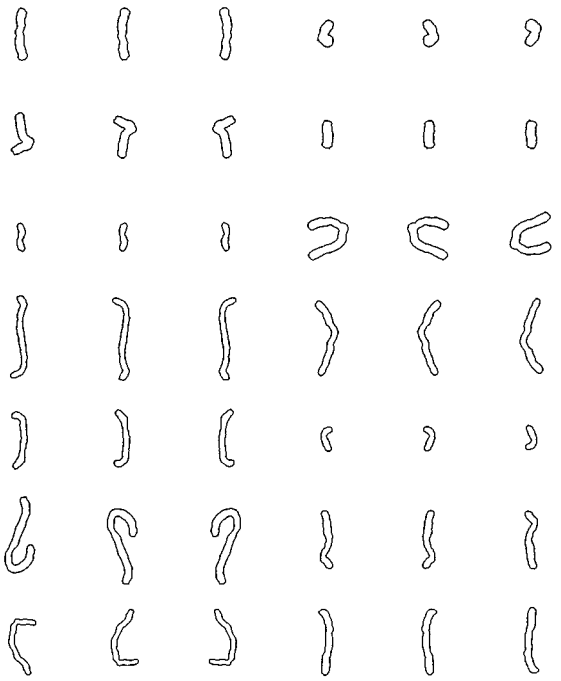


Fig. 4. Some examples of chromosome shapes in the shape training set. Since chromosomes can bend in different directions, the initial set was enlarged by reflecting chromosomes in the x and y directions.

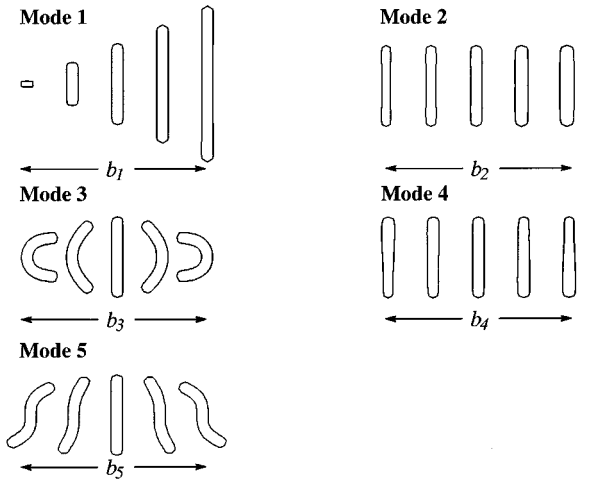


Fig. 5. Five most significant modes of variation of the chromosome training set. In each case, the mean shape appears in the center. The variation results from altering the weight (b) of the first five eigenvectors in turn over the range $-2\sqrt{\lambda_i}$ to $2\sqrt{\lambda_i}$. The modes correspond to intuitive descriptions of shape change. The most significant mode of variation is the chromosome length, followed by a slight variation in width. Bending of the chromosomes is accounted for in the third mode, bending with two axes of curvature appearing as mode 5. The fourth mode reflects change in chromosome width along the axis; chromosomes are not uniformly wide all the way along and, in particular, show a characteristic constriction (the centromere), which can occur at various positions, including the ends. Different weighted combinations of these modes can generate a wide range of "legal" chromosome shapes, i.e., shapes that could occur within the distribution of training shapes, even though individually, they may never have been observed.

vector. The vector $\mathbf{b}$ is equivalent to $\mathbf{x}$ as a shape description. Since the eigenvectors are orthogonal, we can generate $\mathbf{b}$ for any shape using (3). We call $\mathbf{b}$ the *shape vector* of a chromosome as it represents how closely a shape fits the model

$$\mathbf{b} = \mathbf{P}^T(\mathbf{x} - \overline{\mathbf{x}}). \tag{3}$$

The proportion of the variance described by each eigenvector is equal to its eigenvalue. The variance in the training data described by the t most significant eigenvalues is given by $V_t = \sum_{i=1}^{t} \lambda_i$. The total variance in the data is V_n. As there is a high degree of correlation amongst the parameters, the shape variation can be described by t eigenvectors, where t is selected to represent a large proportion of V_n while giving a shape vector with a dimension significantly reduced compared with the original $(2n+1)$. This generates a compact shape description while ensuring that the observed variability is adequately represented.

Fig. 4 shows some examples from the set of 1412 shapes used to train the shape model. These were extracted from chromosome images by interactive thresholding and manual isolation of overlapping chromosomes. Axes were formed by skeletonization followed by manual extension of the skeletons to the boundaries and spline fitting. Forty-five chords, which are normal to the spline axis and equally spaced along it, were generated for each shape. The eight most significant modes of variation describe 92.5% of the total variance in the training set. Fig. 5 shows the effect of varying each one of the first five modes in turn (varying each of the b_i individually over the range $-2\sqrt{\lambda_i}$ to $2\sqrt{\lambda_i}$).

The quality of fit of a given chromosome shape to the distribution of shapes can be assessed from the Mahalanobis distance $(D^2 = \sum_{i=1}^{t} (b_i^2/\lambda_i))$ from the mean shape using t modes of variation. We normalize this to a probability measure using the Chi-square distribution of D^2 with t degrees of freedom. Since the PCA removes linear correlations from the features, we take the b_i to be independent measures of shape, making the use of Chi-square appropriate. In addition, since we truncate the number of dimensions (modes) of the shape distribution, the Chi-square calculation allows straightforward comparison of results with different values of t.

Having expressed the shape in terms of a probability, it can be used to evaluate the candidate chromosome either on its own or combined with classification evidence.

IV. Combining Evidence

In resolving a shape, neither the classification evidence, which is outlined briefly in Section II, nor the shape evidence described in Section III is totally reliable. The experience of analysis of chromosomes by human expert observers is that the uncertainty in resolving the overlap should be reduced by using both. Fig. 2 represents the possible solutions of a two-chromosome overlap generating four identifiable segments—an "X-shaped" overlap. Two overlapping chromosomes can also form a three-segment, "T-shaped," cluster (see Fig. 6). The following argument is constructed in terms of the "X-shaped" overlap but can equally be applied in the "T-shaped" case. Fig. 2(b)–(d) represent three possible hypotheses for the resolution of the overlap. We have constructed the two sources of evidence so that each provides an estimate of the probability that a given hypothesized chromosome is valid. We can therefore conduct the evidence combination in a probabilistic fashion. Cast in Bayesian terms, we wish to identify the hypothesis with the maximum *a posteriori* probability. That is, we wish to choose

$$i^* = \arg \max_i [\max_c \Pr(H_i, \mathbf{C}|\mathbf{r})] \tag{4}$$

where H_i are separate hypotheses, $\mathbf{C} = (C_1, C_2)$ is a pair of possible class assignments, and $\mathbf{r}$ is a data vector. Rearranging this and following Bayes theorem, in terms of the class conditional probabilities and priors, we have

$$i^* = \arg \max_i [\max_c \Pr(\mathbf{r}|H_i, \mathbf{C}) \Pr(H_i, \mathbf{C})]. \tag{5}$$

We take the prior probabilities of the individual classes to be equal. We have noted that this is true to a close approximation. We also assume that all cluster geometries are equally probable. (It is generally accepted that chromosomes lie in random positions and orientations in the image.) Using these assumptions, we need to maximize the likelihood of the data, subject to the hypothesis and the assigned classes

$$i^* = \arg \max_i [\max_c \Pr(\mathbf{r}|H_i, \mathbf{C})]. \quad (6)$$

The data $\mathbf{r}$ for a given hypothesis consists of a pair of shape vectors $\mathbf{b_1}$ and $\mathbf{b_2}$ (see Section III) and four classification vectors $\mathbf{u_1} \cdots \mathbf{u_4}$ (see Section II). Each hypothesis consists of an association between each of the shape vectors and two of the classification vectors. Equation (6) becomes

$$i* = \arg \max_i [\max_c \Pr(\mathbf{b_1}, \mathbf{b_2}, \mathbf{u_1}, \mathbf{u_2}, \mathbf{u_3}, \mathbf{u_4}|H_i, \mathbf{C})]. \quad (7)$$

This is, in principle, a very large problem as we need to maximize over all shape configurations, each over all pairs of classes. We make the problem tractable by assuming that the shape and classification evidence are independent. This is a reasonable assumption up to a point. The banding pattern is largely independent of shape. (In fact, some distortion of the banding pattern occurs due to bending, but the effect of this on classification is minimal and has never been taken into consideration in the design of chromosome classifiers.) The shape is also at least partly independent of the class. One aspect of a chromosome's shape that is correlated with class is its length, which forms part of the shape descriptor (see Section III). Chromosome length is a feature commonly used along with the banding pattern in chromosome classification [5]. Our hypothesis testing becomes greatly simplified if we can factor out the shape from the classification to give

$$i^* = \arg \max_i [\Pr(\mathbf{b_1}, \mathbf{b_2}|H_i) \max_c \Pr(\mathbf{u_1}, \mathbf{u_2}, \mathbf{u_3}, \mathbf{u_4}|H_i, \mathbf{C})]. \quad (8)$$

This simplifying assumption can be justified on the grounds that we are assessing the probability that a hypothesized object is a chromosome of *any* class. For the purpose of resolving the overlap, we are not interested in knowing the class of the chromosome; the classification is merely a convenient mechanism for pairing up segments of the banding pattern. We therefore pool the length variability as part of the shape model. In principle, one could propose using the assigned classes of the hypothesized chromosomes to provide more information about the resolution. However, the resolution of the overlap is then influenced by the classes of all other chromosomes in the cell, which may in turn be involved in overlaps, leading rapidly to a problem of intractable complexity [6].

We further assume that the banding patterns and shapes of the candidate chromosomes arising from a given hypothesis are independent of each other. This is also a close approximation to the truth. Certainly, knowing the shape of one chromosome in an overlap provides no constraint on the shape of the other. There is some small interaction between the classes of the hypothesized chromosomes. It is less likely that they will belong to the same class than to different classes. Assuming independence of classes introduces a small bias into the classification evidence. However, the assumption allows us to write the probability of a given outcome as

$$i^* = \arg \max_i [\Pr(\mathbf{b_1}|H_i) \Pr(\mathbf{b_2}|H_i) \cdot \max_{c_1} \Pr(\mathbf{u_1}, \mathbf{u_2}|H_i, C_1) \max_{c_2} \Pr(\mathbf{u_3}, \mathbf{u_4}|H_i, C_2)]. \quad (9)$$

The probabilities in the left-most product are the shape probabilities described in Section III. The probabilities in the right-most product are derived using (1).

In summary, to resolve an "X-shaped" overlap such as that in Fig. 2, we visit each of the three hypotheses in turn and identify the two candidate chromosomes by generating a shape connecting the free ends. For each generated chromosome, we calculate the shape probability for each (see Section III) and the maximum class-conditional probabilities based on the combination of fragmentary banding patterns. Finally, we multiply all four probabilities, choosing the hypothesis that yields the largest result. A "T-shaped" overlap can be resolved in the same way. There are three hypotheses for each overlap in this case also.

V. Evaluation

Chromosome Data: We used a publicly available data set (the "600-band" data set) consisting of 6177 chromosomes from 136 G-banded blood cells, which has been used in a number of previous classification studies (see, e.g., [1] and [7]). For each chromosome, there is, among other things, an isolated image, the class specified by a cytogeneticist, and an identifier for its cell of origin. The chromosomes have been selected so that very few have banding patterns corrupted by overlaps in the original images.

Simulation of Overlaps: We used the isolated "clean" chromosome images to form simulated overlaps. The method of selecting the chromosomes to be involved in the overlaps was the same as that reported in our earlier paper [1]. In addition to the image and classification data, each chromosome record in the data set contains a banding profile, which is a linear representation of the banding pattern in which the chromosome density is projected onto the central axis. The length of the profile corresponds to the length of the chromosome. To select chromosomes to be involved in overlaps, the banding profiles from a given cell are appended and two random positions selected along the length of the aggregate cell profile. By selecting chromosomes in this way, the likelihood of a chromosome being involved in an overlap is proportional to its length. The isolated images of the selected chromosomes were added to a target image at randomly selected positions and orientations. Only some random configurations result in overlaps. If an overlap was present, it was evaluated as a "T-shaped" (three-segment) or "X-shaped" (four-segment) cluster according to whether or not four segments with skeletons longer than 15 pixels could be isolated. (Classification evidence cannot be generated for segments smaller than 15 pixels due to the size of the smallest template in the PCM set, which is defined to be half the size of the shortest chromosome class in the 600-band data set [1].)

Chromosome shapes and smooth axes were obtained for the simulated cluster, as described in Section III. The location of the overlap was defined to be a node in the skeleton of the cluster. This was confirmed manually, as we sought to evaluate the evidence combination rather than the preprocessing. Banding patterns, for use by the PCM classifier, were obtained by integrating density normal to the axis. A single overlap was generated from each alternate cell in the evaluation sets, resulting in 56 "T-shaped" clusters and 13 "X shapes." The small number of "X-shapes" arises from the limit on the smallest segment that can be used for classification evidence. Crossing chromosomes having one segment less than 15 pixels long were classified as "T-shapes."

Evaluation Experiments: We performed the following experiments. Each experiment was a cross-validation in which the data were split into reversible training and evaluation sets. PCM training was performed, and as described previously [1], CDMs were trained as in Section III.

Experiment 1: Resolution using classification evidence by setting all the $\Pr(\mathbf{b}_j|H_i)$ *to be equal.*

TABLE I
RESULTS OF OVERLAP RESOLUTION EXPERIMENTS. THE COLUMNS CORRESPOND TO THE THREE EXPERIMENTS: RESOLUTION ON CLASSIFICATION EVIDENCE, ON SHAPE EVIDENCE, AND COMBINED EVIDENCE. IN EACH CASE, THE PERCENTAGE OF CORRECTLY RESLOVED OVERLAPS IS SHOWN FOR X-SHAPED (FOUR-SEGMENT) AND T-SHAPED (THREE-SEGMENT) CLUSTERS, AS WELL AS THE OVERALL TOTALS

Overlap Type	Evidence Type		
	Classification	**Shape**	**Combined**
T-shape	**66.1%**	**78.6%**	**84.8%**
X-shape	**53.9%**	**92.3%**	**92.3%**
All	**64.5%**	**81.2%**	**86.2%**

Experiment 2: Resolution using shape evidence by setting all the $\Pr(\mathbf{u}_i, \mathbf{u}_j | H_i C)$ *to be equal.*

Experiment 3: Resolution using combined evidence by using trained values for both shape and classification evidence. For this experiment, we set a threshold such that if the classification probabilities of all segments were less than 0.05, they were set to be equal and the resolution took place on shape evidence alone. Classification evidence becomes extremely unreliable when all segments have low classification probability.

The results are shown in Table I. Aside from the encouragingly high proportion of correct resolutions, we can make several observations. "X-shaped" overlaps are well resolved on shape information alone. This is intuitively reasonable as "X-shapes" formed from the coincidence of very bent chromosomes [as in Fig. 2(c)] are rare. Conversely, the classification evidence is much less useful for "X-shapes" than for "T-shaped" clusters, making no apparent contribution to the resolution. The greater contribution of classification evidence in the case of "T-shapes" probably occurs because the segments are longer with more of the banding pattern visible than in the "X-shapes." The shape evidence is often more misleading for the "T-shapes" (see Fig. 6), and the combination of evidence makes a noticeable contribution to the resolution. Some examples of errors made on shape and classification evidence are shown in Fig. 6.

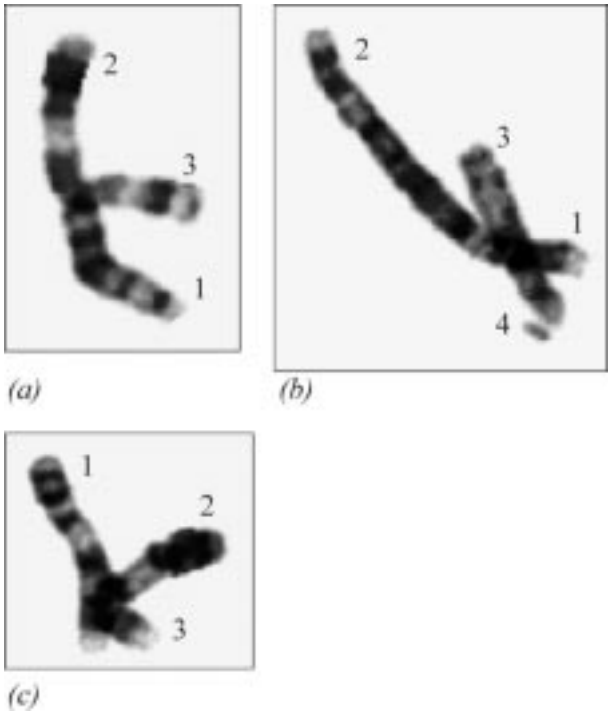


Fig. 6. Examples of overlaps, showing how evidence is combined. (a) Example in which the shape evidence is misleading, giving the incorrect solution that segment 3 is a single chromosome and that segments 1 and 2 should be joined (we use the notation {1+2;3}). The correct solution, which is generated by the classification evidence, is {2; 1+3}. (b) Showing the opposite effect. The classification evidence generates the wrong solution ({1+3; 2+4}), largely due to the paucity of banding information in segments 1 and 4. The shape evidence corrects this in the combined result to give {1+2; 3+4}. This is a fairly typical result in X-shaped overlaps. (c) Only example among the simulated overlaps where all the classification, shape, and combined evidence resulted in an incorrect solution. The correct solution is {1; 2+3}. The incorrect solution of {2; 1+3} arises because the skeleton for the combination {1+3} results in a more likely shape than {2+3}. The classification evidence suggests the wrong solution as segment 3 is extremely short, and the banding pattern at the end of segment 2 is corrupted due to an overlap at that position in the original data. As we have noted, the number of overlaps in the data was kept to a minimum but not totally eliminated; there is a small residue of corrupted banding patterns.

VI. CONCLUSIONS AND DISCUSSION

In this paper, we have dealt with ways of using available evidence to resolve clusters of chromosomes. Our premise is that clusters can be recognized as such by some other process. Furthermore, we assume that it is straightforward to locate the positions where chromosomes cross and, hence, identify the "uncorrupted" segments away from the overlap region. It has been shown in other studies [2], [8] that this "cueing" process can be achieved straightforwardly and robustly by analysis of the shape of the cluster. Ji [2] used an analysis of the skeleton and boundary based on curvature and the convex hull. (Our semi-manual cueing method was intended to approximate Ji's method.) Popescu *et al.* [8] proposed a method of analyzing the boundary and axis shape that is less heuristic than that of Ji, giving comparable results for axis and overlap location on small numbers of clusters. We have also restricted ourselves to dealing with the resolution of overlapping rather than "touching" chromosomes. In the latter case, the separation can often be achieved by following a "pale path" between appropriate boundary points. Ji used this method for unbanded chromosomes. The method is more difficult to apply to banded chromosomes as there are many "pale paths" across the chromosomes. Nevertheless, Graham [9] has addressed this issue using a split-and-merge approach in the context of an interactive system, and Popescu *et al.* [8] have proposed a mechanism for "pale path cutting" that involves evaluating the resulting candidate chromosomes in a "hypothesize and test" strategy. The same study also deals with overlapping chromosomes, assessing hypothesized objects constructed from segments on the basis of their banding pattern. This is one of a number of recent studies [10], [11] that use different methods of assessing the banding evidence for candidate chromosomes isolated from a cluster and are directly comparable with our earlier study (see [1, Sec.]). The studies of Stanley [11] and Popescu [8] are particularly relevant in dealing specifically with overlaps in this way.

We believe that this is the first attempt to combine evidence from shape and classification for resolving overlapping clusters. Ji's [2] method of resolving overlaps using geometric evidence provides an interesting comparison with the use of shape evidence reported here. In a trial on 46 two-chromosome overlaps, which in our terminology would be "X-shaped," he achieved a correct resolution of 94.6%. This is nearly identical to the resolution rate obtained here on our small set of "X-shapes" (Table I), which was also almost entirely based on geometric evidence, tending to confirm the intuition that "X-shaped" overlaps of two chromosomes can be reliably resolved on the basis of shape.

The studies by Agam and Dinstein [3] and Lerner *et al.* [10] are relevant in using, respectively, shape and banding evidence for resolving clusters, although both are restricted to touching or "slightly overlapping" configurations ("T-shaped" in our terminology). Agam and Dinstein [3] evaluate hypotheses using simple, heuristic models of shape. A

correct resolution rate of 82% overall and 88% on two-chromosome clusters is reported. These clusters contain an unspecified, but probably high, proportion of "touches." Lerner *et al.* [10] assess the hypothesized chromosomes by classification of the banding pattern with an MLP neural network. They evaluate the method on 46 images of two-chromosome clusters and achieve a "probability of correct segmentation" of 82.6%. This method, however, is subject to the serious constraint that the classes of the chromosomes involved need to be known *a priori*. The authors claim that this knowledge can be obtained using a "simple elimination criterion" after classification of isolated chromosomes. We feel that this claim grossly underestimates the difficulty of the problem. There would need to be no more than one overlap per cell, and all other chromosomes would need to be classified with very high accuracy; neither condition is reasonable. This constraint might be dispensed with by adopting the optimization strategy of Popescu *et al.* [8], who appear to achieve a lower correct recognition rate for overlapping chromosomes. However, their method applies a much more realistic analysis to more difficult clusters. Despite the close relationship between [3] and [10], these studies use shape and classification evidence as *alternatives*, perhaps due to the difficulty of combining essentially heuristic information. Our use of *trainable* models of shape allows this evidence to be combined with probabilistic classification evidence to resolve difficult cases.

We have not sought, with the small evaluation set used in this study, to infer the proportion of overlaps that could be resolved in a larger set of real images, particularly in the case of "X-shaped" overlaps. Our results are, however, consistent with those of previous studies. We can conclude that for an important class of overlap (the "T-shapes"), the combination of shape and banding evidence provides an advantage over the use of either in isolation and that the use of a trainable shape model provides a natural mechanism for the evidence combination.

In this work, we have restricted our attention to overlap resolution. The segmentation and classification of isolated chromosomes is a more straightforward problem that has been the subject of a number of studies (see, for example, [5] and [7]). Furthermore, we have limited ourselves to two-chromosome clusters, which is a limitation we share with other similar studies [2], [10]. The same evidence could, in principle, be used for larger clusters, but the complexity of analysis increases with the numbers of chromosomes involved. Even "X-shaped" or "T-shaped" objects could be composed of more than two chromosomes due to unfortunate alignments of small chromosomes. This has a rather lower likelihood of arising in practice than the two-chromosome overlaps investigated here, although such configurations are observed in the analysis of real metaphases. Again, the same evidence could be used to evaluate individual segments as for combinations of segments. It is possible to imagine a strategy for analyzing clusters by taking these possibilities into account. One might proceed heuristically (in the manner of Ji) by re-examining cases where the probability of combining segments to form a chromosome is low. In the current study, these have simply been resolved on shape evidence alone, assuming that two chromosomes are present. A more robust and computationally sound approach would be to use an optimization strategy such as that described by Popescu *et al.* [8]. It would take a longer study to quantify the error in segmentation and classification arising from such cases.

In evaluating our method on simulated overlaps, we have sought to avoid the problem of acquiring a sufficient number of expertly annotated clusters to form ground truth for both evaluation and training. This problem restricted Ji and Lerner *et al.* to a relatively small evaluation sets, which was somewhat larger in the study by Popescu *et al.* We have two models to train—a banding model and a shape model. The shape models were trained on a set of 1412 chromosomes independent of the 600-band images (see Section III). The banding model was trained and evaluated as a separate exercise, as described in our earlier paper [1]. In that case, 136 simulated overlaps were used in a cross-validation experiment, where the model for each overlap was trained on an independent set of about 4000 chromosomes. (However, the figures shown in Table I for "classification only" resolution correspond to banding profiles extracted from the simulated images generated in this study.) For the evaluation of evidence combination, the use of synthetic clusters lets us make use of an existing resource: a body of preclassified chromosome images. Furthermore, we can, in principle, generate as many overlaps of known true resolution as we like (although the process is rather time consuming). By using randomly selected and positioned images of real isolated chromosomes, we believe that we avoid the danger of introducing bias into the evaluation set. Visual inspection of simulated clusters (e.g., Fig. 6) suggests that in doing so, we do not generate objects of unusual appearance.

Automated karyotyping systems are now in common use in clinical cytogenetics laboratories and are available commercially from a number of vendors. These systems deal very well with the segmentation and classification of isolated chromosomes. While complex overlaps involving several chromosomes are observed in metaphase images, they tend to occur most frequently in images that are unsuitable for visual analysis in any case. The most frequently occurring type of cluster in images used in practice consists of a pair of touching chromosomes, and some commercially available systems have functions for dealing with these. However, separation of *overlaps* is left to operator interaction. The majority of overlaps to be resolved consist of pairs of chromosomes, and solving that problem reliably is the single most important improvement in functionality that could be brought to these systems. The fact that a number of recent publications have addressed this issue is evidence of both its practical importance and technical challenge.

REFERENCES

[1] G. C. Charters and J. Graham, "Trainable grey level models for disentangling overlapping chromosomes," *Pattern Recognit.*, vol. 32, pp. 1335–1349, 1999.

[2] L. Ji, "Decomposition of overlapping chromosomes," in *Automation of Cytogenetics*, C. Lundsteen and J. Piper, Eds. New York: Springer-Verlag, 1989, pp. 177–190.

[3] G. Agam and I. Dinstein, "Geometric separation of partially overlapping nonrigid objects applied to automatic chromosome classification," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 1212–1222, Nov. 1997.

[4] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models—Their training and application," *Comput. Vis. Image Understand.*, vol. 61, pp. 38–59, 1995.

[5] J. Graham and J. Piper, "Automatic karyotype analysis," in *Chromosome Analysis Protocols, Methods in Molecular Biology*, J. R. Gosden, Ed. Totowa, NJ: Humana, 1994, vol. 29, pp. 141–185.

[6] J. Piper, R. Baldock, S. Towers, and D. Rutovitz, "Toward a knowledge-based chromosome analysis system," in *Automation of Cytogenetics*, C. Lundsteen and J. Piper, Eds. New York: Springer-Verlag, 1989, pp. 275–289.

[7] P. Kleinschmidt, I. Mitterreiter, and J. Piper, "Improved chromosome classification using monotonic functions of Mahalanobis distance and the transportation method," *ZOR—Math. Meth. Oper. Res.*, vol. 40, pp. 305–323, 1994.

[8] M. Popescu, P. Gader, J. Keller, C. Klein, J. Stanley, and C. Caldwell, "Automatic karyotyping of metaphase cells with overlapping chromosomes," *Comput. Biol. Med.*, vol. 29, pp. 61–82, 1999.

[9] J. Graham, "Resolution of composites in interactive karyotyping," in *Automat. Cytogenet.*, C. Lundsteen and J. Piper, Eds. New York: Springer-Verlag, 1989, pp. 191–203.

[10] B. Lerner, H. Guterman, and I. Dinstein, "A classification-driven partially occluded object segmentation (CPOOS) method with application to chromosome analysis," *IEEE Trans. Signal Processing*, vol. 46, pp. 2841–2847, Oct. 1998.

[11] R. J. Stanley, J. Keller, P. Gader, and C. W. Caldwell, "Homologue matching applications: Recognition of overlapped chromosomes," *Pattern Anal. Applicat.*, vol. 1, pp. 206–217, 1998.

20. **The application of artificial neural networks to Doppler ultrasound waveforms for the classification of arterial disease.** J. H. Smith, J. Graham and R. J. Taylor, *International Journal of Clinical Monitoring and Computing, 13: 85-91, 1996.* doi:10.1007/BF02915843

International Journal of Clinical Monitoring and Computing **13**: 85–91, 1996.

© 1996 *Kluwer Academic Publishers. Printed in the Netherlands.*

The application of an artificial neural network to Doppler ultrasound waveforms for the classification of arterial disease

Julia H. Smith[1], Jim Graham[1] & Robert J. Taylor[2]
[1] *Department of Medical Biophysics, University of Manchester, Oxford Road, Manchester M13 9PT, United Kingdom;* [2] *Department of Medical Physics, Salford Royal Hospitals NHS Trust, Hope Hospital, Stott Lane, Salford M6 8HD, United Kingdom*

Received 14 May 1996; accepted 26 May 1996

Key words: arterial disease classification, artificial neural networks, doppler ultrasound, multi-layer perceptron

Abstract

In this study we have investigated the application of an Artificial Neural Net classifier to the diagnosis of vascular disease using Doppler ultrasound blood-velocity/time waveforms. A multi-layer perceptron network was trained with waveforms from control subjects and from patients with arterial disease. The diseased cases were confirmed by angiography and allocated to three groups according to the location of the stenosis: proximal or distal to the site of measurement or multi-segmental. We compared network classification results with a Bayesian classifier following a Principal Component Analysis of the waveforms. Versions of both classifiers were trained to discriminate two classes (normal v. abnormal) and four classes. In both cases the neural networks gave superior discrimination to the Bayesian classifier. While the four-class network was unable to provide useful discrimination among the stenosis sites, discrimination between abnormal and normal classes was obtained which is comparable to that achieved by a human expert observer.

Introduction

Vascular disease

Vascular disease is the most frequent cause of morbidity and mortality in the western world [1]. One of the commonest forms of the disease is atherosclerosis [2], a condition which affects the intima of the aorta and larger distributing arteries, and is responsible for approximately 60% of deaths from cardiovascular disease. The disease is characterised by the presence of raised plaques of fibrous fatty material which encroach on the lumen of the vessel and lead to the impairment of the arterial circulation. When the disease is present in the lower systemic circulation and the patient is at rest, it has been shown [3] that a reduction of 40% in the lumen diameter will significantly inhibit blood flow; the ischaemic effects of such a stenosis is enhanced when the limb is exercised.

Doppler ultrasound

Arteriography has been widely used as a diagnostic technique to assess the patency of diseased arteries. However, whilst being anatomically informative, it fails to give any indication of the haemodynamic function of the circulation. The recording and analysis of Doppler ultrasound blood-velocity/time waveforms at the site of the common femoral artery for the purpose of haemodynamic assessment of atherosclerotic diseased arteries has been well documented. It has been shown that quantitative analysis of the shape of the waveform can, in some cases, distinguish between partially and totally occluded vessels both proximal and distal to the site of measurement. The changes in waveform shape as a function of disease can be quite subtle. A range of methods of analysis have been applied from the straightforward, such as the Pulsatility Index [4], through to more complex approaches, for example Laplace Transform Analysis [5] and Principal Component Analysis (PCA) [6]. Various levels of success

have been reported for all these methods, but to date no single method has achieved both the sensitivity and specificity required to eliminate the need for invasive traumatic investigations such as angiography.

Artificial neural networks

Neural networks are devices, used mainly for recognition and classification tasks, consisting of a number of computing elements (nodes or neurones) highly connected by weighted links. The analogy with a biological network of neurones lies not only in the physical connectivity, but also in the fact that the function of the network is largely determined by the pattern of weights on the links (analogous to synaptic strength) and that this pattern is determined entirely by training the network on examples of the classes of objects it is to recognise. Many different network architectures have been designed, varying in the topology of the connections between nodes, the type of input data used (binary or continuous-valued) and the exact nature of the task being applied [7, 8].

Objectives

The aim of this study was to investigate the efficacy of an artificial neural network (ANN) applied to the disease classification of Doppler waveforms recorded at the site of the common femoral artery of normal and atheromatous arteries. In particular we wished to evaluate the performance of the ANN in comparison with that of PCA in conjunction with a Bayes classifier [9] when applied to Doppler signals obtained in routine clinical investigations. A previous preliminary study (Taylor, unpublished) had shown that Bayesian classification following PCA was capable of some discrimination among disease categories using data of this form. To be sure of ground-truth, our study was carried out retrospectively on waveforms derived from arteries whose disease classification had been confirmed by angiography in the course of normal clinical investigations.

Methods

Digitisation of the doppler waveforms

Waveforms in the clinical archive are recorded as paper charts. Digitisation of this archive data involved tracing the outline of each waveform over one complete cardiac cycle using a digitisation tablet attached to a personal computer. This data was then normalised both in amplitude (the lowest amplitude assumed the value 0, the highest 1) and time, and digitised into thirty three samples. The final data were stored in an ASCII data file on the same personal computer.

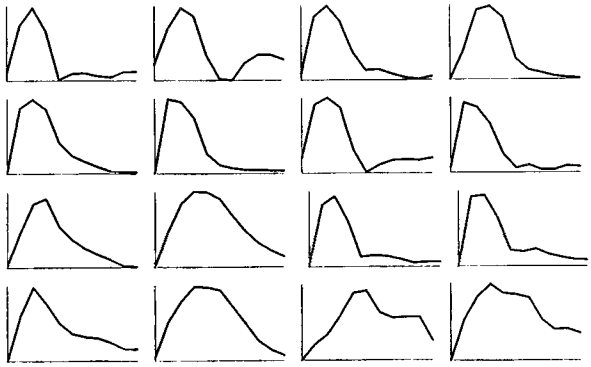


Figure 1. Selected examples of sampled waveforms indicating the range of appearance in the data set. Top row: Controls; Second row: Distal Disease: Third row: Proximal Disease; Bottom row: Multi-segmental disease.

In order to reduce the number of data input nodes to the ANN, and so reduce the processing time required, each waveform had the number of data points representing its profile reduced to eleven. Examples of the resulting sampled waveforms are shown in Figure 1. This data reduction was deemed acceptable as Fourier transforms of the original waveforms revealed negligible energy above the fifth harmonic.

Disease classification and clinical groups

Patients were grouped according to the site and severity of their arterial disease. Four groups, one control and three disease groups, were defined from a total of 408 Doppler waveforms taken from a cohort of 219 subjects (161 males, mean age 61 years, age range 33 to 84 years). Classification was based on both a clinical and angiographic assessment of the patients' arteries. The control group comprised waveforms recorded from subjects clinically judged to show no evidence of stenoses. The second group comprised data obtained from patients presenting with predominantly significant stenoses distal to the site of measurement (disease in the superficial and/or common femoral arteries). Group 3 included patients with prevalent disease proximal to the common femoral artery (disease in the aorta and/or iliac arteries). Patients with multi-segmental

Table 1. Number of Doppler waveforms in each disease classification group

Disease classification	Number of waveforms
Control	128
Distal	108
Proximal	88
Multi-segmental	84

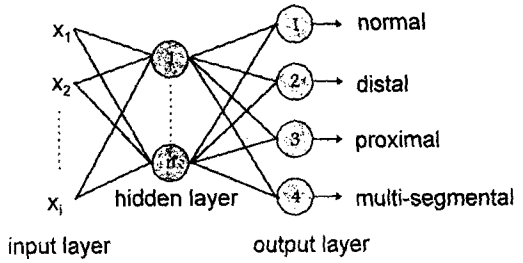


Figure 2. The Network Structure.

disease comprised the final group (significant stenoses both proximal and distal to the site of measurement). Examples of Doppler waveforms associated with each of these groups are presented in Figure 1 and the number of waveforms included in each group are listed in Table 1.

Artificial neural network

The Doppler waveform data set described above, in which continuous valued data has previously been assigned to known disease classes, lends itself naturally to the use of a multi-layer perceptron (MLP) network architecture. The MLP was simulated on a personal computer and coded in the C++ programming language. Figure 2 shows the configuration of the simulated network. The input layer consists of eleven nodes; the number of nodes in the single hidden layer is a variable parameter. The output layer comprises either two or four nodes depending on the classification experiment being performed (see below), each output node corresponding to a disease classification.

The combining and transfer functions are the same for each node, being the weighted sum of the inputs and the sigmoid function respectively. The output y_j from a node in layer j, which receives input from layer i, is represented by equation 1.

Table 2. The optimum network parameters for each of the discrimination tasks

Parameter	2 class network	4 class network
N° hidden nodes	10	10
Gain	0.09	0.03
Momentum	0.9	0.9

$$y_j = \frac{1}{1 + e^{-k.\sum_{i=0}^{n} w_{ij} x_i}} \quad (1)$$

where w_{ij} = weight from node i to node j; x_i = output from node i and k is a factor which controls the 'spread' of the sigmoid function (a value of $k = 1$ was used in this study).

The classical 'back-propagation' training algorithm was employed throughout the analysis [10, 11]. This algorithm is the most commonly used training mechanism for the MLP network. Weights are initially assigned random values, and are then altered in succeeding passes of the training data in a way which minimises the difference between the expected and observed outputs for each pattern in the training set [7]. Following training, the weights are configured so as to define decision regions in the space of the input vector. Thus an unseen waveform on presentation to the network, is classified according to the node in the output layer which yields the highest output value.

Network optimisation

The network is defined by a number of adjustable parameters: the number of nodes in each layer, the gain (or learning rate) η and the momentum α. The learning rate governs the amount by which the weights change on each pass of the training data and the momentum stabilises the convergence by encouraging consecutively similar weight changes and damping oscillations. The number of nodes in the input layer is determined by the resolution of the Doppler waveform and the number of output nodes is determined by the number of disease classes defined. These parameters are therefore fixed for a particular study. The number of intermediate hidden nodes may be varied to alter network performance. An optimum set of parameters was found empirically for each network by initially applying a coarse search followed by a more detailed inspection in the vicini-

ty of the combinations which gave discrimination on the training data set. The parameter values used in this study are shown in Table 2.

Network training and assessment

The ability of the ANN to discriminate between various disease groups was assessed as both a two-class and four-class problem (2 and 4 output nodes respectively). Initially, the data were collapsed into two classes, one containing normal waveforms the other waveforms from arteries with significant disease; the latter included equal portions of data from patients with distal, proximal and multi-segmental diseased limbs. Each class was represented by 125 waveforms. These were divided into 5 blocks, four being used for training and one as an unseen cross-validation set for testing discrimination power. The experiment was repeated with each of the blocks in turn acting as the cross-validation set for a network trained with the remainder of the data. The overall result is of a network trained using 200 waveforms and assessed using a cross-validation data set of 250 waveforms (125 in each class).

The four-class problem used the data preserved in four disease groups (Table 1). The same training strategy was used, the data being divided into six blocks in this case. The net effect was that of training on a total of 280 waveforms (70 in each class) and testing on a cross-validation set of 336 waveforms (84 in each class).

Principal component analysis

The use of principal component analysis (PCA) as a method of extracting salient features from Doppler ultrasound blood velocity waveforms has been well established [6, 12–15]. Application of a Bayesian classifier to the features generated by the PCA gives probabilities of each observation belonging to each of the disease classes, allowing decision surfaces to be constructed. This method has been discussed in detail by Evans [9].

PCA was applied to the same data sets as those used in the ANN two-class and four-class class analysis problems. Initially, the first two principal components were calculated from data in the training sets. (The first two principal components explained 82% and 86% of the variance in the data for the two-class and four-class problems respectively.) These were then used to calculate the first two principal component coefficients (PCC) for each of the Doppler waveforms in the cross-validation data sets. The Bayes Classifier technique was finally applied to these first two PCC in order to categorise each Doppler waveform into the clinical group with the greatest derived probability.

Table 3. Decision matrices for the two-class problem: a) Multi-Layer Perception discriminator; b) PCA using a Bayesian discriminator

a		Angiographic classification	
		Normal	Disease
Derived Class	Normal	116	22
	Disease	9	103

b		Angiographic classification	
		Normal	Disease
Derived Class	Normal	59	38
	Disease	66	87

Expert visual assessment of the doppler waveforms

In order to provide some measure of the information available in the Doppler waveforms, the vascular flow laboratory technicians were asked, as experts, to classify a set of waveforms according to one of the four disease groups. The classification had to be based only on visual content of the waveform's profile, as the patient's history and angiograms were not provided. Only typical reference waveforms were provided in order to aid them in their analysis. One hundred and ten waveforms, taken at random from all the four disease groups, were provided for this study.

Results

Class discrimination

Figure 3 shows the learning curves for the networks trained on the Doppler waveforms using the parameters listed in Table 2 for both two-class (solid line) and four-class (broken line) discrimination. The curves show the percentage of disease group misclassifications for the cross-validation data set as a function of the number of presentations of the training data set. Network stability occurs within about 100 passes of the training data. The lowest misclassification rates appear after 75 passes for the two-class network and after 250 passes for the four-class network. Tables 3 and 4 compare the network's performance at its best configuration, against

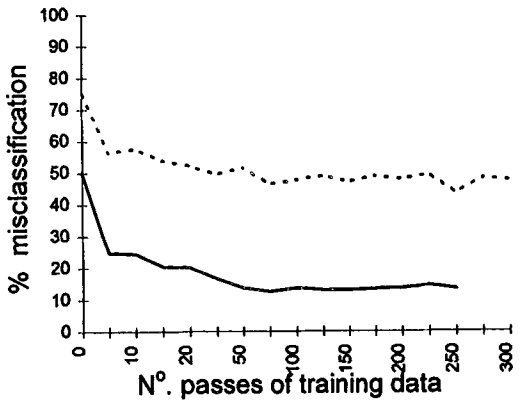


Figure 3. Learning Curves for the two networks using optimal network parameters, showing the proportion of the cross-validation data set misclassified as training proceeds. Solid Line = Two-class network; Dotted Line = Four=class network.

Bayesian classification on Principal Components in the two-class and four-class tasks. Total misclassification rates together with sensitivity and specificity values are shown in Tables 5 and 6. In calculating the sensitivities and specificities for Table 6, the three disease groups were considered as one group, as in the case of the two-class problem. However, the overall misclassification rate was calculated for all four groups.

Expert visual assessment of the doppler waveforms

Results of this study revealed that using visual inspection alone, it was only possible to reliably distinguish between normal and abnormal waveforms, and so results were only compiled for the two-class problem. An overall misclassification rate of 13.6% was achieved (82.9% sensitivity and 96.4% specificity).

Discussion

Previous studies investigating methods of quantitative assessment of Doppler waveforms have compared PCA, Pulsatility Index (PI) and Laplace Transform Analysis (LTA). Overall these studies have favoured PCA to be more sensitive and specific in terms of disease classification [6, 13, 16]. Evans et al. [13], for example, found that PCA was able to distinguish stenoses of less than 78% area reduction in the dog model, whilst PI and LTA were only able to distinguish stenoses of greater than 85% area reduction. If these data was categorised into groups of stenosis severity, (0–51%, 65 or 77%, 85 or 88% , 92 or 95% area reduction), it was found that PCA had an accuracy of 75%, the remaining 25% of the data being classified into a category one more severe than it actually was [6]. Evans et al. [16] found all three techniques to be successful at separating badly diseased arteries from normal arteries in the human model. However, only PCA worked well in the presence of less severe disease. This appeared to be due to its ability distinguish between limbs with and without distal disease. Sherriff and Barber [15] used PCA on Doppler waveforms recorded at the site of the carotid artery in patients with extracranial carotid artery disease. Their work demonstrated a sensitivity of 90%, a specificity of 77% and an overall accuracy of 85%.

Our principal aim in this study was to investigate the performance of a MLP classifier in this problem. Previous studies, while not directly comparable to this particular investigation, had suggested that PCA would probably be more effective than other approaches. In our hands, the neural network classifier significantly outperforms the Bayesian classification based on Principal Components.

As might be expected, the discriminating power of both classifiers in the four-class problem is inferior to that in the two-class problem. In the case of the MLP, the overall misclassification rate increases dramatically and the specificity is reduced in attempting to discriminate four groups. The principal source of error contributing to the misclassification rate in the four-class problem is the increase in misassignments among the disease groups. There is also a tendency (rather less pronounced) to assign controls to disease groups, presumably because there are a greater number of output classes to which ambiguous normal waveforms may be assigned. This increased false positive count is expressed as a reduction in specificity. Discrimination among the disease groups on the basis of the waveforms alone is clearly difficult, as can be supported by the attempt at visual classification by expert inspection. However, whilst this discrimination is poor, it is by no means random. Calculation of Cohen's Kappa statistic [17] gives a value of 0.32. (A value of 0.0 equates to random assignment and a value of 1.0 infers perfect discrimination.) The network has learned some basis for distinguishing the various disease groups.

The MLP is clearly the classifier of choice in this application. Its performance in the two-class problem is comparable to that found in the studies carried out by Allen and Murray [18, 19], in which ANN inputs were trained on photoelectric plethysmographic waveforms. The size of both their training and test data sets

Table 4. Decision matrices for the four-class problem. a) Multi-Layer Perceptron discriminator; b) PCA using a Bayesian discriminator

a		Angiographic classification			
		Normal	Distal	Proximal	Multi-seg
Derived	Normal	70	18	8	6
class	Distal	5	38	14	11
	Proximal	2	7	33	18
	Multi-seg	7	21	29	49

b		Angiographic classification			
		Normal	Distal	Proximal	Multi-seg
Derived	Normal	34	22	35	15
class	Distal	33	39	10	20
	Proximal	14	12	5	10
	Multi-seg	3	11	34	39

Table 5. Misclassification rates, sentivity and specificity for the two-class problem

Disease classifier	Misclassifications (%)	Sensitivity (%)	Specificity (%)
MLP	12.4	82.4	92.8
Bayesian, PCA	41.6	69.6	47.2

Table 6. Misclassification rates, sensitivity and specificity for the four-class problem

Disease classifier	misclassifications (%)	sensitivity (%)	specificity (%)
MLP	43.4	87.3	83.3
Bayesian, PCA	65.2	71.4	40.5

are similar to those used here; however, their disease groups were classed according to disease severity (normal, significant and major peripheral vascular disease), rather than the site of disease. Generally, our results are comparable to theirs although our values of overall misclassification rate and specificity are better than the 20% and 63% achieved in their prospective study. Allen and Murray do not report the use of a network trained specifically on two classes; our experience suggests that this is useful if only two-class discrimination is required.

Whilst the ANN proved to be of little use in localising the site of disease in the peripheral circulation, it shows a greater potential in discriminating between those patients whose arteries are normal and those who need further investigation. Every patient who attends the vascular clinic has to undergo a Doppler ultrasound test. The extra few minutes taken to process the results through the ANN may be rewarded by savings made, in trauma, time and cost, on further unnecessary invasive diagnostic tests. Obviously, in this mode the test requirement is 100% sensitivity with some acceptable specificity less than 100%, such that few normals are further referred. The MLP used in the two-class problem currently falls short of the sensitivity requirement, but achieves discrimination comparable to expert humans. This is an encouraging result, particularly as we note that the size of the training data set is rather limited. The appropriate size of training set is difficult to estimate directly. James [20] suggests that a 'large' training set is one in which the number of examples is more than ten times the number of features. For our eleven-feature classifier, this implies that 110 data items would be sufficient, and that the classifier in this study is sufficiently trained. This conclusion however is contrary to common experience with neural networks. In a study using statistical classification methods on a real classification problem for which very large quantities of training data were available, Piper [21] has shown that the optimum size of training set is great-

ly in excess of James' estimate, and is related to the complexity of the classifier. He proposes that a suitable estimate may be ten times the number of parameters to be determined. Errington [22] has applied a neural network classifier to the same data as Piper and shown that the size of training set required for best generalisation of an MLP is about the same as that required for unbiased statistical classification. Furthermore, taking the number of parameters in the network to be the number of weights, Piper's rule of thumb gives an estimate of the number of training data needed for Errington's network which conformed closely with the empirical findings. Extrapolating (circumspectly) from the results of these studies to the current problem, we reach the conclusion that to achieve optimum generalisation our network of 11 input nodes 10 hidden nodes and two output nodes would require a training set of rather more than ten times that used in this study. While obtaining this quantity of training data would be a significant exercise in itself, the experience of Errington[22] and Piper [21] suggests that substantial gains in sensitivity and specificity could be achieved.

Acknowledgements

The authors wish to extend their grateful thanks to Mr R.W. Marcuson (Consultant Surgeon, Hope Hospital) for allowing them access to his patients' data. Also they are indebted to the staff of the Vascular Flow Laboratory, Hope Hospital for their help in collecting patients' individual Doppler waveform traces and medical records.

References

Forrest APM, Carter DC, Macleod IB. Principles and Practice of Surgery. Edinburgh: Churchill-Livingstone, 1992.

Souhami RL, Moxham J, editors. Textbook of Medicine. Edinburgh: Churchill-Livingstone, 1990.

Gosling RG, Dunbar G, King DH, Newman DL, Side CD, Woodcock JP, Fitzgerald DE, Keates JS, MacMillan D. The quantitative analysis of peripheral arterial disease by a non-intrusive ultrasonic technique. Angiology. 1971; 22: 52–5.

May AG, Dewesse JA, Rob CG. Haemodynamic effects of arterial stenosis. Surgery 1963; 53: 513–24.

Skidmore R, Woodcock JP. Physiological Interpretation of Doppler-shift Waveforms – I. Theoretical considerations. Ultrasound in Med. & Biol. 1980; 6: 7–10.

Prytherch DR, Evans DH, Smith MJ, Macpherson DS. On-line classification of arterial severity using principal component analysis applied to Doppler ultrasound., Clin. Phys. Physiol Meas. 1982; 3: 191–200.

Beale R, Jackson T. Neural Computing: An Introduction. Bristol: Institute of Physics Publishing Ltd, 1992.

Clark JW. Neural Network Modelling. Phys. Med. Biol. 1991; 36: 1259–317.

Evans DH. The interpretation of continuous wave ultrasonic Doppler blood velocity signals viewed as a problem in pattern recognition. J. Biomed. Eng. 1984; 6: 272–80.

Rumelhart DE, Hinton GE, Williams RJ. Learning Representations by back-propagating errors. Nature 1986; 323: 533–6.

McClelland JL, Rumelhart DE. Exploration in Parallel Distributed Processing: A handbook of Models, Programs and Exercises. Cambridge, MA: MIT Press, 1988.

Martin TRP, Barber DC, Sherriff SB, Prichard DR. Objective feature extraction applied to the diagnosis of carotid artery disease using a Doppler ultrasound technique. Clin. Phys. Physiol. Meas. 1980; 1: 71–81.

Evans DH, MacPherson DS, Bentley S, Asher MJ, Bell PRF. The effect of proximal stenosis on Doppler waveforms: a comparison of three methods of waveform analysis in an animal model. Clin. Phys. Physiol. Meas. 1981; 2: 17–25.

Barber DC, Sherriff SB. Carotid artery blood flow: single factor classification of Doppler waveforms. Clin. Phys. Physiol. Meas. 1986 ; 7: 271–5.

Sherriff SB, Barber DC. A simple quantitative screening test for the detection of extracranial carotid artery disease. Clin. Phys. Physiol. Meas. 1989; 10 (Suppl. A): 23–32.

Evans DH, Macpherson DS, Bell PRF. A comparison of three methods of analysis of Ultrasonic Doppler waveforms recorded from the common femoral artery of patients with vascular disease. Ann. 8th Brazilian Biomed. Engng. Cong. 1983: 112–7.

Bland M. An Introduction to Medical Statistics. Oxford: Oxford University Press, 1993.

Allen J, Murray A. Development of a neural network screening aid for diagnosing lower limb peripheral vascular disease from photoelectric plethysmography pulse waveforms. Physiol. Meas. 1993 ; 14: 13–22.

Allen J, Murray A. Prospective assessment of an artificial neural network for the detection of peripheral vascular disease from lower limb pulse waveforms. Physiol. Meas. 1995 ; 16: 29–38.

James M. Classification Algorithms. Chichester: Wiley, 1985.

Piper J. Variability and bias in experimentally measured classification error rates. Patt. Recog. Lett. 1992; 13: 685–92.

Errington PA. Application of Artificial Neural Networks to Chromosome Analysis, (PhD Thesis), Manchester (UK): Univ. of Manchester, 1995.

Address for correspondence: Jim Graham, Department of Medical Biophysics, University of Manchester, Oxford Road, Manchester M13 9PT United Kingdom

Statistical Models of Shape and Appearance

21. **Locating overlapping flexible shapes using geometric constraints.** D.H. Cooper, C.J. Taylor, J. Graham and T.F. Cootes, *Proceedings of the British Machine Vision Conference, Glasgow, 1991. Springer-Verlag. pp 185-192.* doi:10.5244/C.5.24

Locating Overlapping Flexible Shapes Using Geometrical Constraints

David H. Cooper, Christopher J. Taylor, Jim Graham, Tim F. Cootes.

Department of Medical Biophysics
University of Manchester
Oxford Rd.
Manchester M13 9PT

Abstract

In an earlier paper [1] we have proposed a shape representation called the CLD (Chord Length Distribution) which possesses many of the often-quoted desirable properties of a shape representation. It also captures shape variability and complements an object location method using belief updating which integrates low-level evidence and shape constraints. Promising results on synthetic and real rigid objects were given. This paper describes a development to the original definition which makes the location method robust with respect to clutter. We give experimental results which demonstrate the performance of the revised scheme on a class of flexible shapes, both singly and overlapping.

We are currently engaged in a research project [see acknowledgements] concerned with automated 2-D inspection of complex (industrial) assemblies. In common with many machine vision applications we seek to exploit object shape and other geometrical constraints to assist in locating objects in scenes and evaluating interpretations with respect to expected appearance. To this end we need suitable representations for shape (intra-object) and inter-object relationships together with location and verification schemes capable of exploiting such representations. Ideally we seek a scheme capable of addressing both shape and inter-object relationships in a uniform manner.

We have argued [1] that a shape representation not only needs to satisfy often-quoted [2,3] properties of being easily computable, unique, and exhibiting proportional behaviour, but must also describe expected variability and invariance within a class of shapes and be capable of describing a wide range of shape classes. We have proposed such a representation called a Chord Length Distribution (CLD) and an associated object location scheme which exploits and integrates geometrical (shape) constraints with low-level (edge) evidence in a principled way, originally based on ideas derived from probabilistic reasoning using networks [4].

Unlike many reported methods of applying shape models [5,6,7,8] our approach *does not work by matching image primitives to related model elements*. Rather, it seeks to label each point in an ordinate space with a likelihood of correspondence to the

model. This likelihood is maximised with respect to the image evidence (edge data) and the shape constraints in the model. The advantage of this approach is the late commitment to an interpretation – the highest level primitive used is the pixel. This is particularly important in the context of overlapping or occluded objects.

This paper presents further developments and investigations into the properties of the CLD but first we give a brief outline of the CLD and the object location method. The reader is referred to [1] for a detailed description.

1 CLD and OBJECT LOCATION

A shape is first defined by a set of n points x_1 .. x_n. These may be equally spaced around the boundary but this is not necessary and it may be the case that, for a given value of n, an unequally spaced set of points may provide a more stable description, particularly for man–made objects. The only requirement is that there is a consistent method of selecting the points when the shape or family of shapes is defined. A reference point x_0 is also defined for the object. The shape representation consists of the set of probability distributions $P(r_{ij}) : i, j = 0 .. n, i \neq j$ for the distances r_{ij} between all pairs of points x_i, x_j. The arrangement is illustrated in Fig. 1.

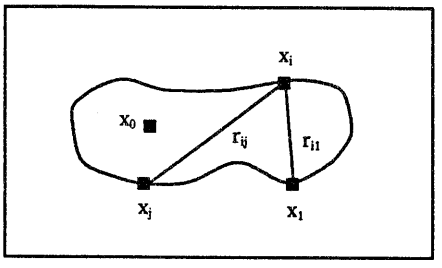


Fig. 1: Geometry of the CLD representation.

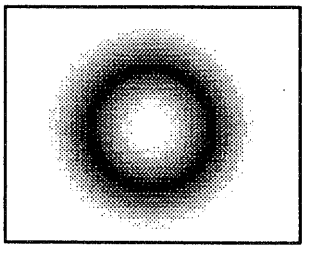


Fig. 2: $P(x_j | x_i)$

The probability distributions can be estimated from a set of example images in which the correct locations of the shape–defining points have been established independently, usually via an interactive training procedure. When the objects of interest are rigid, all the $P(r_{ij})$ will have low variance and the shape will be highly constrained. When the objects of interest are variable, some, though generally not all, of the $P(r_{ij})$ will have high variance and some aspects of the shape will be less constrained. Various other properties are discussed in [1] but the only one of relevance here is that the representation is unique except with respect to mirror symmetry.

Object location depends on the fact that the radial distributions $P(r_{ij})$ allow us to predict where x_j is given the position of x_i by rotating the radial distribution about the origin $x_i = 0$ as shown in Fig. 2.

The key to our method is to store a probability map $P(x_i)$ for each of the n points which define the shape. Each location in the map is labelled with a likelihood of finding x_i there. We can compute a prediction for x_j at all points by correlating $P(x_i)$

with $P(x_i|x_j)$. For each x_i in turn we compute (n–1) predictions for $P(x_i)$ from each of the other x_j and combine them with the original x_i to produce new estimates for the locations of each x_i. This belief updating process is repeated until a stable, maximally consistent interpretation is reached.

The initial values of the maps are generated by combining predictions made from the expected (prior) position of the reference point x_0 and edge data obtained from the image.

1.1 Behaviour With Clutter

The scheme outlined above is very successful in locating single instances of an object in a field in the presence of noise [1]. However, the method can sometimes converge to an incorrect result for multiple objects in the circumstance where the distance between the objects is comparable with or less than the chord lengths of the objects. It is easy to see how this can arise.

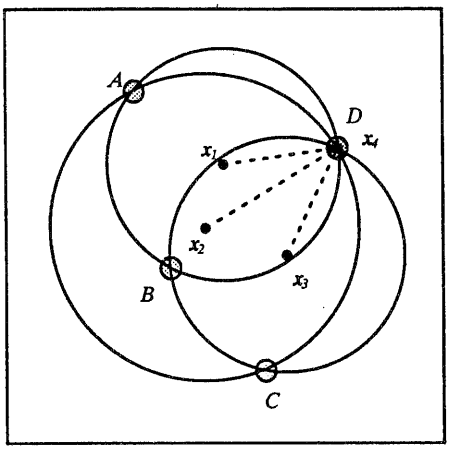


Fig 3: predictions using chords only

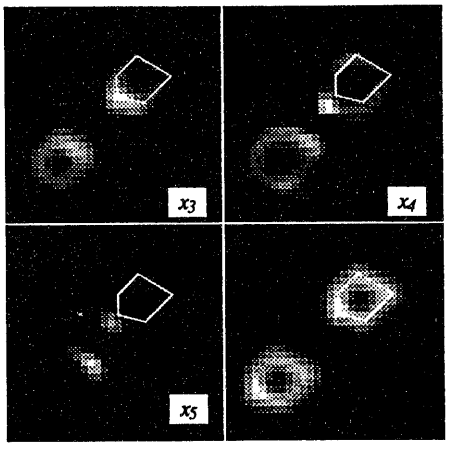


Fig 4: maps after 4 iterations for two similar polygonal objects (a) top = x_3, x_4 (b) bottom = x_5, original map

Fig. 3 depicts 3 points x_1, x_2, x_3 at known positions. The circles represent the predictions for a 4th point x_4, given x_1, x_2, x_3. In this case the predictions combine in a fashion analogous to a voting scheme as used by Hough transforms [9]. In our case, the belief in the location of x_4 is also weighted by the edge evidence, which may be stronger at A,B,C than at D resulting in incorrect convergence of the updating scheme.Fig. 4 shows an example with two similar polygonal objects whose vertices are labelled x_1 to x_5 clockwise from the bottom. Only one polygon icon is drawn. The initial maps are bottom right. The maxima in the top left, top right and bottom left diagrams should correspond to vertices x_3, x_4 and x_5 but clearly do not. (compare this with Fig. 7).

This problem can be overcome by developing the CLD to remove the reflectional symmetry ambiguity.

1.2 The Revised CLD Representation

The modification is illustrated in Fig. 5. We have introduced angles θ_{ij} which describe the the angle that the normal to the boundary at $\mathbf{x}_i$ has to be rotated anti-clockwise to indicate the direction to $\mathbf{x}_j$. The choice of object-related direction is arbitrary – in fact we use the direction of the image gradient at $\mathbf{x}_i$ in our experiments. We record the distributions $P(\theta_{ij})$ as part of the model.

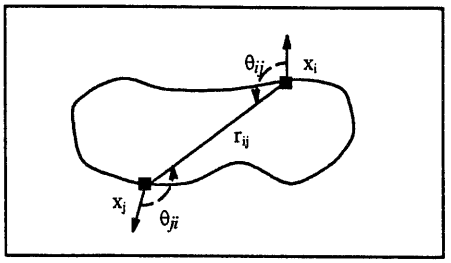


Fig. 5: CLD with direction information

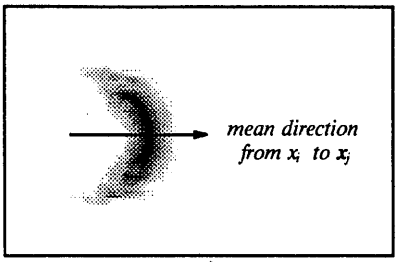


Fig. 6: $P(x_j|x_i)$

The conditional probability maps $P(\mathbf{x}_i|\mathbf{x}_j)$ become reduced annulli as depicted in Fig. 6, where the angular dispersion is determined by the variance in θ_{ij}. These new maps produce far more constrained predictions and result in faster and more stable convergence.

2 EXPERIMENTAL RESULTS

2.1 Nearby Objects

It is easy to see that the situation depicted in Fig. 3 is far less likely to occur when conditional maps incorporating angle statistics are used. The results for the same polygon pair as in Fig. 4 are shown in Fig. 7.

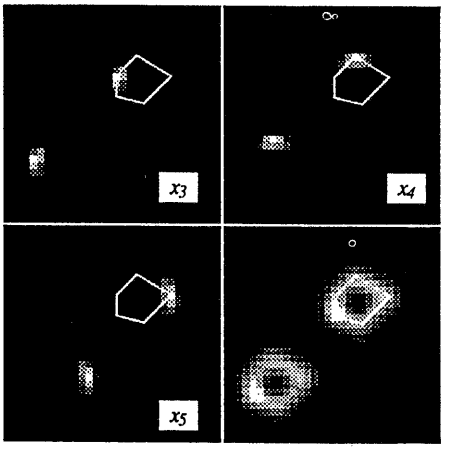


Fig 7: maps after 4 iterations for two similar polygonal objects (a) top = x_3 , x_4 (b) bottom = x_5, original map (revised scheme)

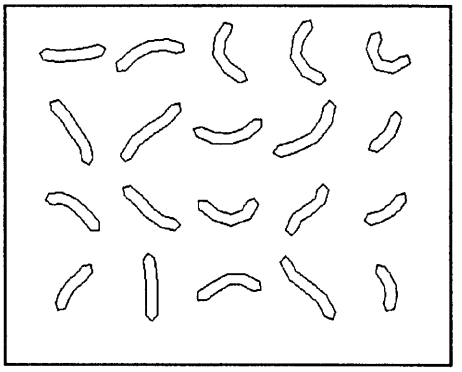


Fig 8: a family (can) of worms

In this case the local maxima in the maps correspond closely to the polygon vertices. There is one maximum for each polygon vertex. Note also the improvement in the rate of convergence using the revised scheme – the maxima are much better localised.

2.2 Flexible Objects

We wish to demonstrate our claim that we can locate objects whose expected shape is allowed to vary. To this end we have generated a set of axially symmetric ribbons (worms) whose axes can bend and be of different lengths but whose widths are fixed. Twenty examples taken from this set are shown diagrammatically in Fig. 8 to indicate the kind of variation present. The CLD in this experiment uses 12 points, one at each end of the worm and 5 pairs equally spaced along its length.

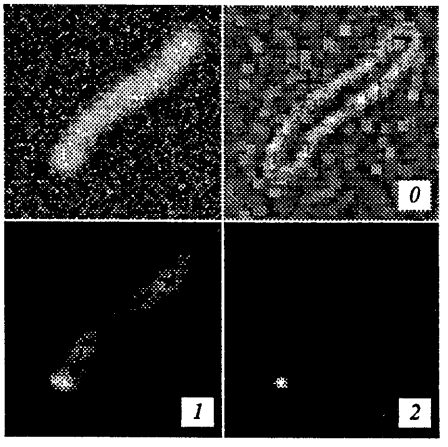


Fig 9(a): noisy worm + $P(x_1)$ after 0,1,2 iterations

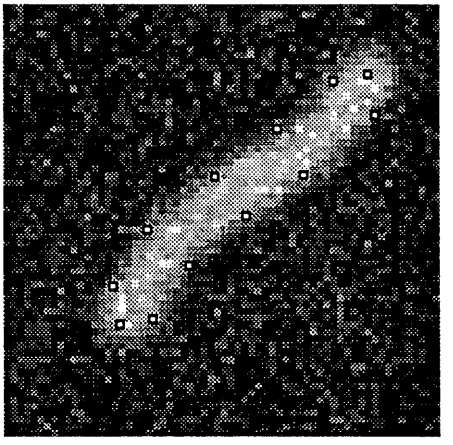


Fig 9(b): noisy worm with located points superimposed

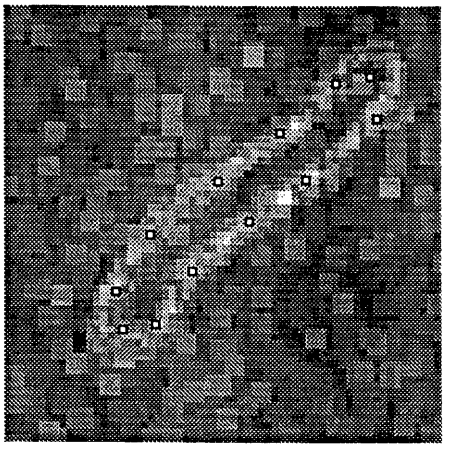


Fig 9(c): initial map data with located points

Fig. 9(a) shows a typical worm with 20% noise added. Also shown are the states of the maps for point x_1 after 0,1 and 2 iterations of the updating scheme. Fig. 9(b) shows

the worm at a larger scale. The 12 located points are superimposed. Fig. 9(c) also shows the initial state of the maps $P(x_i)$, which were generated via a morphological edge operator [10]. It is an indication of the power of the method that the points have been located so well considering that no prior integration of the obviously poor edge data has been made.

2.3 Overlapping Objects

We have investigated the behaviour of the revised scheme by applying the model to images of overlapping worms. Fig. 10(a) shows an example in which 4 possibilities arise for the position of x_1. Figs. 10(b) and 10(c) show 2 solutions obtained by selecting the south and west candidate positions for x_1 and continuing the iterations. The other 2 solutions are similar and differ only in the labelling of the points.

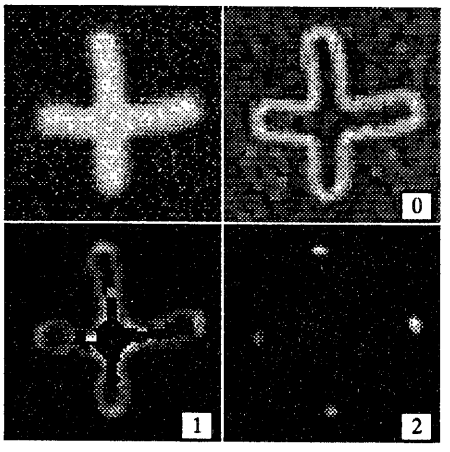


Fig 10(a): crossed worms and maps for point x_1 after 0,1,2 iterations

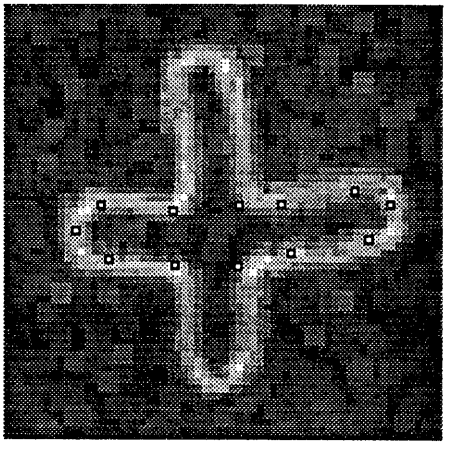


Fig 10(b): first solution

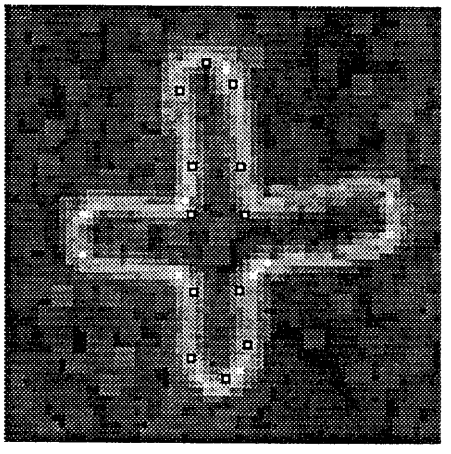


Fig 10(c): second solution

Fig. 11 shows a second case where one of the two solutions fails to include one of the extreme ends of the worm. Probable causes for this behaviour are that (a) the true

distributions for the model parameters are nearer to uniform over an interval than normal as assumed by the model, and so the predictions are weighted against examples at the edges of the distribution as is the case here, and (b) the object edge directions are corrupted in regions of overlap, giving rise to misleading predictions. We have yet to verify whether either of these possibilities is responsible.

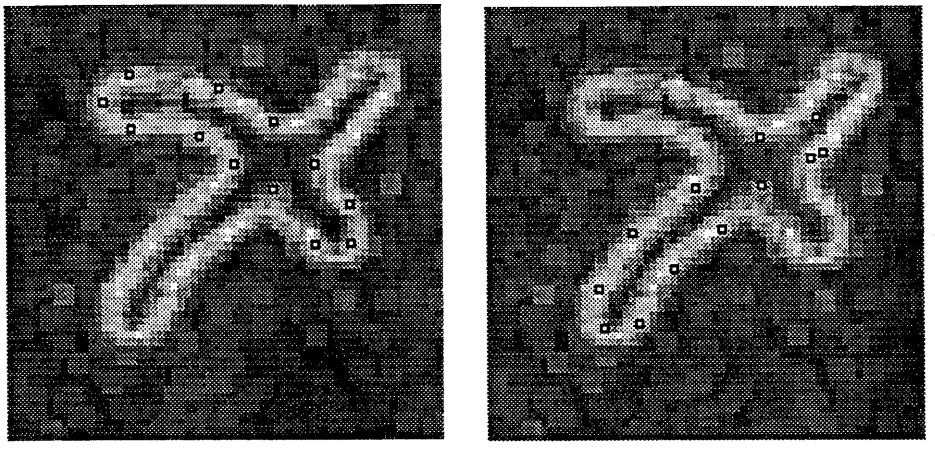


Fig. 11: a second example showing a failure to locate an extreme end in one case

3 DISCUSSION

As the figures above show, the revised CLD representation shows encouraging behaviour in locating objects whose shape is difficult to model explicitly, both in the presence of noise and clutter. Some further work is required to evaluate robustness when occlusion is present, but the results are promising. The method copes with both rigid and flexible objects. As expected, convergence is faster for rigid objects because of the more constraining predictions. Experiments (not described here due to lack of space) indicate that location performance increases with the number N of points in the model, and that digitisation errors can occur if the inter–point distances r_{ij} are small, typically 5 pixels or less. These factors limit the size of the smallest object that can be located.

The main drawback of the method is that it is slow – $O(N^2a^2b^2)$ where N is the number of points in the model, a is the typical prediction mask size and b is the region of interest size in pixels. On a SUN3/160 the 12–point model above on a 64 * 64 region typically takes hours per iteration (66 convolution–type predictions). The predictions can be expensive because the mask size is determined by the size of the object and can be large. Although we can propose a number of ad–hoc tricks to reduce this complexity, we are unlikely even on a modern workstation to achieve execution times which are practical for a working inspection system.

Despite this we can fruitfully apply the technique to multiple objects. By choosing object–defining points or derived reference points (the $\mathbf{x}_0$ in Fig. 1) for several

objects in a scene, we can capture inter-object spatial relationships using a CLD and exploit the arrangement in a top-down (predictive) way to limit search regions. The ability of the CLD to capture variability is being investigated mathematically with a view to applications using other search techniques which use shape generation. Early work in this direction is described in a companion paper [11] submitted to BMVC91.

4 ACKNOWLEDGEMENT

This work has been funded by DTI/SERC as project ref. IED3/1/2114 "VISAGE: Visual Inspection System Application Generation Environment".

5 REFERENCES

[1] **Taylor, C.J., Cooper, D. H.,** *Shape Verification Using Belief Updating* Proceedings: British Machine Vision Conference BMVC90 (Oxford); pp 61–66, 1990.

[2] **Mokhtarian, F., Mackworth, A.** *Scale–based description and recognition of planar curves and two dimensional shapes.* IEEE PAMI Vol. 8 p 34–43 ,1986.

[3] **Brady, M.** *Criteria for Representations of Shape* Human and Machine Vision. Academic Press , 1983.

[4] **Pearl, J.** *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufman (Publishers) ,1988.

[5] **Bolles, R. C.** *Robust Feature Matching Through Maximal Cliques* SPIE , Bellingham, Wash. Vol 182 pp 140–149. , 1979.

[6] **Chin,R.T.,Dyer, C.R.** *Model–Based Recognition in Robot Vision* Computing Surveys Vol 18 No 1 , 1986.

[7] **Grimson, W.E.L., Lonzano–Perez, T.** *Model–Based Recognition and Localisation from Sparse Range or Tactile Data* Int. J. Robotics Research Vol 3 No 3 pp 3–35 , 1984.

[8] **Grimson, W.E.L., Lonzano–Perez, T.** *Localising Overlapping Parts by Searching the Interpretation Tree* IEEE PAMI Vol 9 No 4 pp 469–482 , 1987.

[9] **Ballard, D., Brown, C.** *Computer Vision.* Prentice Hall , 1982.

[10] **Maragos, P.** *Tutorial on Advances in Morphological Image Processing and Analysis* Optical Engineering Vol 28 No 7 pp 623–632 , 1987.

[11] **Cootes, T.F., Cooper, D. H., Taylor, C.J., Graham, J.** *A Trainable Method of Parametric Shape Description* Proceedings: British Machine Vision Conference BMVC91 (Glasgow) , 1991.

22. **Trainable method of parametric shape description.** T.F. Cootes, D.H. Cooper, C.J. Taylor and J. Graham, *Image and Vision Computing 10: 289-294, 1992.* doi:10.1016/0262-8856(92)90044-4

Trainable method of parametric shape description

T F Cootes, D H Cooper, C J Taylor and J Graham

We have developed a trainable method of shape representation which can automatically capture the invariant properties of a class of shapes, and provide a compact parametric description of variability. We have applied the method to a family of flexible ribbons (worms), and to heart shapes in echocardiograms. We show that in both cases a natural parameterization of shape results.

Keywords: flexible shape models, deformable templates

Shape models have been used widely to achieve robust interpretation of complex images. They allow image evidence to be organized into plausible interpretations which can then be verified[1–3]. We are interested in the class of problems where shapes are variable. Important examples are the inspection of complex manufactured assemblies where relative motion between subparts is possible, and medical image interpretation where biological variation is present. In such applications it is generally the case that some aspects of shape are invariant whilst others are subject to constrained variability. The problem of adequately modelling such behaviour in a general way has not been solved. We have previously described a method of shape representation based on modelling the statistical distributions of chord lengths between control points placed in a consistent manner on each shape in a training set[12,13]. The objectives of the work we describe here were to significantly develop this basic idea to automatically:

1. Make shape invariants more explicit.
2. Identify and parameterize the significant degrees of freedom in a set of training shapes.

Our principal motivation was the wish to develop efficient methods of image interpretation based on flexible template matching. This requires generalization from a training set, to generate plausible instances of shapes which satisfy the constraints exhibited in the training set, yet can be controlled using a small number of parameters. The basic approach is as follows:

1. Gather chord statistics from the training set.
2. Calculate invariant and covariant sets of chords.
3. Generate new sets of chords from mean chords + weighted sums of covariant sets (varying the weights varies the form of the shape reconstructed).
4. Reconstruct shape from new set of chords.

Shapes are generated by varying the lengths of chords around their mean values in such a way that the shape invariants are maintained. This is achieved by deriving, from the training set, a form of relationship between the shape controlling parameters and the chord lengths which is guaranteed (to a first approximation) not to modify the invariant properties of the chord set. A shape is reconstructed by finding the positions of control points which are most consistent with the given set of chord lengths.

Many people use flexible models or deformable templates to aid the interpretation of images. Yuille *et al.*[4] and Lipson *et al.*[5] use models built by hand from various subparts. Unfortunately, these have to be individually tailored for each application. Kass *et al.*[6] described 'active contour models', flexible snakes which can stretch and deform to fit an image feature. These have been extended to apply constraints to their deformation[7,8], but not necessarily ones learnt from an example set. Pentland and Sclaroff[9] model objects as lumps of elastic clay, generating different shapes using combinations of the modes of vibration of the clay. However, this does not always lead to a very compact description of the sorts of deformation which can occur to a particular object. Bookstein[10,11] has studied the statistics of shape deformation by representing the objects as a set of 'landmark points', but does not use this information to generate flexible models with small numbers of parameters.

MODELLING SHAPE

Taylor and Cooper[12] describe a method of shape representation called the Chord Length Distribution

Department of Medical Biophysics, University of Manchester, Oxford Road, Manchester M13 9PT, UK

Paper received: 10 October 1991; revised paper received: 16 January 1992

0262–8856/92/005289–06 © 1992 Butterworth-Heinemann Ltd

(CLD). The CLD is trained on a set of s shapes which are represented by a set of n-vertex polygons. For each polygon in the training set the $m = \frac{1}{2}n(n-1)$ chord lengths R_i between all pairs of points are calculated, giving a m-vector $\underline{R} = \{R_i\}$.

A set of chord lengths is said to be 'Euclidean' in two dimensions if it can be generated from a real set of points, the vertices of an n-gon[14]. The lengths of many chords are inter-related, so one cannot adjust them arbitrarily and retain a Euclidean set. It is possible to estimate the correlation between pairs of chords by calculating their covariance over the training set.

The covariance (C_{ij}) between pairs (i, j) of chords is given by:

$$C_{ij} = \frac{1}{s} \sum_{t=1}^{s} (R_i^{(t)} - \mu_i)(R_j^{(t)} - \mu_j) \quad (1)$$

where s is the number of shapes in the training set, $R_i^{(t)}$ is the ith chord length of the tth shape in the training set, and μ_i is the mean length of R_i.

This gives a $m \times m$ covariance matrix $\boldsymbol{C} = \{C_{kl}\}$ for a given training set. We can find the (m) normalized eigenvectors $\underline{r}_k$ of $\underline{\boldsymbol{C}}$ such that:

$$\underline{\boldsymbol{C}}\underline{r}_k = \lambda_k \underline{r}_k \quad (\underline{r}_k^T \underline{r}_k = 1)$$

$$(\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_m) \quad (2)$$

These eigenvectors are combinations of variations in chord lengths which are linearly independent. If we make the assumption that the dependencies between chord lengths are linear, the eigenvectors may be treated as a totally independent parameterization of shape variability. The eigenvectors corresponding to large eigenvalues represent significant degrees of freedom in the family of shapes from which statistics have been obtained, whilst the vectors with small (or zero) eigenvalues represent shape invariants. Although the assumption of linear dependence does not strictly hold in the system we describe, it is a reasonable approximation, particularly where shape variability is modest.

The eigenvectors are mutually orthgonal and span the m-dimensional chord space. if $\underline{\boldsymbol{r}}$ is the $(m \times m)$ matrix of eigenvectors $\underline{\boldsymbol{r}} = (\underline{r}_1 | \underline{r}_2 | \underline{r}_3 | \ldots | \underline{r}_m)$, any set of chords R can be written:

$$\underline{R} = \underline{\mu} + \underline{\boldsymbol{r}}\underline{b} \quad (3)$$

where b is a $(m \times 1)$ column vector $(b_1\ b_2\ b_3 \ldots b_m)^T$:

$$\underline{b} = \underline{\boldsymbol{r}}^T(\underline{R} - \underline{\mu}) \quad (4)$$

(since the columns of $\underline{\boldsymbol{r}}$ are orthogonal, $\underline{\boldsymbol{r}}^T = \underline{\boldsymbol{r}}^{-1}$).

Each shape in the training set can thus be represented by a vector $\underline{b}$ in a new m-dimensional space. It can be shown that in this space the variance of the parmameter b_k over the set of training shape $\underline{b}$-vectors is λ_k, the kth eigenvalue of the covariance matrix $\underline{\boldsymbol{C}}$[15]. Thus the vector of chord deviations r_k explains λ_k of the variance of the chords in the training set.

We wish to choose a reduced set of chord vectors which can explain most of the variation in shape. This will allow us to generate shapes similar to those in the training set by varying only a small number of parameters. The variance explained by the first t eigenvectors is:

$$V_t = \sum_{k=1}^{t} \lambda_k \quad (5)$$

If t is chosen such that V_t is a suitably large proportion of V_m, the total variance the first t eigenvectors will then be able to explain most of the variability in the training set. Almost Euclidean sets of chords ($\underline{R}$) can be generated by taking the mean chord lengths and adding weighted combinations of the first t eigenvectors corresponding to the large eigenvalues:

$$\underline{R} = \underline{\mu} + \sum_{k=1}^{t} b_k \underline{r}_k$$

$$R = \underline{\mu} + \boldsymbol{r}' b' \quad (6)$$

where $\underline{\boldsymbol{r}}'$ is the $(m \times t)$ matrix of eigenvectors; $\boldsymbol{r}' = (r_1 | r_2 | \ldots | r_t)$, b' is a $(t \times 1)$ (column) vector of parameters b_k.

If a shape is reconstructed from the new set of chords $\underline{R}$, the parameters $b_k (k = 1 \ldots, t)$ will control the variations in the shape.

GENERATING POLYGONS FROM CHORD SETS

Because the above method uses a linear approximation to the possibly non-linear relationship between chord lengths, and is statistical in nature, the sets of chords produced as $\underline{b}'$ is varied may not be Euclidean in 2D (will not precisely correspond to a polygon). We define the polygon which best fits a set of chords ($\underline{R}_0$) constructed from a set of parameters (b_0) as the set of points $x = \{(x_i, y_i)\} (i = 0 \ldots, n-1)$ which minimizes the weighted sum:

$$F(x) = \sum_{i=1}^{m} w_i (R_i - R_{i,0})^2 \quad (7)$$

where w_i is a weight:

$$w_i = 1/\sigma_i \quad \text{if} \quad \sigma_i > 0.01$$

$$w_i = 100 \quad \text{otherwise.}$$

(σ_i is the standard deviation of the length of the ith chord over the training set.)

This weighting scheme ensures that those chords which are invariant in the training set have the same length in the reconstructed shape. Such chords have small standard deviations σ_i, so will have large weights w_i encouraging $R_i = R_{i0}$.

Equation (7) is non-linear in the $2n$ variables $\{x_0, y_0, \ldots, x_{n-1}, y_{n-1}\}$, and an analytical solution for the minima may not exist. The minimization is currently achieved using a 'steepest decent' optimization method[16]. The surface defined by equation (7) can have many local minima, so it is necessary to start with a good first guess. We obtain this by finding the example in the training set which gives the lowest value of $F(x)$.

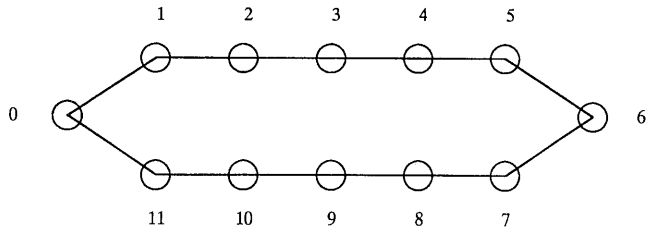


Figure 1. Labelling of 12 point 'worm' polygon used in training

Table 1. Eigenvalues of covariance matrix derived from 'worm' data set

Eigenvalue	*Value*	*% total sum of eigenvalues*
λ_1	959	75
λ_2	216	17
λ_3	44	3
λ_4	30	2
λ_5	25	2
λ_i $(i>5)$	<5	1

EXPERIMENTAL RESULTS

We have tested the method described above by applying it to two shape parameterization problems. The first involves a synthesized set of 'worm' shapes whose shape invariants and modes of variability are both known. The second relates to a practical application in medical image interpretation, and involves the outline of the left ventricle as seen in echocardiograms. In both cases, a training set of shapes has been used to automatically generate a parametric model with only a few degrees of freedom.

'Worms'

A set of 21 'worm' shapes, each described by 12 boundary points (see figures 1 and 2), was used to train the system. The worms varied in length and axial curvature but were of constant width. This example represents a challenging problem for the method; the important invariant is an axial symmetry which we wish the system to discover from the example shapes. At the same time, we wish the parametrization to allow significant variations in length and curvature.

The eigenvalues of the 66×66 covariance matrix $\boldsymbol{C}$ are shown in Table 1. These results suggest that the worms have two significant degrees of freedom, though a more comprehensive model could use the first five. The ranges of parameters b_1 and b_2 for the shapes in the training set (derived using equation (4)) are shown in Figure 3. The apparently random scatter suggests b_1 and b_2 can be varied independently to generate new shapes.

Examples of shapes generated by varying b_1 and b_2 are shown in Figure 4. The most significant eigenvector r_1 appears to affect the length of the worm. The next eigenvector r_2 affects the degree of overall curvature in the worm. In both cases, axial symmetry is preserved. The third parameter determines the amount to which the ends of the worms curve in opposite directions, but its effects are small as there were few examples of this in the training set.

Heart data set

A project currently being undertaken by Hill and Taylor[17] involves finding the boundary of the left ventricle in echocardiograms, ultrasound images of the heart. A 'hand crafted' parameterized model of the boundary is matched to the image using a genetic algorithm. We have investigated the possibility of constructing the model automatically from a set of examples. A training set was generated by manually drawing the heart boundary on each 66 images (see Figure 5). Each boundary was represented by an 18-vertex polygon. Four control points were placed on

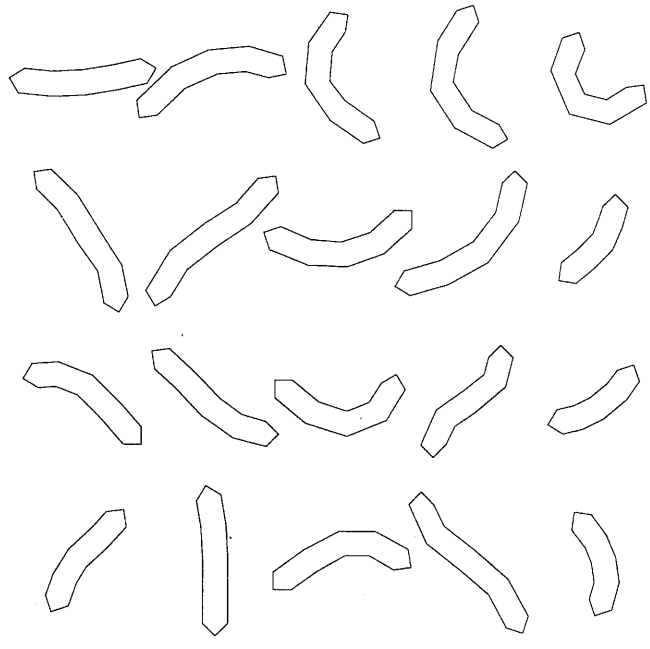


Figure 2. Examples of 12 point 'worms' training set

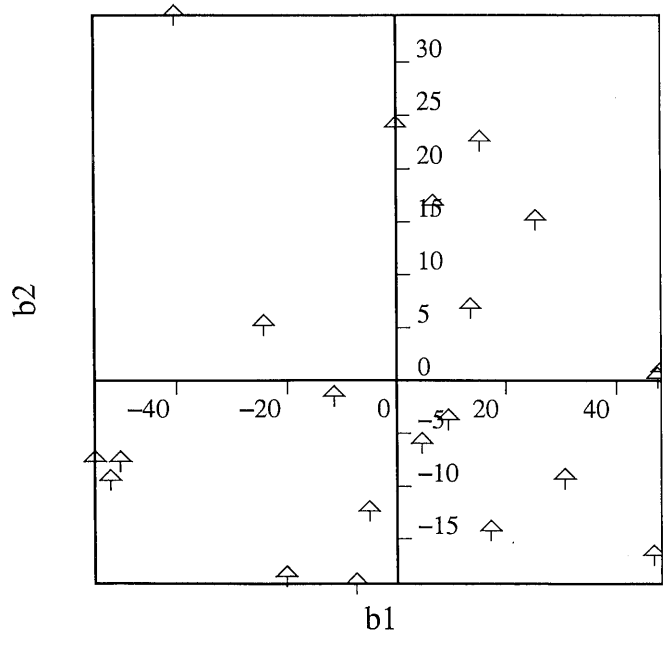


Figure 3. Scattergram of b_1 versus b_2 for shapes in the 'worm' training set

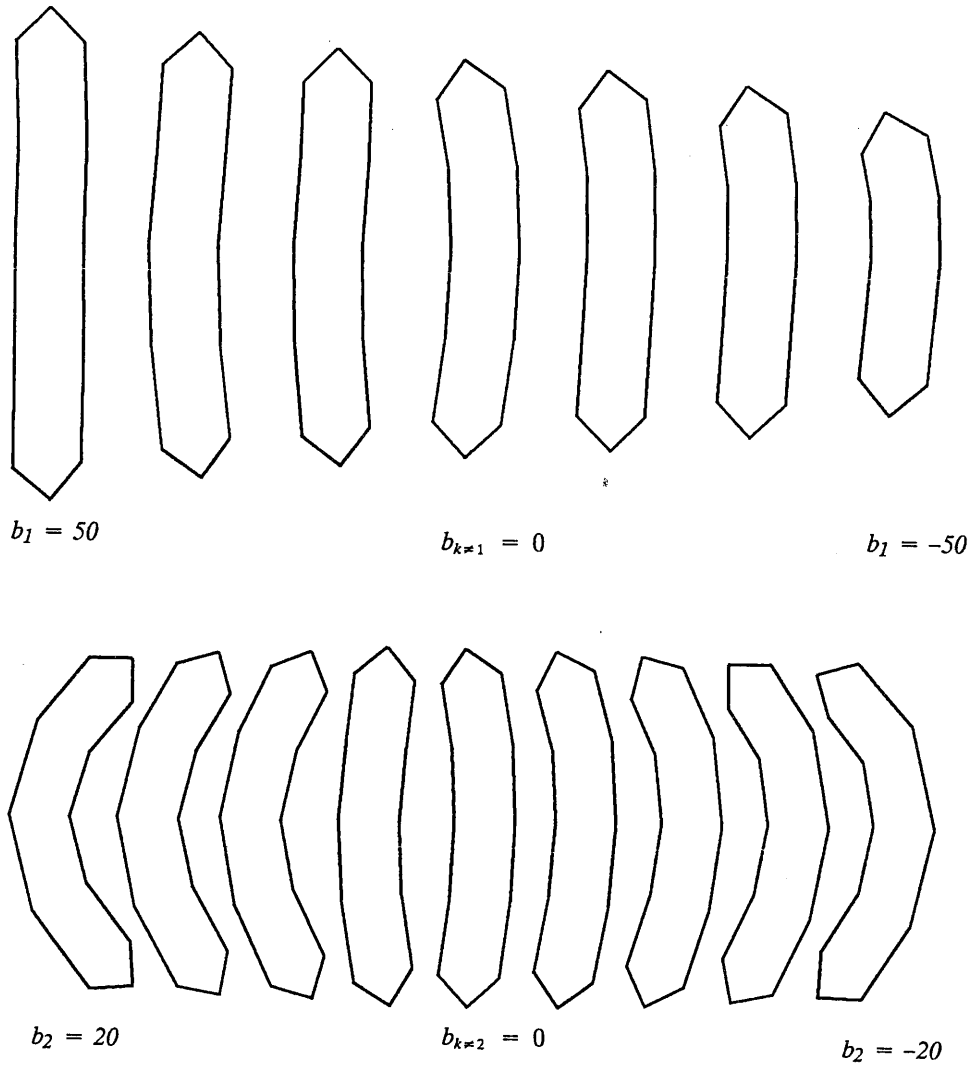


Figure 4. Effects of varying parameters corresponding to the largest two eigenvalues in the 'worm' model

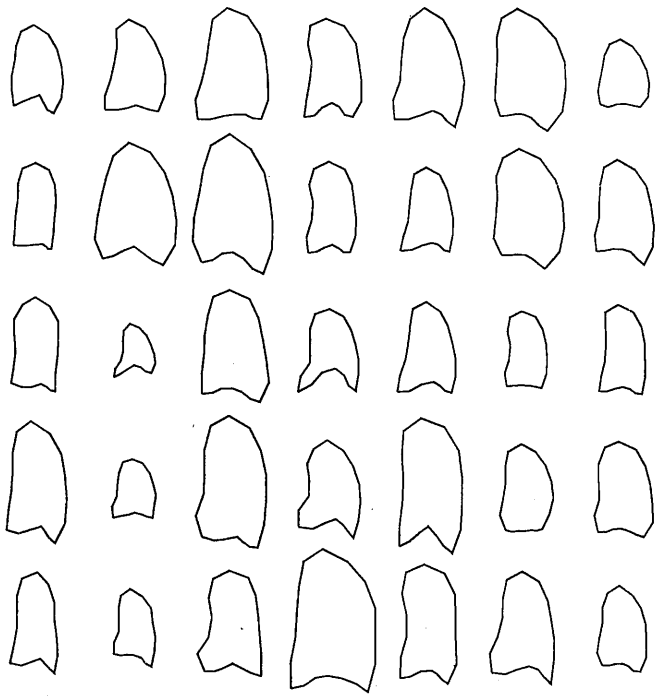


Figure 5. Examples from the training set of 18-vertex heart shapes

Table 2. Eigenvalues of covariance matrix derived from 'heart' data set

Eigenvalue	*Value*	*% total sum of eigenvalues*
λ_1	24080	91
λ_2	950	4
λ_3	465	2
λ_4	320	1
λ_5	210	1
λ_i $(i>5)$	<200	1

each boundary by hand, the 14 other points were equally spaced along the boundary between the control points.

The eigenvalues of the covariance matrix derived from the shapes in the training set are shown in Table 2. The ranges of parameters b_1 and b_2 for the shapes in the training set (derived using equation (4)) are shown in Figure 6.

The results of varying the first four prameters are shown in Figure 7. The parameter associated with the largest eigenvalue, b_1, controls the scale of the shape. The second parameter, b_2, seems to affect the width of the shape. The third and fourth parameters seem to affect the width and the form of the base of the shape,

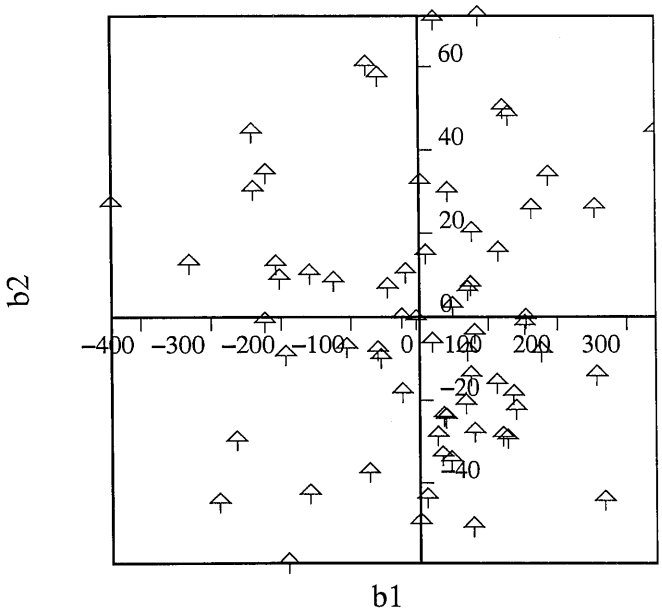


Figure 6. Scattergram of b1 versus b2 *for heart shapes in the training set*

which corresponds to the opening and closing of the mitral valve in the heart. Further parameters have more subtle effects.

DISCUSSION

The method has been applied to two cases and in both gives a parameterized model of shape. The resulting models have both been used to find boundaries in noisy images.

For some types of variability (e.g. rotation of one subpart around another) the assumption of linear dependence between chords does not hold. Consider the pair of rectangles rotating around one another shown in Figure 8. Although there is only one degree of freedom, one obtains two non-sero eigenvectors suggesting two parameters. However, a plot of b_1 against b_2 (see Figure 9) shows that they are not independent. In a more complicated example, the relationship may not be so clear, but would cause problems when parameters are chosen.*

In a case where the parameters are not independent, the iterative optimization technique can be though of as finding the point in parameter space which corresonds to a Euclidean set of chords (one that can be exactly reconstructed) which is closest to the point defined by the desired parameters.

To choose a suitable number of modes of variation for a model one can find how many modes are required to explain a large proportion (perhaps 99%) of the variance in the chord lengths. Alternatively, the effects of each mode in turn could be examined, and only

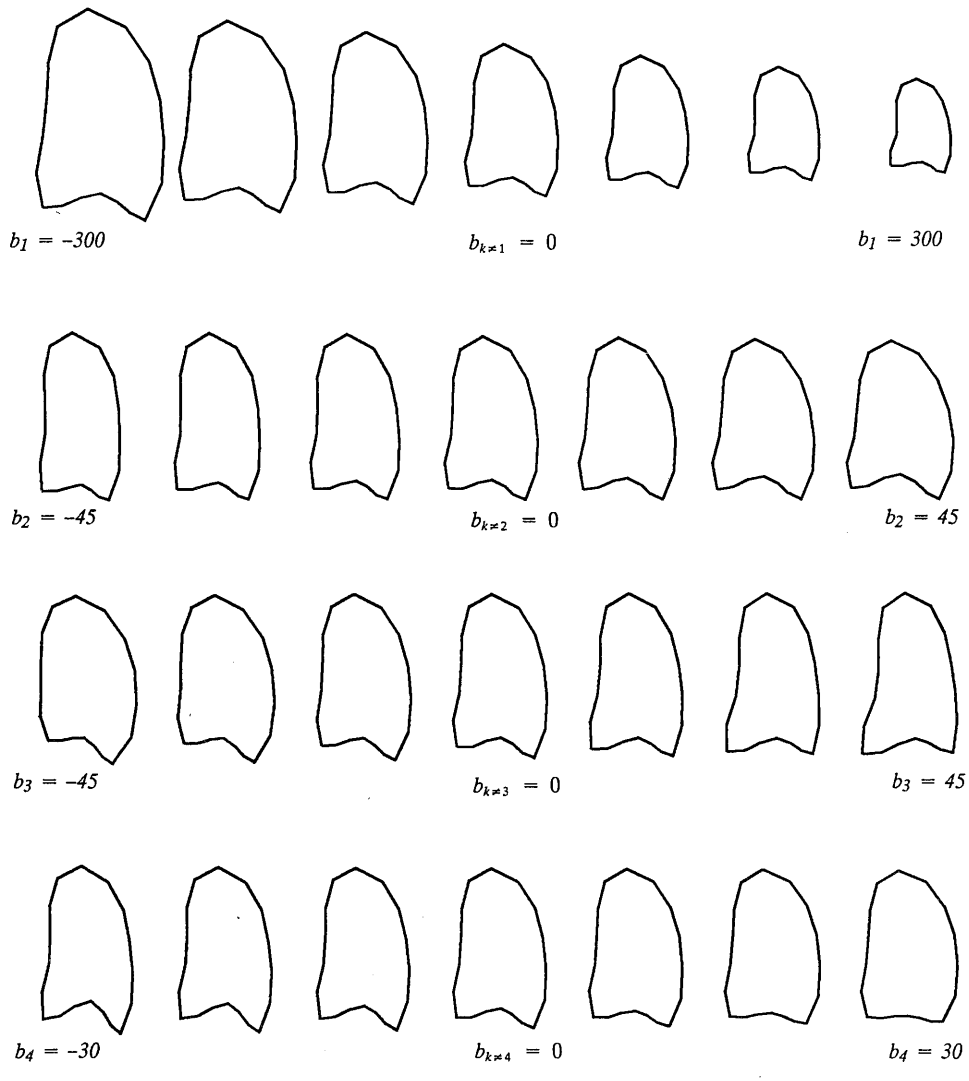


Figure 7. Effect of varying parameters associated with the four largest eigenvalues in the heart model

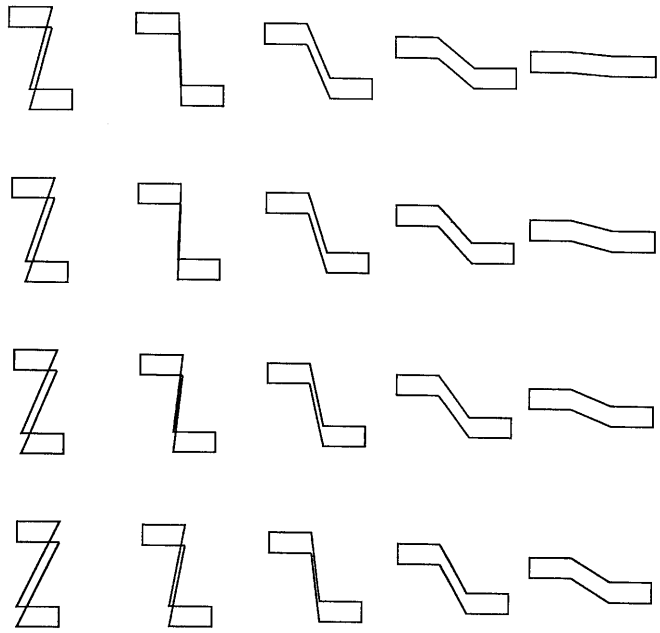


Figure 8. Examples from set of training shapes derived from a pair of rectangles

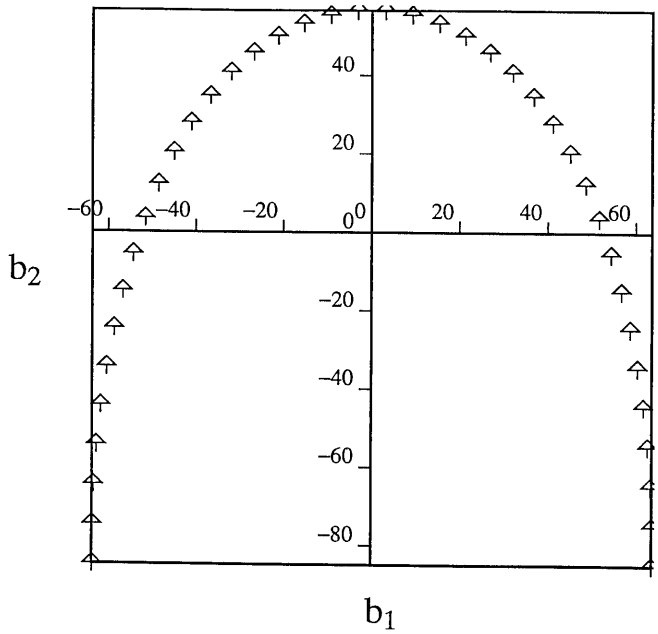


Figure 9. Plot of b_1 versus b_2 for set of rectangles

those which moved parts of the shape more than a chosen distance would be included in the model.

Although the examples given above are of shapes, it is the position of the points which are modelled. These can represent vertices of polygons or the positions of subparts equally easily, allowing spatial relationships between parts to be modelled.

The method could be used to model the image variability due to small changes in viewpoint and lighting conditions in industrial inspection.

The choice of points in the training set is important. Each point must be in a position which can be reproduced in each training shape. Choosing points can be a way of using human expertise in the training phase, though it would be useful to automate the procedure of positioning the points on training images as much as possible.

ACKNOWLEDGEMENTS

This project is funded by SERC, under the IEATP initiative (Project No. 3/2114). The authors would like to thank Andrew Hill and David Bailes for their help in preparing the training set for the heart model, and the reviewers for their suggestions.

REFERENCES

1 **Chin, R and Dyer C R** 'Model-based recognition in robot vision', *Comput Surv.*, Vol 18 No 1 (1986)

2 **Grimson, W E L** *Object Recognition by Computer: The Role of Geometric Constraints*, MIT Press, Cambridge, MA (1990)

3 **Cooper, D H, Bryson, N and Taylor, C J** 'An object location strategy using shape and grey-level models', *Image & Vision Comput.*, Vol 7 No 1 (1989) pp 50–56

4 **Yuille, A L, Cohen, D S and Hallinanmm, P** 'Feature extraction from faces using deformable templates', *Proc Comput. Vision & Patt. Recogn.*, San Diego, CA (1989) pp 104–109

5 **Lipson, P, Yuille, A L, O'Keeffe, D, Cavanaugh, J, Taaffe, J and Rosenthal, D** 'Deformable templates for feature extraction from medical images', *Proc. Euro. Conf. on Comput. Vision.*, Antibes, France (1990) pp 413–417

6 **Kass, M, Witkin, A and Terzopoulos, D** 'Snakes: active contour models', *Proc. 1st Int. Conf. on Comput. Vision.*, IEEE Press, NY (1987) pp 259–268

7 **Staib, L H and Duncan, J S** 'Parametrically deformable contour models', *Proc. Comput. Vision. & Patt. Recogn.*, San Diego, CA (1989) pp 98–103

8 **Terzopoulos, D and Metaxas, D** 'Dynamic 3D models with local and global deformations: deformable Superquadrics', *IEEE Trans. PAMI,* Vol 13 No 7 (1991) pp 703–714

9 **Pentland, A and Sclaroff, S** 'Closed-form solutions for physically based modelling and recognition', *IEEE Trans. PAMI,* Vol 13 No 7 (1991) pp 703–714

10 **Bookstein, F L** 'Principle warps: Thin-plate splines and the decomposition of deformations', *IEEE Trans. PAMI,* Vol 11 No 6 (1989) pp 567–585

11 **Bookstein, F L** *Morphometric Tools for Landmark Data,* Cambridge University Press, UK (1991)

12 **Taylor, C J and Cooper, D H** 'Shape verification using belief updating', *Proc. Br. Mach. Vision Conf.*, Oxford, UK (1990) pp 61–66

13 **Cooper, D H, Taylor C J, Graham, J and Cootes, T F** 'Locating overlapping flexible shapes using geometrical constraints', *Proc. 2nd Br. Mach. Vision Conf.*, Glasgow, UK (1991) pp 185–192

14 **Gower, J C** 'Euclidean distance geometry', *Math. Scientist,* Vol 7 (1982) pp 1–14

15 **Fukunaga, K and Koontz, W L G** 'Application of the Karhunen-loeve Expansion to feature selection and ordering', *IEEE Trans. Comput.*, Vol 19 No 4 (April 1970)

16 **Gill, P, Murray, W and Wright, M** *Practical Optimisation,* Academic Press, New York (1981)

17 **Hill, A and Taylor, C J** 'Model based image interpretation using genetic algorithms', *Proc. 2nd Br. Mach. Vision Conf.*, Glasgow, UK (1991)

23. **Training models of shape from sets of examples.** T.F. Cootes, C.J. Taylor , D.H. Cooper and J. Graham, *Proceedings of the British Machine Vision Conference, Leeds, 1992. Springer-Verlag. pp 9-18.* doi:10.5244/C.6.2

Training Models of Shape from Sets of Examples

T.F.Cootes, C.J.Taylor, D.H.Cooper and J.Graham

Department of Medical Biophysics
University of Manchester
Oxford Road
Manchester M13 9PT
email: bim@wiau.mb.man.ac.uk

Abstract

A method for building flexible shape models is presented in which a shape is represented by a set of labelled points. The technique determines the statistics of the points over a collection of example shapes. The mean positions of the points give an average shape and a number of modes of variation are determined describing the main ways in which the example shapes tend to deform from the average. In this way allowed variation in shape can be included in the model. The method produces a compact flexible 'Point Distribution Model' with a small number of linearly independent parameters, which can be used during image search. We demonstrate the application of the Point Distribution Model in describing two classes of shapes.

1 Introduction

We have previously described a method for modelling two dimensional shape, based on the statistics of chord lengths over a set of examples [12]. Although this provided a means of automatically parameterising shape variability, the method was difficult to use, requiring an iterative procedure to reconstruct a shape given a set of parameters. The method has computational complexity $O[n^2]$ where n is the number of points used to describe the shape. In this paper we present a new method which produces a more compact representation, allows direct reconstruction of a shape from a set of parameters and offers $O[n]$ computational complexity.

Image interpretation using rigid models is well established [1,2]. However, in many practical situations objects of the same class are not identical and rigid models are inappropriate. This is particularly true in medical applications, but also many industrial applications involve assemblies with moving parts, or components whose appearance can vary. In such cases flexible models, or deformable templates, can be used to allow for some degree of variability in the shape of the imaged object.

Yuille, Cohen and Hallinan [3] and Lipson *et al* [4] use deformable templates for image interpretation. Unfortunately their templates are hand-crafted with modes of variation which have to be individually tailored for each application. Kass, Witkin and Terzopoulos [5] described 'Active Contour Models', flexible snakes which can stretch and deform to image features. These have been extended to apply constraints to their deformation by adjusting the elasticity and stiffness of the model [6,7]. Pentland and Sclaroff [8] model objects as lumps of elastic clay, generating different shapes using combinations of the modes of vibration of the clay. However this does not always lead to a very compact description of the variability within a particu-

lar class of objects. Bookstein [9] has studied the statistics of shape deformation by representing objects as sets of 'landmark points', but has not applied this to the problem of shape modelling. Mardia, Kent and Walder [10] represent the boundary of a shape as a sequence of points with distributions related by a covariance matrix. To fit a model to an image they cycle through the points to find the most likely position given the image and the current shape. The examples given seem to be local models, in that deforming one part of the boundary does not affect the rest of it until the change has been propagated round the boundary by the updating method.

In this paper we describe a new method of shape modelling based on the statistics of labelled points placed on a set of training examples. The sets of points are automatically aligned so that their mean positions and main modes of variation can be calculated. Aligning the shapes allows the positions of equivalent points in different examples to be compared simply by examining their co–ordinates. A model consists of the mean positions of the points and a number of vectors describing the modes of variation.

2 Point Distribution Models

Suppose we wish to derive a model to represent the shape of resistors as they appear on a printed circuit board, such as those shown in Figure 1. Different examples of resistor have sufficiently different shapes that a rigid model would not be appropriate. Figure 2 shows some examples of resistor boundaries which were obtained from backlit images of individual resistors. Our aim is to build a model which describes both typical shape and allowed variability, using the examples in Figure 2 as a training set.

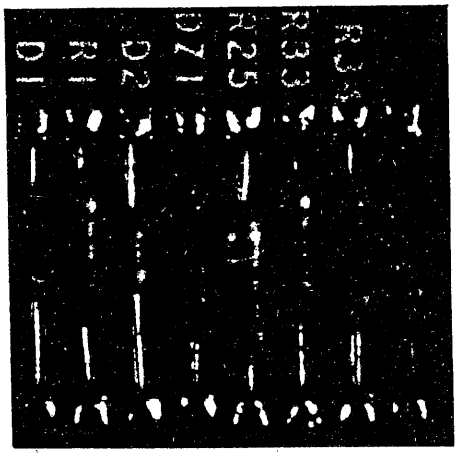


Figure 1 : Image of printed circuit board showing examples of resistors.

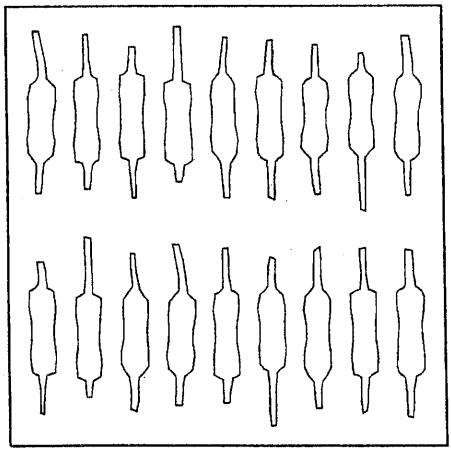


Figure 2 : Examples of resistor shapes from a training set.

2.1 Labelling The Training Set

In order to model a shape, we represent it by a set of points. For the resistors we have chosen to place points around the boundary, as shown in Figure 3. This must be done for each shape in the training set. The labelling of the points is important, each la-

belled point represents a particular part of the object or its boundary. For instance, in the resistor model, points 0 and 31 always represent the ends of a wire, points 3, 4 and 5 represent one end of the body of the resistor and so on. The method works by modelling how different labelled points tend to move together as the shape varies. If the labelling is incorrect, with a particular point placed at different sites on each training shape, the method will fail to capture shape variability.

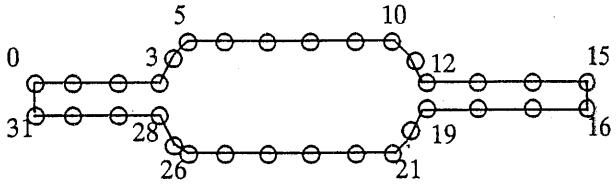


Figure 3 : 32 point model of the boundary of a resistor.

It is important that the points are placed correctly on each example image. This will usually require someone familiar with the application to choose the most appropriate set of points and to be able to reproducably place them on different examples. This procedure can be time consuming, though we are developing tools to speed up the process. It should be noted that though the labelling of the training set is done manually, finding the mean shape and main modes of variation is automatic. Deducing such a set of modes would be very difficult by hand, particularly for more complex biological shapes.

2.2 Aligning The Training Set

Our modelling method works by examining the statistics of the co–ordinates of the labelled points over the training set. In order to be able to compare equivalent points from different shapes, they must be aligned in the same way with respect to a set of axes. If they are not, we would not be comparing like with like and any statistics derived would be meaningless. We achieve the required alignment by scaling, rotating and translating the training shapes so that they correspond as closely as possible. We aim to minimise a weighted sum of squares of distances between equivalent points on different shapes. This is a form of Generalised Procrustes Analysis [11].

We will first consider aligning a pair of shapes. Let $\mathbf{x}_i$ be a vector describing the n points of the i^{th} shape in the set;

$$\mathbf{x}_i = (x_{i0},\ y_{i0},\ x_{i1},\ y_{i1},\ \ldots,\ x_{ik},\ y_{ik},\ \ldots,\ x_{in-1},\ y_{in-1})^T$$

Let $M_j[\mathbf{x}_j]$ be a rotation by θ_j and a scaling by s_j. Given two similar shapes, $\mathbf{x}_i$ and $\mathbf{x}_j$ we can choose θ_j , s_j and a translation $(t_x,t_y)_j$ mapping $\mathbf{x}_i$ onto $M_j[\mathbf{x}_j]$ so as to minimise the weighted sum

$$E_j = (\mathbf{x}_i - M_j(\mathbf{x}_j))^T \mathbf{W} (\mathbf{x}_i - M_j(\mathbf{x}_j)) \qquad (1)$$

where
$$M_j\begin{pmatrix} x_{jk} \\ y_{jk} \end{pmatrix} = \begin{pmatrix} (s_j \cos\theta)x_{jk} - (s_j \sin\theta)y_{jk} + t_{jx} \\ (s_j \sin\theta)x_{jk} + (s_j \cos\theta)y_{jk} + t_{jy} \end{pmatrix} \qquad (2)$$

and W is a diagonal matrix of weights for each point.

Details are given in Appendix A.

The weights can be chosen to give more significance to those points which tend to be most 'stable' over the set – the ones which move about least with respect to the other points in a shape. We have used a weight matrix defined as follows: let R_{kl} be the distance between points k and l in a shape; let $V_{R_{kl}}$ be the variance in this distance over the set of shapes; we can choose a weight, w_k, for the k^{th} point using

$$w_k = \left(\sum_{l=0}^{n-1} V_{R_{kl}} \right)^{-1} \tag{3}$$

If a point tends to move around a lot with respect to the other points in the shape, the sum of variances will be large, and a low weight will be given. If, however, a point tends to remain fixed with respect to the others, the sum of variances will be small, a large weight will be given and matching such points in different shapes will be a priority.

In order to align all the shapes in a set we use the following algorithm.

1) Rotate, scale and translate each of the shapes in the set to align to the first shape.

Repeat

2) Calculate the mean of the transformed shapes

3) Either

a) Adjust the mean to a default scale, orientation and origin,

b) Rotate, scale and translate the mean to align to the first shape

4) Rotate, scale and translate each of the shapes again to match to the adjusted mean.

Until convergence.

Stage 3 inside the iteration loop is required to renormalise the mean. Without this the algorithm is ill-conditioned – there are in effect $4(N_s-1)$ constraints on $4N_s$ variables (θ , s, t_x,t_y for each shape) – and will not converge – the mean will shrink, rotate or slide off to infinity. Constraints on the pose and scale of the mean allow the equations to have a unique solution. Either the mean is scaled, rotated and translated so it matches the first shape, or an arbitrary default setting can be used, such as choosing an origin at its centre of gravity, an orientation so that a particular part of the shape is at the top and a scale so that the distance between two points is one unit.

The convergence condition can be tested by examining the average difference between the transformations required to align each shape to the recalculated mean and the identity transformation. Experiments suggest that the method converges to the same result independent of which shape is aligned to in the first stage, though a formal proof of convergence has yet to be devised.

2.3 Capturing the Statistics of a Set of Aligned Shapes

Once a set of aligned shapes is available the mean shape and variability can be found. The mean shape, $\bar{\mathbf{x}}$, is calculated using

$$\bar{\mathbf{x}} = \frac{1}{N_s}\sum_{i=1}^{N_s}\mathbf{x}_i \tag{4}$$

The modes of variation, the ways in which the points of the shape tend to move together, can be found by applying principal component analysis to the deviations from the mean as follows.

For each shape in the training set we calculate its deviation from the mean, $d\mathbf{x}_i$, where

$$d\mathbf{x}_i = \mathbf{x}_i - \bar{\mathbf{x}} \tag{5}$$

We can then calculate the $2n$ x $2n$ covariance matrix, S, using

$$\mathbf{S} = \frac{1}{N_s}\sum_{i=1}^{N_s} d\mathbf{x}_i d\mathbf{x}_i^T \tag{6}$$

The modes of variation of the points of the shape are described by the unit eigenvectors of S, $\mathbf{p}_i$ *(i = 1 to 2n)* such that

$$\mathbf{S}\mathbf{p}_i = \lambda_i \mathbf{p}_i \tag{7}$$

(where λ_i is the i'th eigenvalue of S, $\lambda_i \geq \lambda_{i+1}$)

$$\mathbf{p}_i^T \mathbf{p}_i = 1 \tag{8}$$

It can be shown that the eigenvectors of the covariance matrix corresponding to the largest eigenvalues describe the most significant modes of variation in the variables used to derive the covariance matrix, and that the proportion of the total variance explained by each eigenvector is equal to the corresponding eigenvalue [13]. Most of the variation can usually be explained by a small number, t, modes. One method for calculating t would be to chose the smallest number of modes such that the sum of variance explained was a sufficiently large proportion of λ_T, the total variance of all the variables, where

$$\lambda_T = \sum_{i=1}^{2n} \lambda_i \tag{9}$$

The i'th eigenvector affects point k in the model by moving it along a vector parallel to (dx_{ik}, dy_{ik}), which is obtained from the k'th pair of elements in $\mathbf{p}_i$,

$$(dx_{i0},\ dy_{i0},\ \ldots\ , \underbrace{dx_{ik},\ dy_{ik}},\ \ldots\ ,\ dx_{in-1},\ dy_{in-1}\ ) \tag{10}$$

Any shape in the training set can be approximated using the mean shape and a weighted sum of these deviations obtained from the first t modes

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}\mathbf{b} \tag{11}$$

where $\mathbf{P} = (\mathbf{p}_1\ \mathbf{p}_2\ \ldots\ \mathbf{p}_t)$ is the matrix of the first t eigenvectors,

$\mathbf{b} = (b_1\ b_2\ \ldots\ b_t)^T$ is a vector of weights for each eigenvector

the eigenvectors are orthogonal, $\mathbf{P}^T\mathbf{P} = \mathbf{I}$ so

$$\mathbf{b} = \mathbf{P}^T(\mathbf{x} - \bar{\mathbf{x}}) \tag{12}$$

The above equations allow us to generate new examples of the shapes be varying the parameters (b_i) within suitable limits. The parameters are linearly independent, though there may be non–linear dependencies still present. The limits for b_i are derived by examining the distributions of the parameter values required to generate the training set. Since the variance of b_i over the training set can be shown to be λ_i, suitable limits are likely to be of the order of

$$-3\sqrt{\lambda_i} \leq b_i \leq 3\sqrt{\lambda_i} \tag{13}$$

since most of the population lies within three standard deviations of the mean.

3 Practical Examples

The techniques described above have been used to generate shape models for both manufactured and biological objects. We present results for the set of resistor shapes shown in Figure 2 and a set of hand shapes.

3.1 Resistor Example

The resistor shapes were aligned using the method described above, arranging the mean shape to be horizontal and scaling so the average distance of each point of the mean from its centre of gravity is one unit. The most significant eigenvalues of the covariance matrix derived are shown in Table 1.

Table 1 : Eigenvalues of the covariance matrix derived from a set of resistor shapes.

Eigenvalue	λ_i	$\frac{\lambda_i}{\lambda_T} \times 100\%$	$\sqrt{\lambda_i}$
λ_1	0.207	66%	0.46
λ_2	0.026	8%	0.16
λ_3	0.017	5%	0.13
λ_4	0.013	4%	0.11
λ_5	0.010	3%	0.10
λ_6	0.008	3%	0.09

Figure 4 shows the plot of b_1 against b_2 for the training set. The lack of structure in the scatter plot suggests that the parameters can be treated as independent. We are currently working on deriving more formal tests of independence. Any dependencies between the parameters would imply non–linear relationships between the original point positions and would results in some combinations of parameters generating 'illegal' shapes. By varying the first three parameters separately we can generate examples of the shape as shown in Figures 5–7. Each parameter 'represents' a mode of variation of the shape which can frequently be associated with an intuitive

description of the deformation. Compare Figures 5–7 with Figure 2. Varying the first parameter (b_1) adjusts the position of the body of the resistor up and down the wire. The second parameter varies the shape of the ends of the main body of the resistor, between tapered and square. The third parameter affects the curvature of the wires at either end. Subsequent parameters have smaller effects, including the wires bending in opposite directions. These modes of variation effectively capture the variability which was present in the training set.

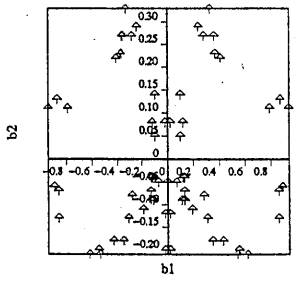
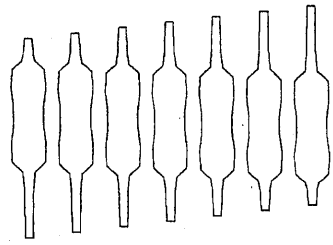
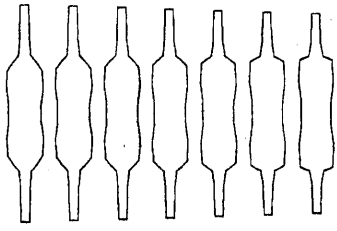


Figure 4 : Plot of b_1 vs b_2 for a training set of resistor shapes.



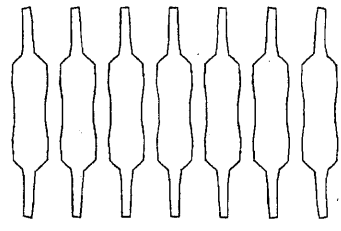
–0.9 ← b_1 → 0.9

Figure 5 : Effects of varying the first parameter of the resistor model.



–0.3 ← b_2 → 0.3

Figure 6 : Effects of varying the second parameter of the resistor model.



–0.25 ← b_3 → 0.25

Figure 7 : Effects of varying the third parameter of the resistor model.

3.2 Hand Example

A set of 18 hand shapes was generated from images of the right hand of one of the authors (Figure 8). Each was represented by 72 points around the boundary. These were planted on the examples by locating 12 control points at the ends and joints of the fingers and filling in the rest equally along the connecting boundaries. A model was trained on the data, and it was found that 96% of the variance could be explained by the first 6 modes of variation. The first three modes are shown in Figure 9, and consist of combinations of movements of the fingers. Again, a compact parameterised model has been generated.

4 Discussion and Conclusions

The method outlined above allows a compact, flexible shape model to be built, representing a class of shapes by the mean positions of a set of labelled points and a small number of modes of variation about the mean. The model points do not have to lie only on the boundary of objects, they can represent internal features, and even sub–components of a complex assembly. In the latter case the model describes both

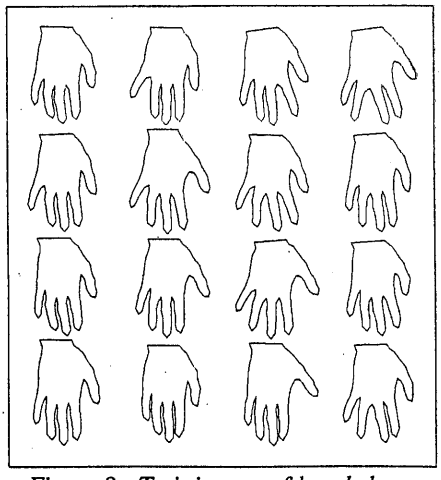


Figure 8 : Training set of hand shapes, each defined by 72 points.

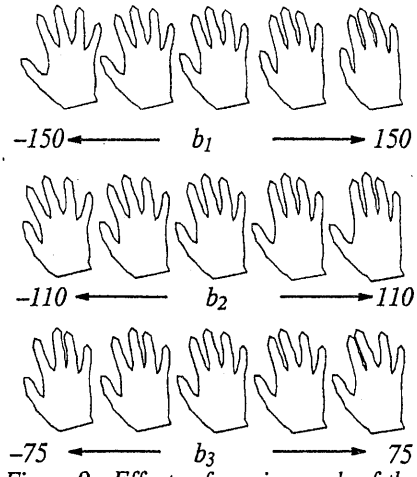


Figure 9 : Effects of varying each of the first three parameters of the hand model individually.

the variations in the shapes of the sub-components and the geometric relationships between components. Such a model, representing a section through the ventricles in the brain in MR scans is described by Hill *et al* in [16].

It is important to arrange that all the examples used to train the model are similarly aligned with respect to a set of axes, to ensure that the labelled points in different shapes are being compared correctly. In some cases an obvious alignment is apparent, but in others, particularly medical cases where the shapes of organs are very flexible, the automatic least squares alignment method is essential. The method has been used successfully to model a variety of objects from both industrial and biological domains.

The models we build are linear. Varying each parameter individually moves the points along straight lines. The method is inefficient at modelling non-linear effects such as bending or rotation of one sub-component about another. To deal with such cases a non-linear model of the modes of variation would be required. We have begun experimenting with a system which represents each mode using a polynomial curve rather than a straight line [14]. Some promising results have been produced which will be the subject of a further paper.

Point Distribution Models have been used in image search. A local optimiser called the Active Shape Model has been developed [15] which provides a way of iteratively improving an initial estimate of the position, pose and shape parameters of a model fitted to image data. The model has also been used in conjunction with a generate and test strategy based around Genetic Algorithms [16]. The hand and resistor models described above have been successfully used to find examples in images with both techniques.

The model can also be used in a classifier. Given an example of a shape, an estimate can be made of how likely that example is to be a member of the class of shapes described by a model. If labelled points are placed on the example and the point set aligned with the mean shape, Equation 12 can be used to calculate the model para-

meters required to generate the example. The distributions of the parameters can be estimated from the training set, allowing probabilities to be assigned. This technique has been successfully used in a simple handwritten character recognition application [17].

The models are compact and easy to use. Given a set of parameters an example of the model can be calculated rapidly. The models are well suited to generate-and-test image search strategies in many domains.

Acknowledgements

This work is funded by SERC under the IEATP Initiative (Project Number 3/2114). The authors would like to thank the other members of the Wolfson Image Analysis Unit for their help and advice, particularly D.Bailes and A.Hill.

Appendix : Aligning A Pair of Shapes

Given two similar shapes, x_1 and x_2 we would like to choose a rotation, θ, a scale s and a translation (t_x,t_y) mapping x_2 onto $M(x)$ so as to minimise the weighted sum

$$E = (\mathbf{x}_1 - M(\mathbf{x}_2))^T \mathbf{W} (\mathbf{x}_1 - M(\mathbf{x}_2)) \tag{1}$$

where

$$M\begin{pmatrix} x_{jk} \\ y_{jk} \end{pmatrix} = \begin{pmatrix} (s\cos\theta)x_{jk} - (s\sin\theta)x_{jk} + t_x \\ (s\sin\theta)x_{jk} + (s\cos\theta)x_{jk} + t_y \end{pmatrix} \tag{2}$$

and W is a diagonal matrix of weights for each point.

If we write

$$a_x = s\cos\theta \qquad a_y = s\sin\theta$$

then least squares approach (differentiating with respect to each of the variables a_x, a_y, t_x, t_y) leads to a set of four linear equations;

$$\begin{pmatrix} X_2 & -Y_2 & W & 0 \\ Y_2 & X_2 & 0 & W \\ Z & 0 & X_2 & Y_2 \\ 0 & Z & -Y_2 & X_2 \end{pmatrix} \begin{pmatrix} a_x \\ a_y \\ t_x \\ t_y \end{pmatrix} = \begin{pmatrix} X_1 \\ Y_1 \\ C_1 \\ C_2 \end{pmatrix} \tag{14}$$

where

$$X_i = \sum_{k=0}^{n-1} w_k x_{ik} \qquad Y_i = \sum_{k=0}^{n-1} w_k y_{ik} \tag{15}$$

$$Z = \sum_{k=0}^{n-1} w_k (x_{2k}^2 + y_{2k}^2) \qquad W = \sum_{k=0}^{n-1} w_k \tag{16}$$

$$C_1 = \sum_{k=0}^{n-1} w_k (x_{1k}x_{2k} + y_{1k}y_{2k}) \tag{17}$$

$$C_2 = \sum_{k=0}^{n-1} w_k (y_{1k}x_{2k} - x_{1k}y_{2k}) \tag{18}$$

These can be solved for a_x, a_y, t_x, and t_y using standard matrix methods.

References

[1] R. Chin and C.R. Dyer, Model-Based Recognition in Robot Vision. Computing Surveys 1986; Vol 18, No 1

[2] W.E.L. Grimson, Object Recognition by Computer : The Role of Geometric Constraints, The MIT Press, Cambridge, MA, USA, 1990.

[3] A.L. Yuille, D.S. Cohen and P. Hallinan, Feature extraction from faces using deformable templates, Proc. Computer Vision and Pattern Recognition (1989) pp104–109.

[4] P. Lipson, A.L. Yuille, D. O'Keeffe, J. Cavanaugh, J. Taaffe and D. Rosenthal, Deformable Templates for Feature Extraction from Medical Images, Proceedings of the First European Conference on Computer Vision (Lecture Notes in Computer Science, ed. O. Faugeras, pub. Springer–Verlag) 1990 pp413–417.

[5] M. Kass, A. Witkin and D. Terzopoulos, Snakes: Active Contour Models. First International Conference on Computer Vision, pub. IEEE Computer Society Press, 1987, pp 259–268.

[6] L.H. Staib and J.S. Duncan, Parametrically Deformable Contour Models. IEEE Computer Society conference on Computer Vision and Pattern Recognition, San Diego, 1989

[7] D. Terzopoulos and D. Metaxas, Dynamic 3D Models with Local and Global Deformations : Deformable Superquadrics. IEEE Trans. on Pattern Analysis and Machine Intelligence 1991; Vol.13 No.7 pp703–714.

[8] A. Pentland and S, Sclaroff, Closed–Form Solutions for Physically Based Modelling and Recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence 1991; Vol.13 No.7 pp703–714. 715 - 729

[9] F.L. Bookstein, Morphometric Tools for Landmark Data. Cambridge University Press, 1991.

[10] K.V. Mardia, J.T. Kent and A.N. Walder, Statistical Shape Models in Image Analysis. Proceedings of the 23rd Symposium on the Interface, Seattle 1991, pp 550–557.

[11] J.C. Gower, Generalized Procrustes Analysis. Psychometrika. 40, 1975, 33–51.

[12] T.F. Cootes, D. Cooper, C.J. Taylor and J. Graham, A Trainable Method of Parametric Shape Description. Proc. BMVC 1991 pub. Springer–Verlag, pp54–61.

[13] K. Fukunaga and W.L.G. Koontz, Application of the Karhunen–Loeve Expansion to Feature Selection and Ordering. IEEE Trans. on Computers 1970; 4.

[14] J. Graham, T.F. Cootes, D.Cooper and C.J. Taylor, VISAGE Progress Report – Deliverable D4, Wolfson Image Analysis Unit, Manchester University 1992.

[15] T.F. Cootes and C.J. Taylor, Active Shape Models – 'Smart Snakes'. This Volume.

[16] A. Hill, T.F. Cootes and C.J. Taylor, A Generic System for Image Interpretation Using Flexible Templates. This Volume.

[17] A. Lanitis, Optical Character Recognition of Hand–written Characters using Flexible Templates. Internal Report, Wolfson Image Analysis Unit, Manchester University 1992.

24. **Active Shape Models - Their training and application.** T.F. Cootes , D.H. Cooper, C.J. Taylor and J. Graham, *Computer Vision and Image Understanding 61: 38-59, 1995.* doi:10.1006/cviu.1995.1004

COMPUTER VISION AND IMAGE UNDERSTANDING
Vol. 61, No. 1, January, pp. 38–59, 1995

Active Shape Models—Their Training and Application

T. F. COOTES, C. J. TAYLOR, D. H. COOPER, AND J. GRAHAM*

Department of Medical Biophysics, University of Manchester, Oxford Road, Manchester M13 9PT, England

Received July 29, 1992; accepted April 12, 1994

Model-based vision is firmly established as a robust approach to recognizing and locating known rigid objects in the presence of noise, clutter, and occlusion. It is more problematic to apply model-based methods to images of objects whose appearance can vary, though a number of approaches based on the use of flexible templates have been proposed. The problem with existing methods is that they sacrifice model specificity in order to accommodate variability, thereby compromising robustness during image interpretation. We argue that a model should only be able to deform in ways characteristic of the class of objects it represents. We describe a method for building models by learning patterns of variability from a training set of correctly annotated images. These models can be used for image search in an iterative refinement algorithm analogous to that employed by Active Contour Models (Snakes). The key difference is that our *Active Shape Models* can only deform to fit the data in ways consistent with the training set. We show several practical examples where we have built such models and used them to locate partially occluded objects in noisy, cluttered images. © 1995 Academic Press, Inc.

1. INTRODUCTION

We address the problem of locating examples of known objects in images. Image interpretation using rigid models is well established [1, 2]. However, in many practical situations objects of the same class are not identical and rigid models are inappropriate. In medical applications, for instance, the shape of organs can vary considerably through time and between individuals. In addition, many industrial applications involve assemblies with moving parts, or components whose appearance can vary. In such cases flexible models, or deformable templates, can be used to allow for some degree of variability in the shape of the imaged objects [3–23].

In this paper we present new methods of building and using flexible models of image structures whose shape can vary. The models are able to capture the natural variability within a class of shapes and can be used in image search to find examples of the structures that they represent. Previous approaches have allowed models to deform, but have not tailored the variability to the class of shapes concerned—the models are not specific. Our main contribution is to describe how to create models which allow for considerable variability but are still specific to the class of structures they represent.

Our technique relies upon each object or image structure being represented by a set of points. The points can represent the boundary, internal features, or even external ones, such as the center of a concave section of boundary. Points are placed in the same way on each of a training set of examples of the object. This is done manually, though tools are available to aid the user. The sets of points are aligned automatically to minimize the variance in distance between equivalent points. By examining the statistics of the positions of the labeled points a "Point Distribution Model" is derived. The model gives the average positions of the points, and has a number of parameters which control the main modes of variation found in the training set.

Given such a model and an image containing an example of the object modeled, image interpretation involves choosing values for each of the parameters so as to find the best fit of the model to the image. We describe a technique which allows an initial very rough guess for the best shape, orientation, scale, and position to be refined by comparing the hypothesized model instance with image data, and using differences between model and image to deform the shape. We have previously described how to obtain the initial guess [7]. The method has similarities with the Active Contour Models (or snakes) of Kass *et al.* [3], but differs in that global shape constraints are applied; to make this distinction clear we have adopted the term Active *Shape* Models. The key point is that instances of the models can only deform in ways found in the training set.

Our results demonstrate that the method for constructing models combined with the active matching technique provides a systematic and effective paradigm for the interpretation of complex images. In the remainder of the paper we review some of the relevant literature, describe the modeling method, and show examples of

* E-mail: bim@uk.ac.man.mb.wiau. Fax: 061 275 5145.

1077-3142/95 $6.00
Copyright © 1995 by Academic Press, Inc.
All rights of reproduction in any form reserved.

trained models. The active matching technique is described and results are given, showing how the models can be used to interpret images.

2. BACKGROUND

There is a substantial literature describing the use of flexible models or deformable templates to aid image interpretation. Such models usually have a number of parameters to control the shape and pose of all or parts of the model. We give a brief review of some of the most significant work, which relates mainly to two-dimensional images.

2.1. *"Hand Crafted" Models*

Flexible models can be built up from simple subcomponents, such as circles, lines, or arcs, which are allowed some degree of freedom to move around relative to one another, and possibly change scale and orientation. Yuille *et al.* [5] model parts of the face, such as the eyes and mouth, in this way. When attempting to fit a model to an image they first obtain an approximate fit, which they refine by changing different parts of the model, one at a time. Lipson *et al.* [6] apply a similar scheme to map ellipitical models of vertebrae onto CT images of the spine. Hill *et al.* [7] use a handcrafted model of the heart in combination with Genetic Algorithm search to find the left ventricle in echocardiograms.

Although such models can capture detailed knowledge of expected shapes, the approach lacks generality. It is necessary to design both a new model and a scheme for fitting to images for each application.

2.2. *Articulated Models*

A number of authors consider articulated models built from rigid components connected by sliding or rotating joints. Beinglass and Wolfson [8] describe a scheme for locating such objects using a Generalized Hough Transform with the point of articulation as the reference point for each subpart. Connected subparts then vote for the same reference point. Grimson [2] has extended his "interpretation tree" approach to object recognition to include some articulations, and reviews other work along the same lines. This approach is only applicable to a restricted class of variable shape problems.

2.3. *Active Contour Models ("Snakes")*

Kass *et al.* [3] describe flexible contour models which are attracted to image features. These energy minimizing spline curves are modeled as having stiffness and elasticity and are attracted toward features such as lines and edges. Constraints can be applied to ensure that they remain smooth and to limit the degree to which they can be bent.

Snakes can be considered as parameterized models, the parameters being the spline control points. They are usually free to take almost any smooth boundary with few constraints on their overall shapes. The idea of fitting by using image evidence to apply forces to the model and minimizing an energy function is effective.

Hinton *et al.* [4] describe a type of spline snake governed by a number of control points which have preferred "home" locations to give the snake a particular default shape. Deformations are caused by moving the control points away from their "home" locations. Although the average shape of an object is represented, the modes of shape variation are only coarsely defined by the number and position of control points.

2.4. *Fourier Series Shape Models*

Scott [9] proposes a method of modeling shapes by an expansion of trigonometric functions,

$$\begin{aligned} x &= x_0 + \sum_n a_n \sin(n\theta + \phi_n) \\ y &= y_0 + \sum_n b_n \sin(n\theta + \psi_n). \end{aligned} \tag{1}$$

The shape produced is a function of the parameters a_n, b_n, ϕ_n, ψ_n. By varying the parameters and the number of terms used, different shapes can be generated. Scott shows how to fit such a shape model to image data by varying the parameters so as to minimize an energy term. The model is almost infinitely deformable, and contains no prior shape information. Staib and Duncan [10] describe similar Fourier models, and use them to interpret medical images. They derive distributions for each of the parameters over a training set and while fitting the model to an image maximize a probability measure determining how likely it is that the current example is the desired object. Bozma and Duncan [11] describe how such a technique can be used to model organs in medical images. A given shape is represented by a list of values for the parameters and is deformed by varying the parameters from these values. They describe ways of incorporating relationships between several flexible objects by applying constraints to the parameters of the models.

Trigonometric basis functions are not suitable for describing general shapes; for example, using a finite number of terms, they can only approximate a square corner. The relationship between variations in shape and variations in the parameters of the trigonometric expansion is not straightforward.

2.5. *Statistical Models of Shape*

A number of workers have studied the distributions of sets of "landmark" points which mark significant positions on an object. Goodall [14] discusses the registration

of shapes in arbitrary dimensions and the use of Procrustes analysis for estimating the mean shape and the covariances between landmark point coordinates and for assessing the differences between sets of shapes.

Grenander *et al.* [12] describe a method of representing a shape as a set of boundary points connected by arcs, with a statistical model of the relationships between neighboring arcs. They show how a model of a hand outline can be manipulated to fit degraded images of hands. They do this by considering sections of the boundary, and determine their most probable positions given the rest of the boundary and the local image data. By traversing the boundary in a number of sweeps the process iterates to a solution. Grenander and Miller [13] have extended this work to include gray-level information and multiple models. Mardia *et al.* [15] do something similar, representing the boundary of a shape as a sequence of points with distributions related by a covariance matrix. They too cycle through the points to find the most likely position given the image and the current shape. Both Grenander *et al.* and Mardia *et al.* represent shapes as sets of points in the complex plane. The points can vary about their means following a normal distribution with covariance matrix **S**, where **S** is modeled using either first order conditional autoregressive (CAR) models or Toeplitz covariance matrices. In our work we use a similar underlying model, but avoid any dependence of the sequence of points, thereby capturing more global shape properties. We also use principal component analysis to simplify the structure of the covariance matrix.

2.6. *Finite Element Models*

Finite element methods can be used to model variable image objects as physical entities with internal stiffness and elasticity. Pentland [18] and Pentland and Sclaroff [19] use three dimensional models which act like lumps of elastic clay. They derive modes of vibration of a suitable base shape, such as an ellipsoid, and build up shapes from different modes of vibration. The first modes are large-scale variations of shape; the higher order modes are more localized. To model human heads they use the first 30 modes. They can fit models to range data by an interative process, and can compare different heads by comparing the parameters. Terzopoulos and Metaxas [20] present a similar idea using deformable superquadrics. Nastar and Ayache [21] use a finite element approach using the vibrational modes of an example of the shape to be modelled. Karaolani *et al.* [22, 23] use finite element methods to model two dimensional objects, giving an alternative to the 'snakes' of Kass *et al.* [3].

All these methods have the advantage that the models are relatively easy to construct and allow a compact parametric representation of a family of shapes.

2.7. *The Need for Better Models*

We have described a number of existing methods for creating and using deformable models to interpret images containing structures of variable form. It is important to explain at this point why we believe a new method is required. We argue that the key issue is one of model specificity. It is necessary that a deformable model should be able to accommodate the range of variation found in the objects it is used to represent, but not sufficient. The principal role of a model is to facilitate robust automatic interpretation even in images which are noisy or cluttered or where parts of the objects of interest may be occluded. If the model is nonspecific, in the sense that it is able to deform so as to represent objects which are not valid examples of the class to be recognized, then this robustness is compromised. Our objective has been to develop models which can *only* deform in ways which are characteristic of the objects they represent. In general, the mechanisms which give rise to variability are insufficiently well understood to allow a theoretical model of deformability to be proposed. The only realistic approach is to "learn" specific patterns of variability from a representative training set of the structures to be modeled. Others have attempted something similar by placing limits on model parameters on the basis of their distributions determined from a training set. If, as is generally the case, the model parameters are correlated over the training set, this approach does not effectively restrict the shapes which can be generated to ones similar to those found in the original training set (see Fig. 1). Our approach is to find a basis for shape representation in which the shape parameters are uncorrelated over the training set. In this case simple limits on each parameter constrain the model to generate shapes similar to those in the training set. We also show that it is straightforward to use these models in image interpretation.

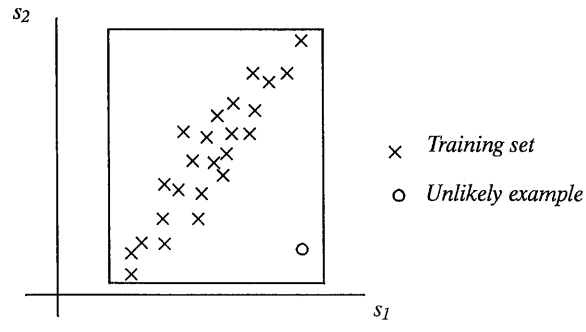


FIG. 1. If two or more shape parameters (s_1 and s_2) are correlated over a set of shapes then simple ranges to the parameters do not restrict shapes to ones similar to those in the original set.

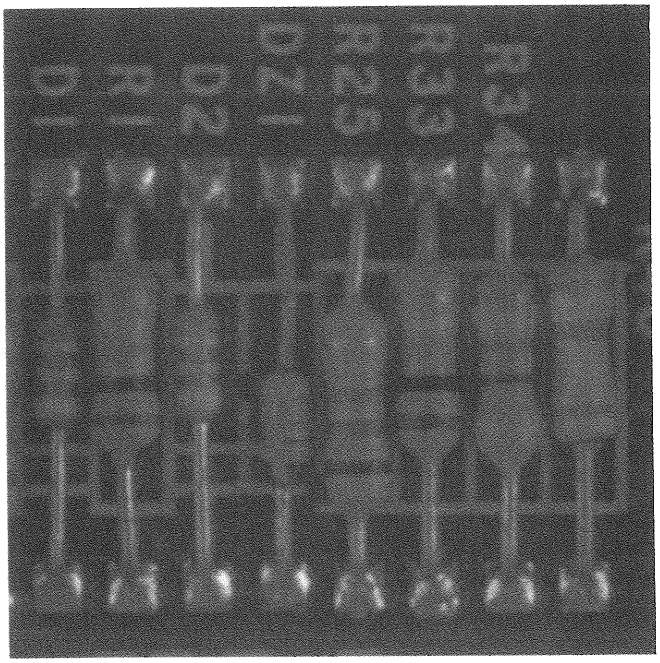


FIG. 2. Image of printed circuit board showing examples of resistors.

3. POINT DISTRIBUTION MODELS

Suppose we wish to derive a model to represent the shapes of resistors as they appear on printed circuit boards, as shown in Fig. 2. Different examples of resistor have sufficiently different shapes so that a rigid model would not be appropriate. Figure 3 shows some examples of resistor boundaries which were obtained from backlit images of individual resistors. Our aim is to build a model which describes both typical shape and typical variability, using the examples in Fig. 3 as a training set. We achieve this by representing each example as a set of labeled 'landmark' points, calculating the mean positions of the points and the main ways in which the points from each example tend to vary from the mean.

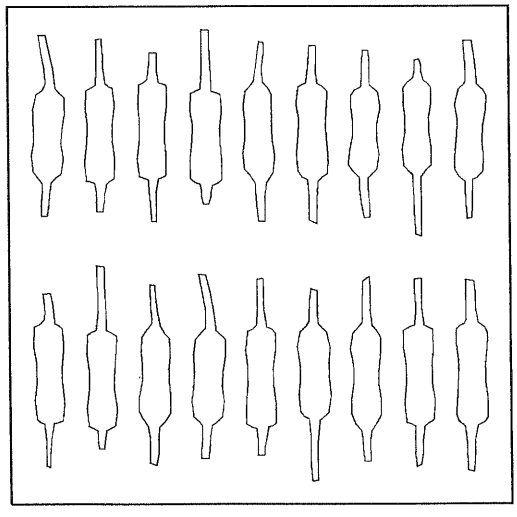


FIG. 3. Examples of resistor shapes from a training set.

3.1. *Labeling the Training Set*

In order to model a shape, we represent it by a set of points. For the resistors we have chosen to place points around the boundary, as shown in Fig. 4. This must be done for each shape in the training set. The labeling of the points is important. Each labeled point represents a particular part of the object or its boundary. For instance, in the resistor model, points 0 and 31 always represent the ends of a wire, points 3, 4, and 5 represent one end of the body of the resistor, and so on. The method works by modeling how different labeled points tend to move together as the shape varies. If the labeling is incorrect, with a particular point placed at different sites on each training shape, the method will fail to capture shape variability reliably. In the examples shown below the points were either placed manually on each image, or tools were used to mark points on boundaries segmented by hand. It is worth noting that the points are only placed manually during the training phase; it is *not* necessary to find these points in advance when the models are used for image interpretation—we describe later how this is achieved implicitly using an automatic method.

Bookstein [16, 17] labeled significant points in images of biological and medical specimens in order to examine and measure shape changes which could be correlated with other factors. We use representative points to capture shape constraints and build models which may be used to construct plausible new examples of the shape for use in image interpretation. Bookstein calls his representative points "landmark points" and describes them in terms of their usefulness. For our purposes they can be reduced to three different types:

1. points marking parts of the object with particular application-dependent significance, such as the center of an eye in the model of a face or sharp corners of a boundary;

2. points marking application-independent things, such as the highest point on an object in a particular orientation, or curvature extrema;

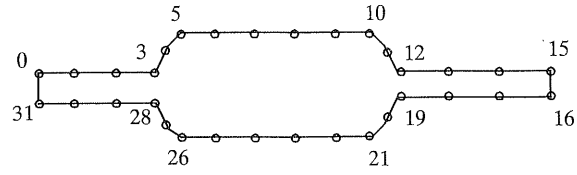


FIG. 4. Thirty-two point model of the boundary of a resistor.

3. other points which can be interpolated from points of type 1 and 2; for instance, points marked at equal distances a round a boundary between two type 1 landmarks.

On the resistor shown in Fig. 4 points 0, 3, 5, 10 and so on mark easily identified features and so are points of type 1. The other points are equally spaced along the boundaries between, and so are of type 3. Landmark points of type 1 are preferable to those of type 2, since they are in general easier to identify precisely. However, points of type 2 and 3 are almost always necessary to define the boundary of a flexible shape in sufficient detail to be useful.

It is important to note that landmark points can be used to describe single objects or sets of spatially related objects—the points can come from several different components of a structure. Typically we use boundary points, and associate boundary segments with appropriate pairs of landmark points. Although we have implemented and investigated the use of interpolating splines to generate boundaries using a minimal set of landmarks we find that simply using enough type 3 points to describe the curve to sufficient accuracy is as effective and is computationally more efficient.

3.2. *Aligning the Training Set*

Our modeling method works by examining the statistics of the coordinates of the labeled points over the training set. In order to be able to compare equivalent points from different shapes, they must be aligned with respect to a set of axes. We achieve the required alignment by scaling, rotating, and translating the training shapes so that they correspond as closely as possible. We aim to minimize a weighted sum of squares of distances between equivalent points on different shapes. This is a modification of the Procrustes method [24].

We first consider aligning a pair of shapes. Let $\mathbf{x}_i$ be a vector describing the n points of the ith shape in the set:

$$\mathbf{x}_i = (x_{i0}, y_{i0}, x_{i1}, y_{i1}, \ldots, x_{ik}, y_{ik}, \ldots, x_{in-1}, y_{in-1})^{\mathrm{T}}. \quad (2)$$

Let $M(s, \theta)[\mathbf{x}]$ be a rotation by θ and a scaling by s. Given two similar shapes, $\mathbf{x}_i$ and $\mathbf{x}_j$ we can choose θ_j, s_j and a translation (t_{xj}, t_{yj}) mapping $\mathbf{x}_i$ onto $M(s_j, \theta_j)[\mathbf{x}_j] + \mathbf{t}_j$ so as to minimize the weighted sum

$$E_j = (\mathbf{x}_i - M(s_j, \theta_j)[\mathbf{x}_j] - \mathbf{t}_j)^{\mathrm{T}}\mathbf{W}(\mathbf{x}_i - M(s_j, \theta_j)[\mathbf{x}_j] - \mathbf{t}_j), \quad (3)$$

where

$$M(s, \theta)\begin{bmatrix} x_{jk} \\ y_{jk} \end{bmatrix} = \begin{pmatrix} (s\cos\theta)x_{jk} - (s\sin\theta)y_{jk} \\ (s\sin\theta)x_{jk} + (s\cos\theta)y_{jk} \end{pmatrix}, \quad (4)$$

$$\mathbf{t}_j = (t_{xj}, t_{yj}, \ldots, t_{xj}, t_{yj})^{\mathrm{T}}, \quad \text{and} \quad (5)$$

$\mathbf{W}$ is a diagonal matrix of weights for each point.

Details are given in Appendix A.

The weights can be chosen to give more significance to those points which tend to be most "stable" over the set—the ones which move about least with respect to the other points in a shape. We have used a weight matrix defined as follows: let R_{kl} be the distance between points k and l in a shape; let $V_{R_{kl}}$ be the variance in this distance over the set of shapes; we can choose a weight, w_k, for the kth point using

$$w_k = \left(\sum_{l=0}^{n-1} V_{R_{kl}}\right)^{-1}. \quad (6)$$

If a point tends to move around a great deal with respect to the other points in the shape, the sum of variances will be large, and a low weight will be given. If, however, a point tends to remain fixed with respect to the others, the sum of variances will be small, a large weight will be given and matching such points in different shapes will be a priority.

We use the following algorithm to align a set of N shapes;

- Rotate, scale, and translate each shape to align with the first shape in the set.
- **Repeat**
  - Calculate the mean shape from the aligned shapes.
  - Normalize the orientation, scale and origin of the current mean to suitable defaults.
  - Realign every shape with the current mean.
- **Until** the process converges.

Normalizing the mean to a default scale and pose during each iteration is required to ensure that the algorithm converges. Without this there are in effect $4(N-1)$ constraints on $4N$ variables (θ, s, t_x, t_y for each of the N shapes) and the algorithm is ill-conditioned—the mean will shrink, rotate, or slide off to infinity. Constraints on the pose and scale of the mean allow the equations to have a unique solution. Either the mean is scaled, rotated, and translated so it matches the first shape, or an arbitrary default setting can be used, such as choosing an origin at its center of gravity, an orientation such that a particular part of the shape is at the top, and a scale such that the distance between two selected points is one unit. Note that normalizing the current mean shape and then aligning the shapes to match is not the same as normalizing each individual shape. If every shape were normalized in scale by setting the distance between a particular two points to be one unit, artificial correlations might be forced upon the set, distorting the model. However, if each shape is aligned with the mean, each will have a scale similar to

that of the mean. In this case the landmark point positions will be chosen to best match the mean, rather than rigidly imposed. This leads to better models.

The convergence condition in the alignment procedure can be tested by examining the average difference between the transformations required to align each shape to the recalculated mean and the identity transformation. Experiments show that the method converges to the same result independent of which shape is aligned to in the first stage, though a formal proof of convergence has yet to be devised. We have considered direct methods of solution but have found problems with numerical stability. Since computational efficiency is not an issue during model construction the iterative method is adequate for our purposes.

3.3. *Capturing the Statistics of a Set of Aligned Shapes*

In Fig. 5 the coordinates of the some of the vertices of the aligned resistor shapes are plotted, with the mean shape overlaid. It can be seen that some of the vertices show little variability over the training set, while others form more diffuse "clouds." The Point Distribution Model (PDM) seeks to model the variation of the coordinates within these clouds. However, it must be remembered that landmarks do not move about independently—their positions are partially correlated.

Each example in the training set, when aligned, can be represented by a single point in a $2n$ dimensional space (see Eq. (2)). Thus a set of N example shapes gives a cloud of N points in this $2n$ dimensional space. We assume that these points lie within some region of the space, which we call the "Allowable Shape Domain," and that the points give an indication of the shape and size of this region. Every $2n$-D point within this domain gives a set of landmarks whose shape is broadly similar to that of those in the original training set. Thus by moving about the Allowable Shape Domain we can generate new shapes in a systematic way. The approach given below attempts to model the shape of this cloud in a high dimensional space, and hence to capture the relationships between the positions of the individual landmark points. We make the assumption that the cloud is approximately ellipsoidal, and proceed to calculate its center (giving a mean shape) and its major axes, which give a way of moving around the cloud. Later we will discuss the implications of this ellipsoid assumption breaking down.

Given a set of N aligned shapes, the mean shape, $\bar{\mathbf{x}}$ (the center of the ellipsoidal Allowable Shape Domain), is calculated using

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \tag{7}$$

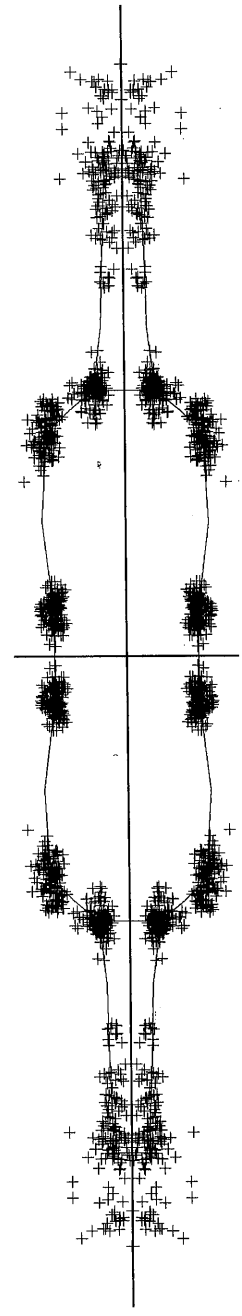


FIG. 5. Scatter of some points from aligned set of resistor shapes, with the mean shape overlaid.

The principal axes of a $2n$-D ellipsoid fitted to the data can be calculated by applying a principal component analysis (PCA) to the data [25]. Each axis gives a "mode of variation," a way in which the landmark points tend to move together as the shape varies. For each shape in the training set we calculate its deviation from the mean, $d\mathbf{x}_i$, where

$$d\mathbf{x}_i = \mathbf{x}_i - \bar{\mathbf{x}}. \tag{8}$$

We can then calculate the $2n \times 2n$ covariance matrix, $\mathbf{S}$, using

$$\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N} d\mathbf{x}_i \, d\mathbf{x}_i^{\mathrm{T}}. \tag{9}$$

The principal axes of the ellipsoid, giving the modes of variation of the points of the shape, are described by $\mathbf{p}_k$ ($k = 1, \ldots, 2n$), the unit eigenvectors of $\mathbf{S}$ such that

$$\mathbf{S}\mathbf{p}_k = \lambda_k \mathbf{p}_k \tag{10}$$

(where λ_k is the kth eigenvalue of $\mathbf{S}$, $\lambda_k \geq \lambda_{k+1}$),

$$\mathbf{p}_k^{\mathrm{T}}\mathbf{p}_k = 1. \tag{11}$$

It can be shown that the eigenvectors of the covariance matrix corresponding to the largest eigenvalues describe the longest axes of the ellipsoid, and thus the most significant modes of variation in the variables used to derive the covariance matrix. The variance explained by each eigenvector is equal to the corresponding eigenvalue [25]. Most of the variation can usually be explained by a small number of modes, t. This means that the $2n$ dimensional ellipsoid is approximated by a t dimensional ellipsoid, where t is chosen so that the original ellipsoid has a relatively small width along axes $t + 1$ and above. One method for calculating t is to chose the smallest number of modes such that the sum of their variances explains a sufficiently large proportion of λ_T, the total variance of all the variables, where

$$\lambda_T = \sum_{k=1}^{2n} \lambda_k. \tag{12}$$

Any point in our Allowable Shape Domain (i.e., any allowable shape) can be reached by taking the mean and adding a linear combination of the eigenvectors. The kth eigenvector affects point l in the model by moving it along a vector parallel to (dx_{kl}, dy_{kl}), which is obtained from the lth pair of elements in $\mathbf{p}_k$:

$$\mathbf{p}_k^{\mathrm{T}} = (dx_{k0}, dy_{k0}, \ldots, \underline{dx_{kl}, dy_{kl}}, \ldots, dx_{kn-1}, dy_{kn-1}). \tag{13}$$

An shape in the training set can be approximated using the mean shape and a weighted sum of these deviations obtained from the first t modes:

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}\mathbf{b}, \quad \text{where} \tag{14}$$

$\mathbf{P} = (\mathbf{p}_1 \, \mathbf{p}_2 \ldots \mathbf{p}_t)$ is the matrix of the first t eigenvectors,

and $\mathbf{b} = (b_1 \, b_2 \ldots b_t)^{\mathrm{T}}$ is a vector of weights.

The above equations allow us to generate new examples of the shapes by varying the parameters (b_k) within suitable limits, so the new shape will be similar to those in the training set. The parameters are linearly independent, though there may be nonlinear dependencies still present. The limits for b_k are derived by examining the distributions of the parameter values required to generate the training set. Since the variance of b_k over the training set can be shown to be λ_k, suitable limits are typically of the order of

$$-3\sqrt{\lambda_k} \leq b_k \leq 3\sqrt{\lambda_k}, \tag{15}$$

since most of the population lies within three standard deviations of the mean.

Alternatively, one can choose sets of parameters $\{b_1, \ldots, b_t\}$ such that the Mahalanobis distance (D_m) from the mean is less than a suitable value, D_{max}:

$$D_m^2 = \sum_{k=1}^{t} \left(\frac{b_k^2}{\lambda_k}\right) \leq D_{max}^2. \tag{16}$$

If each shape parameter is normally distributed then D_m will be chi-squared distributed, and D_{max} can be chosen to include a suitably large proportion of possible realizations.

3.4. *Practical Examples*

The techniques described above have been used to generate point distribution models (PDMs) for both manmade and biological objects. We present results for the set of resistor shapes shown in Fig. 3, a set of heart ventricle shapes, and a set of hand shapes. Other examples have been described elsewhere [33, 34].

3.4.1. *Resistor Model.* The resistor shapes illustrated in Fig. 3 were aligned using the method described above, with the mean shape scaled so the average distance of the points from their center of gravity was one unit. Figure 5 shows the mean shape. The most significant eigenvalues of the covariance matrix derived are shown in Table 1.

Figure 6 shows the plot of b_1 against b_2 for the training set. The lack of structure in the scatter plot suggests that the parameters can be treated as independent. We are currently working on deriving more formal tests of independence. Any dependencies between the parameters would imply nonlinear relationships between the original point positions and would result in some combinations of parameters generating "illegal" shapes. By varying the first three parameters separately we can generate examples of the shape as shown in Figs. 7–9. Each of the parameters represents a mode of variation of the shape which can frequently be associated with an intuitive de-

TABLE 1
Eigenvalues of the Covariance Matrix Derived from a Set of Resistor Shapes

Eigenvalue	$\frac{\lambda_i}{\lambda_T} \times 100\%$
λ_1	66%
λ_2	8%
λ_3	5%
λ_4	4%
λ_5	3%
λ_6	3%

scription of the deformation. Compare Figs. 7–9 with Fig. 3. Varying the first parameter (b_1) adjusts the position of the body of the resistor up and down the wire. The second parameter varies the shape of the ends of the main body of the resistor, between tapered and square. The third parameter affects the curvature of the wires at either end. Subsequent parameters have smaller effects, including the wires bending in opposite directions. These modes of variation effectively capture the variability present in the training set. Note that the apparently large variability in the positions of individual points in Fig. 3 is in fact highly constrained, and the overall variation in shape can be described by a small number of modes. This model has been used to locate resistors in images (see below).

3.4.2. *Heart Model.* Figure 10 shows examples from a set of 66 heart chamber boundaries obtained by asking a cardiologist to draw over echocardiogram images. Each

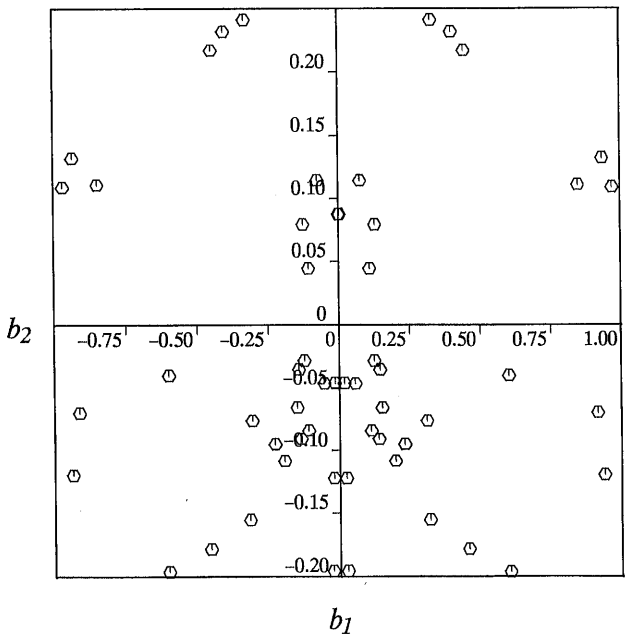


FIG. 6. Plot of b_1 vs b_2 for a training set of resistor shapes.

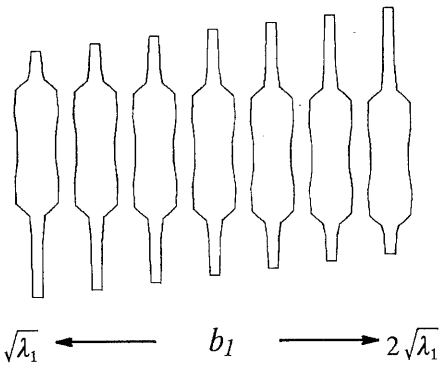


FIG. 7. Effects of varying the first parameter of the resistor model.

structure is represented by 96 points. This example shows how a single model can represent several shapes *and* the spatial relationships between them. The shape variation arises from two sources: the training set was derived from several individuals, and in each case images were taken from different stages in the cardiac cycle, during which the sizes and shapes of the heart chambers can change considerably. The points represent the boundary of the left ventricle, part of the boundary of the right ventricle, and part of the boundary of the left atrium (below the ventricle in the figures). Table 2 shows the eigenvalues of the covariance matrix obtained for the training set. Figure 11 suggests that b_1 and b_2 are again independent, and Fig. 12 shows reconstructed shapes obtained by varying the first four model parameters in turn. The first parameter varies the width of the shape. The second parameter varies the appearance of the septum (the wall separating the left from the right ventricle). The third and fourth parameters vary the shape of the left ventricle and the modeled part of the atrium below. It should be emphasized that these modes are derived entirely automatically, and arise from a statistical analysis of the variation in the data. This model has been used to locate the boundary of a

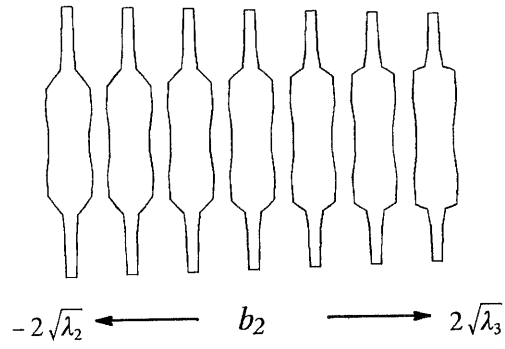


FIG. 8. Effects of varying the second parameter of the resistor model.

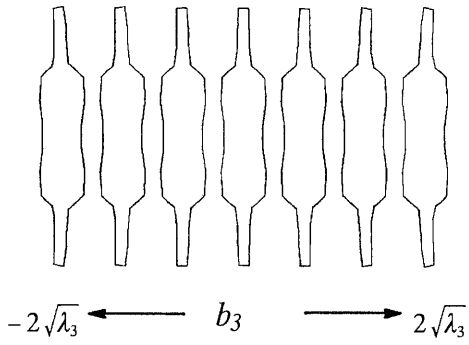


FIG. 9. Effects of varying the third parameter of the resistor model.

TABLE 2
Eigenvalues of the Covariance Matrix Derived from a Set of Heart Ventricle Shapes

Eigenvalue	$\frac{\lambda_i}{\lambda_T} \times 100\%$
λ_1	37%
λ_2	17%
λ_3	13%
λ_4	7%
λ_5	6%
λ_6	4%

heart ventricle in echocardiograms (see below, and also [26]).

3.4.3. *Hand Model.* A set of 18 hand shapes was generated from images of the right hand of one of the authors (Fig. 13). Each was represented by 72 points around the boundary. These were planted on the examples by locating 12 landmark points at the ends and joints of the fingers and filling in the rest equally along the connecting boundaries. A model was trained on the data, and it was found that 96% of the variance could be explained by the first 6 modes of variation. The first three modes are shown in Fig. 14, and consist of combinations of movements of the fingers. Again, a compact parameterized model has been generated, which has been used to locate hands in images (see below).

3.4.4. *Worm Model—Limitations of the PDM.* We have found that the linear models described above are effective in a very broad range of applications. There are, however, some situations where the method breaks down.

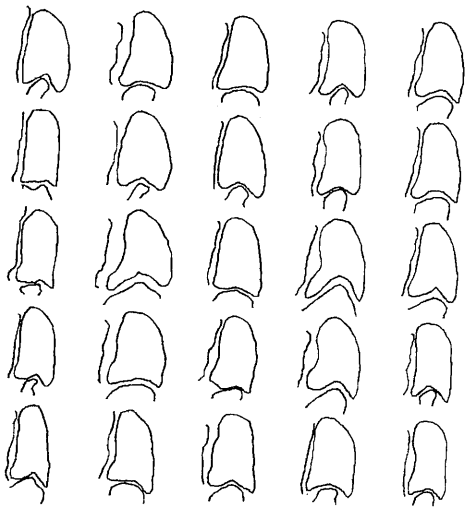


FIG. 10. Examples of heart ventricle shapes, each containing 96 points.

Each mode of variation in a PDM moves the landmark points along straight lines relative to the local coordinate system. In some cases the variations in a class of shapes would be better represented by moving points along curves. This can be especially important when bending or relative rotational effects occur in the class of example shapes. Consider, for instance, the examples from the set of "worm" shapes shown in Fig. 15. The set consists of 84 artificially generated shapes which have a fixed width but varying curvature and length, each represented by 12 labeled points (Fig. 16). Figure 17 shows the scatter of some of the points once the shapes have been aligned. The varying curvature leads to the points at the ends of the shape forming a curved cloud.

The shape formed by the mean positions of the labeled points does not have a constant width and is shorter than the aligned versions of the training shapes. The end points,

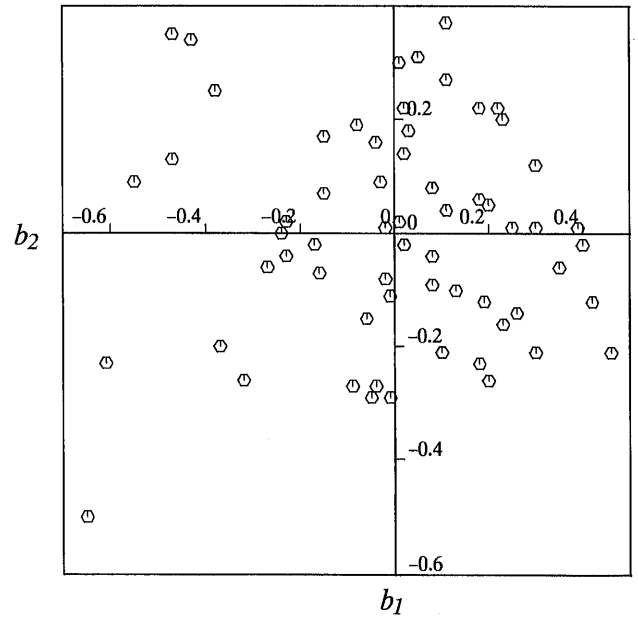


FIG. 11. b_1 vs b_2 for the training set of heart ventricle examples.

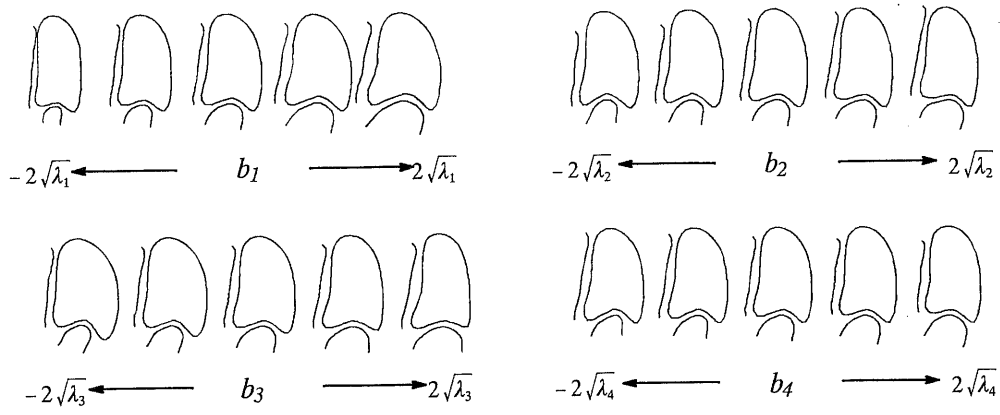


FIG. 12. Effects of varying each of the first four parameters of the heart ventricle model individually.

0 and 6, form curved clouds, the centroids of which do not lie inside the clouds. The mean shape generated in this way is thus not sufficiently similar to the training set to give a satisfactory model. The first three modes of variation of a PDM trained on this data are shown in Fig. 18. Ideally one would expect a model to have the first and second order curvature as its first two modes. The first mode of the PDM is an approximation to bending, generated by fitting straight lines to the curved "clouds" of points. The second mode gives the corrections required because the linear approximation is poor. The third mode of the model gives an approximation to second order bending. Figure 19 shows the relationship between the first two parameters b_1 and b_2. Though they are *linearly* independent, there are clearly nonlinear relationships present. One cannot choose the parameters independently and ex-

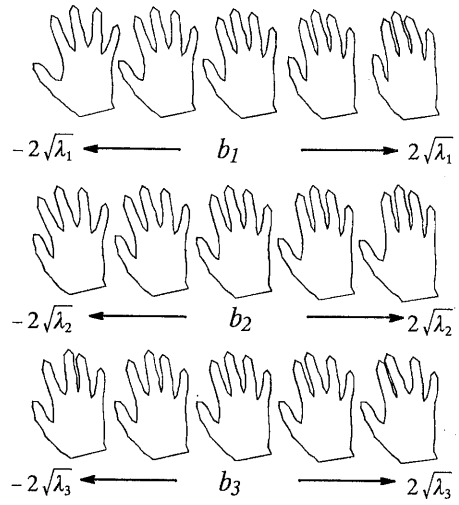


FIG. 14. Effects of varying each of the first three parameters of the hand model individually.

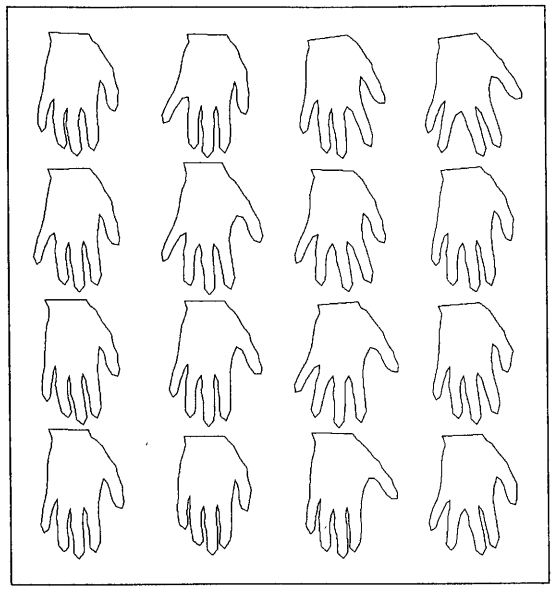


FIG. 13. Training set of hand shapes, each defined by 72 points.

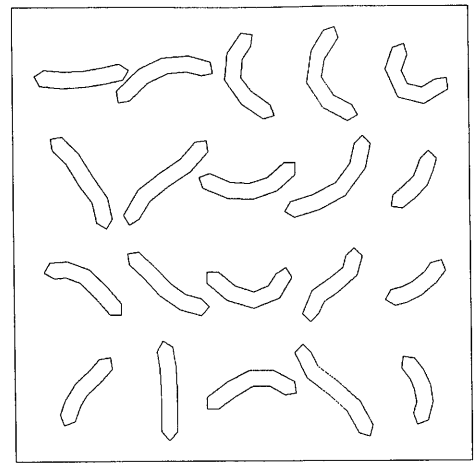


FIG. 15. Examples from a set of "worm" shapes.

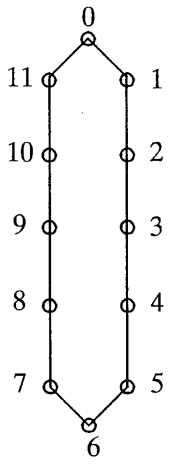


FIG. 16. Labeling of points in "worm" shapes.

pect to get a shape similar to those in the training set. We discuss ways in which this problem might be overcome at the end of the paper.

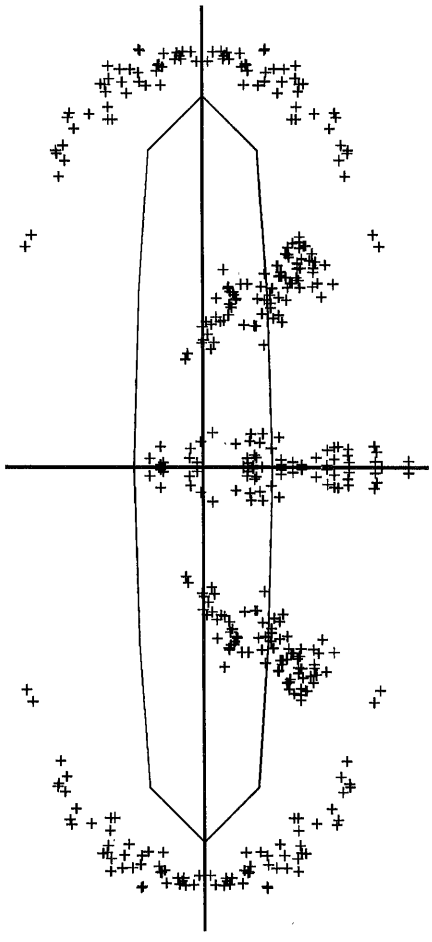


FIG. 17. Scatter of points 0, 2, 3, 4, and 6 from the aligned set of "worms," with the mean shape overlaid.

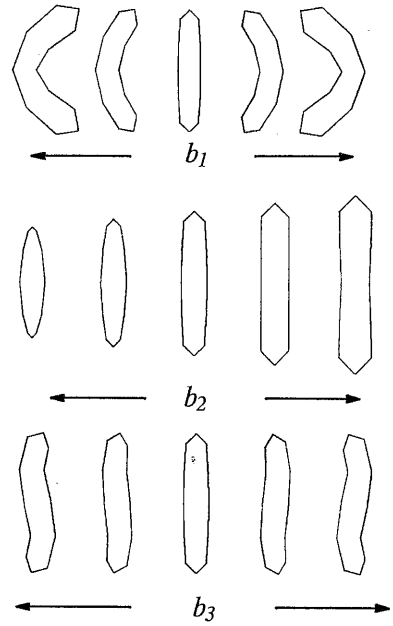


FIG. 18. Effects of varying the first three parameters (b_1, b_2, b_3) of the "worm" model individually.

4. USING POINT DISTRIBUTION MODELS IN IMAGE SEARCH—ACTIVE SHAPE MODELS

Having generated flexible models, we would like to use them in image search, to find new examples of modeled objects in images. This involves finding the shape and pose parameters which cause the model to coincide with

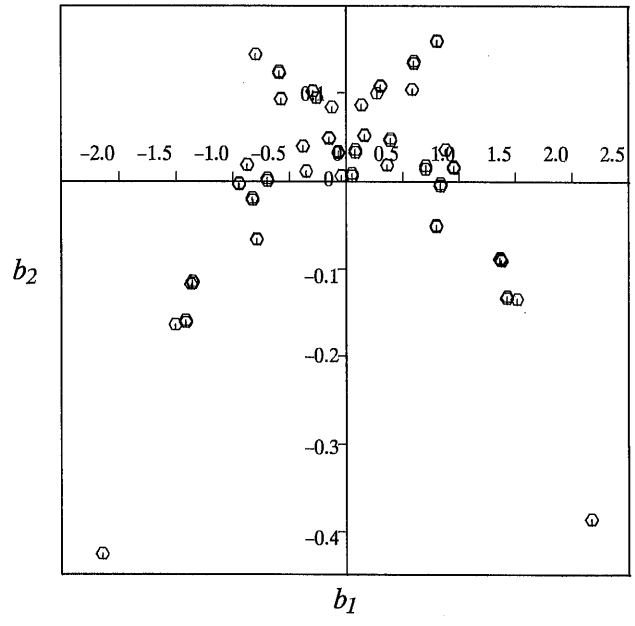


FIG. 19. b_1 vs b_2 for the training set of "worm" examples.

the structures of interest in the image. An instance of the model is given by

$$\mathbf{X} = M(s, \theta)[\mathbf{x}] + \mathbf{X}_c, \quad \text{where}$$
$$\mathbf{X}_c = (X_c, Y_c, X_c, Y_c, \ldots, X_c, Y_c)^{\mathrm{T}}$$
$$M(s, \theta)[\ ] \text{ is a rotation by } \theta \text{ and a scaling by } s, \text{ and} \quad (17)$$
$$(X_c, Y_c) \text{ is the position of the centre of the model in the image frame.}$$

In this section we describe an iterative method for finding the appropriate $\mathbf{X}$ given a very rough starting approximation. Hill *et al.* have described elsewhere how Genetic Algorithm search can be used to find a good starting approximation quite rapidly [26, 7, 27]; this is applicable if there is no prior knowledge of the expected location of objects of interest. In practice, the starting value of $\mathbf{X}$ does not need to be very close to the final solution, so that, for many practical applications, the method below can be used on its own.

The idea of the iterative scheme is to place the current estimate of $\mathbf{X}$ into the image and examine a region of the image around each model point to determine a displacement which moves it to a better location. These local deformations are transformed into adjustments to the pose, scale, and shape parameters of the PDM. By enforcing limits on the shape parameters, global shape constraints can be applied ensuring the shape of the model example remains similar to those of the training set. The procedure is repeated until no significant changes result. Because the models attempt to deform to better fit the data, but only in ways which are consistent with the shapes found in the training set, we call them "Active Shape Models" or "Smart Snakes."

4.1. *Calculating a Suggested Movement for Each Model Point*

Given an initial estimate of the positions of a set of model points which we are attempting to fit to an image object we need to find a set of adjustments which will move each point toward a better position. When the model points represent the boundaries of objects (Fig. 20) this involves moving them toward the image edges. There are various approaches that could be taken. In the examples we describe below we use an adjustment along a normal to the model boundary toward the strongest image edge, with magnitude proportional to the strength of the edge (Fig. 21).

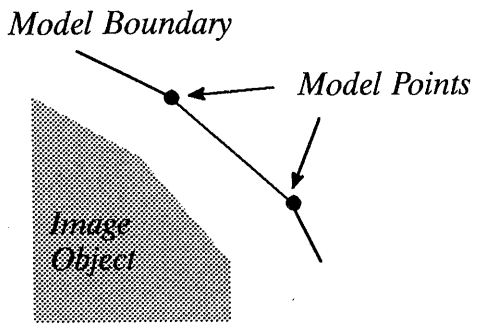


FIG. 20. Part of a model boundary approximating to the edge of an image object.

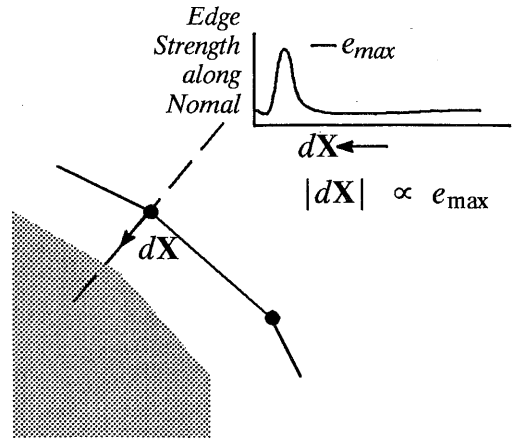


FIG. 21. Suggested movement of point is along normal to boundary, proportional to maximum edge strength on normal.

An alternative approach is to generate potential images such as those described by Kass *et al.* [3], possibly one for each model point, describing the likelihood of each point in the image being the model point. Adjustments to each point position can then be derived from the gradient of the potential image at the current estimate of the point's position.

However they are obtained, we denote the set of adjustments (Fig. 22) as a vector $d\mathbf{X}$, where

$$d\mathbf{X} = (dX_0, dY_0, \ldots, dX_{x-1}, dY_{n-1})^{\mathrm{T}}.$$

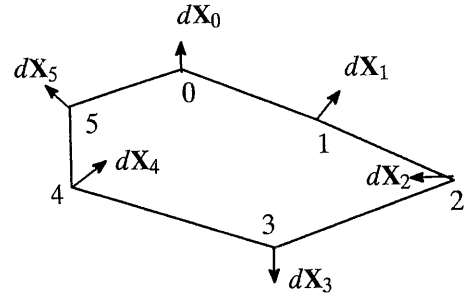


FIG. 22. Adjustments to a set of points.

4.2. *Computing Changes in the Pose and Shape Parameters*

We aim to adjust the pose and shape parameters of the model to move the points from their current locations in the image frame, $\mathbf{X}$, to be as close to the suggested new locations ($\mathbf{X} + d\mathbf{X}$) as can be arranged while still satisfying the shape constraints of the model. If the current estimate of the model is centered at (X_c, Y_c) with orientation θ and scale s we would like first to calculate how to update these parameters to better fit the image. This is achieved by finding the translation (dX_c, dY_c), rotation $d\theta$ and scaling factor $(1 + ds)$ which best map the current set of points, $\mathbf{X}$, onto the set of points given by ($\mathbf{X} + d\mathbf{X}$) using the method given in Appendix A.

Having adjusted the pose variables there remain residual adjustments which can only be satisfied by deforming the shape of the model. We wish to calculate the adjustments, $d\mathbf{x}$, in the local coordinate frame required to cause the points $\mathbf{X}$ to move by $d\mathbf{X}$ when combined with the effect of the new scale, rotation and translation parameters.

The initial position of the points in the image frame is given by Eq. (17),

$$\mathbf{X} = M(s, \theta)[\mathbf{x}] + \mathbf{X}_c.$$

We wish to calculate a set of residual adjustments $d\mathbf{x}$ in the local model coordinate frame such that

$$M(s(1 + ds), (\theta + d\theta)[\mathbf{x} + d\mathbf{x}] + (\mathbf{X}_c + d\mathbf{X}_c) = (\mathbf{X} + d\mathbf{X}). \tag{18}$$

Thus

$$M(s(1 + ds), \theta + d\theta)[\mathbf{x} + d\mathbf{x}] = (M(s, \theta)[\mathbf{x}] + d\mathbf{X}) - (\mathbf{X}_c + d\mathbf{X}_c)$$

and since

$$M^{-1}(s, \theta)[\ ] = M(s^{-1}, -\theta)[\ ]$$

we obtain

$$d\mathbf{x} = M((s(1 + ds))^{-1}, -(\theta + d\theta))[\mathbf{y}] - \mathbf{x}, \quad \text{where } \mathbf{y} = M(s, \theta)[\mathbf{x}] + d\mathbf{X} - d\mathbf{X}_c. \tag{19}$$

Equation (19) gives a way of calculating the suggested movements to the points $\mathbf{x}$ in the local model coordinate frame. These movements are not in general consistent with our shape model. In order to apply the shape constraints we transform $d\mathbf{x}$ into model parameter space, giving $d\mathbf{b}$, the changes in model parameters required to adjust the model points as closely to $d\mathbf{x}$ as is allowed.

Equation (14) gives

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}\mathbf{b}.$$

We wish to find $d\mathbf{b}$ such that

$$\mathbf{x} + d\mathbf{x} \approx \bar{\mathbf{x}} + \mathbf{P}(\mathbf{b} + d\mathbf{b}). \tag{20}$$

Since there are only t ($<2n$) modes of variation available and $d\mathbf{x}$ can move the points in $2n$ different degrees of freedom, we can only achieve an approximation to the deformation required.

Subtracting (14) from (20) gives

$$d\mathbf{x} \approx \mathbf{P}(d\mathbf{b})$$

so

$$d\mathbf{b} = \mathbf{P}^{\mathrm{T}} d\mathbf{x} \tag{21}$$

since $\mathbf{P}^{\mathrm{T}} = \mathbf{P}^{-1}$, as the columns of $\mathbf{P}$ are mutually orthogonal and of unit length.

It can be shown that Eq. (21) is equivalent to using a least-squares approximation to calculate the shape parameter adjustments, $d\mathbf{b}$.

4.3. *Updating the Pose and Shape Parameters*

The equations above allow us to calculate changes to the pose variables and adjustments, dX_c, dY_c, $d\theta$, and ds, to the shape parameters $d\mathbf{b}$ required to improve the match between an object model and image evidence. We apply these to update the parameters in an iterative scheme,

$$X_c \to X_c + w_t\, dX_c \tag{22}$$

$$Y_c \to Y_c + w_t\, dY_c \tag{23}$$

$$\theta \to \theta + w_\theta\, d\theta \tag{24}$$

$$s \to s(1 + w_s\, ds) \tag{25}$$

$$\mathbf{b} \to \mathbf{b} + \mathbf{W}_b\, d\mathbf{b}, \tag{26}$$

where w_t, w_s, and w_θ are scalar weights, and $\mathbf{W}_b$ is a diagonal matrix of weights, one for each mode. This can be the identity, or each weight can be proportional to the standard deviation of the corresponding shape parameter over the training set. The latter allows more rapid movement in modes in which there tend to be larger shape variations. We can ensure that the model only deforms into shapes consistent with the training set by placing limits on the values of b_k. A shape can be considered acceptable if the Mahalanobis distance D_{m} is less than a suitable constant, $D_{\max}$, say 3.0 (See Eq. (16)). This limit is calculated so that almost all the training examples satisfy Eq. (16).

The vector **b** should lie within a hyperellipsoid about the origin. If updating **b** using (26) leads to an implausible shape, i.e., $D_m > D_{max}$ and the point lies outside the ellipsoid, **b** can be rescaled to lie on the closest point of the allowed volume using

$$b_k \rightarrow b_k \cdot \left(\frac{D_{max}}{D_m}\right) (k = 1, \ldots, t). \quad (27)$$

Note that we have already applied implicit limits of zero to the weights of the eigenvectors truncated from our representation (i.e., $b_i = 0 \; \forall \; i > t$). Once the parameters have been updated, and limits applied where necessary, the updated positions of the model points can be calculated, and new suggested movements derived for each point. The procedure is repeated until no significant change results.

4.4. EXAMPLES USING ACTIVE SHAPE MODELS

The techniques described above have been used successfully in a number of applications, both industrial and medical [26, 27, 33]. Here we show results using the resistor, heart, and hand models described above.

In each case initial estimates of the position, orientation, and scale are made, and the shape parameters of the Active Shape Model (ASM) are initialized at zero ($b_i = 0 \; (i = 1, \ldots, t)$). Suggested movements for each model point are calculated by finding the strongest edge (of the correct polarity) along the normal to the boundary at the point (see 4.1 and Fig. 21). Adjustments to the parameters are calculated and applied, and the process is repeated.

4.4.1. *Locating Resistors.* We have constructed a Point Distribution Model of a resistor, representing its boundary using 32 points (Section 3.4.1). Figure 23 shows an image of part of a printed circuit board with the resistor boundary model superimposed as it iterates toward a component in the image. We interpolate an additional 32 points, one between each pair of model points around the boundary, and calculate adjustments to each point by finding the strongest edge along profiles 20 pixels long centred at each point. We use a shape model with 5 degrees of freedom. Each iteration of the ASM takes about 0.015 s on a Sun Sparc 10 Workstation.

The method is effective in maintaining the global shape constraints of the model and works well, given a sufficiently good starting approximation; we discuss methods of obtaining such initial hypotheses elsewhere [26, 27].

4.4.2. *Locating Heart Ventricles.* Figure 24a shows an example of an echocardiogram. The left ventricle is in the top right of the imaged region. Figure 24b shows the initial placement of an instance of the 96 point heart chamber model described above (Section 3.4.2). Figure 24c shows the ASM after 80 iterations. After 200 iterations (Fig. 24d) the model gives a good fit to the data. The shape model used has 12 degrees of freedom. The adjustments to each point are calculated using the strongest edge in a smoothed image along a profile 40 pixels long centered on the point. Each ASM iteration takes about 0.03 s on a Sun Sparc 10 workstation. In this example the model is able to infer the position of the parts of the boundary where there are missing data (for example, the top of the ventricle) by using the knowledge of the expected shape combined with information from the areas of the image where good evidence for the ventricle wall can be found. Without the prior knowledge of the shape given by the model it would not be possible to delineate the ventricle boundary accurately. Further medical applications of the method are described in [33].

4.4.3. *Locating Hands.* We have constructed a Point Distribution Model of a hand, representing the boundary using 72 points (Section 3.4.3). Figure 25 shows an image of one of the author's hands amid some clutter and occlusion, and an example of the model iterating towards it. We calculate adjustments to each point by finding the strongest edge on a profile 35 pixels long centred on the point. The shape model has 8 degrees of freedom, and each ASM iteration takes about 0.02 seconds on a Sun Sparc 10 Workstation. The result demonstrates that the method can deal with clutter and limited occlusion.

5. DISCUSSION

The examples given above illustrate the main features of our approach. Using a single method, specialized only by training with an appropriate set of examples, we have been able to locate automatically a range of structures in complex, noisy, and cluttered images. Other examples reported elsewhere include faces [36], handwritten characters [36], anatomical structures in magnetic resonance images of the brain and abdomen [33], vertebrae in radiographs [33], parts of the foot in pressure images [38] and all the parts in an automobile brake assembly [34]. We discuss below some of the issues which arise from this work, including areas where further development is required.

5.1. *Point Distribution Models*

5.1.1. *Choice of Model Points and Training Examples.* It is important that landmark points be placed on the training images as accurately as possible. If a point is not in the correct position on each shape, the model will be unable to correctly represent the position of that point—it will include terms describing the noise caused by errors in point location. It is equally important to arrange that all the examples used to train the model are

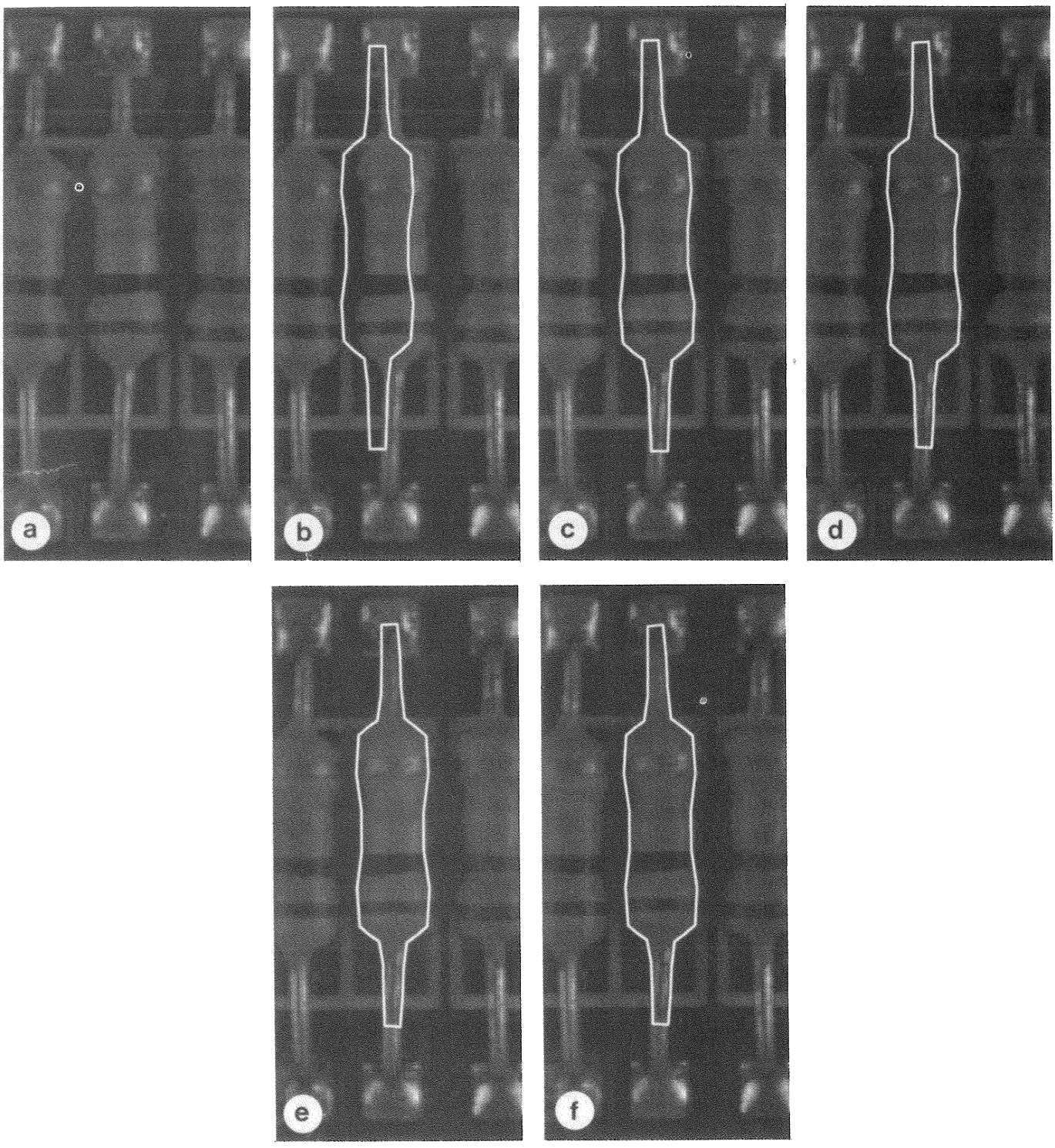


FIG. 23. Section of printed circuit board with resistor model superimposed, showing its initial position and its location after 30, 60, 90, and 120 iterations.

similarly aligned with respect to a set of axes, to ensure that the labeled points in different shapes are being compared correctly. In some cases an obvious alignment is apparent, but in others, particularly medical examples where the shapes of organs are very flexible, the automatic least-squares alignment method is essential.

Of course, placing every point by hand on every training image can be very time consuming. We are developing

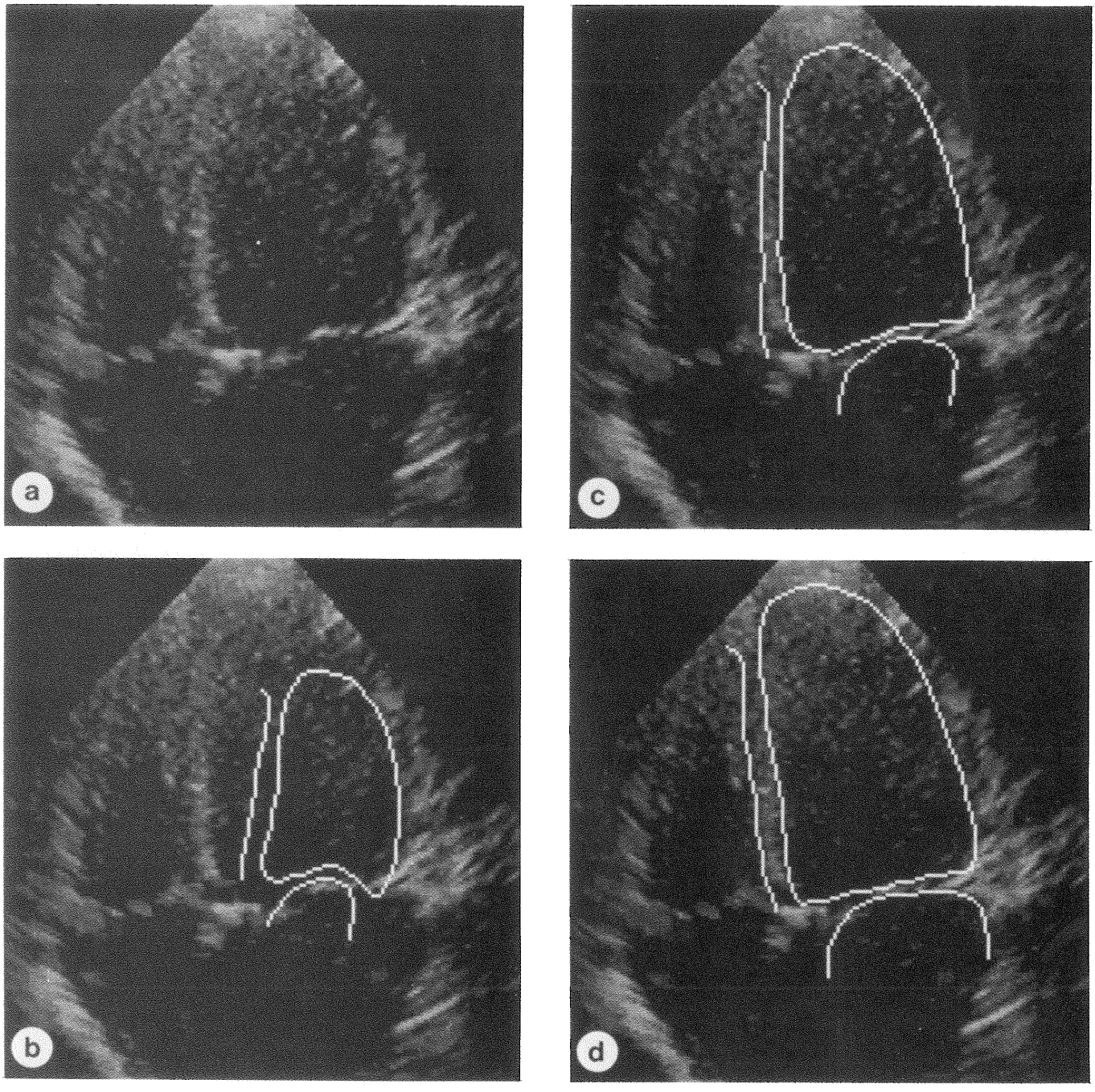


FIG. 24. Echocardiogram image with heart chamber boundary model superimposed, showing its initial position and its location after 80 and 200 iterations.

tools to ease the procedure. Techniques such as those described by Burr [29] and the Finite Element Models of Sclaroff and Pentland [30] or Nastar and Ayache [21] may be able to assist the user in locating point correspondences during this training phase.

In some cases occlusion and noise will lead to images in which some points cannot be accurately located. It is straightforward to adjust the calculation of mean shape (7) and the covariance matrix (9) to give a weighting to each point in each example in the training set. When some points are missing, the weights for known points can be set to unity; those for unknown points can be set to zero. As long as only a small proportion of points are missing in any one example, and no points are missing from all examples, it is still possible to build useful models.

In principle it is possible to "overtrain" a model. Sup-

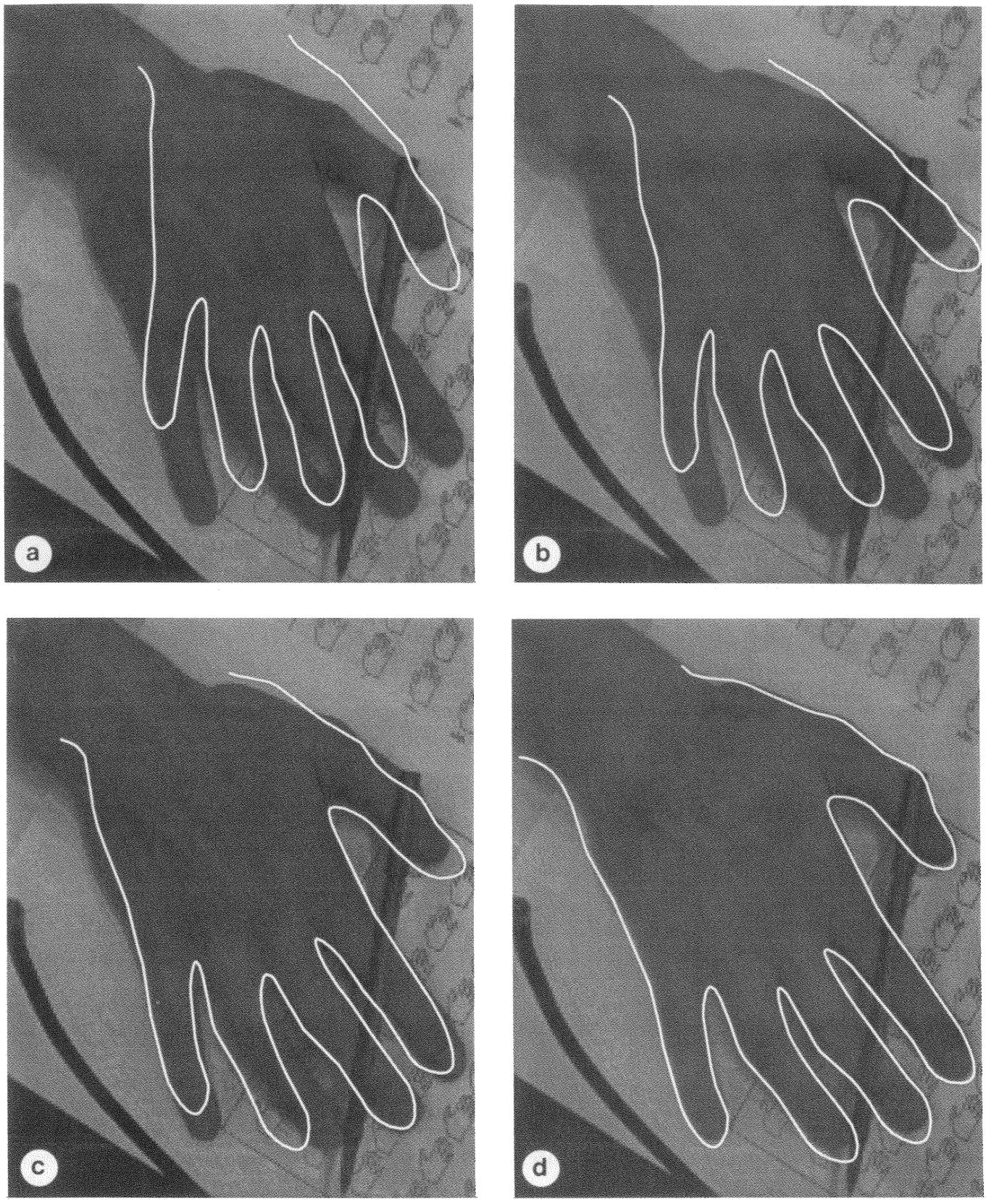


FIG. 25. Image of author's hand with hand model superimposed, showing its initial position and its location after 100, 200, and 350 iterations.

pose that a large proportion of the examples were close to the mean and there were only one or two examples demonstrating some particular form of shape variation. It is possible that when the number of modes to be used, t, is chosen, the mode which best describes the infrequent shape variation will be truncated, since it will explain only a small amount of the total variance. However, since the training examples are typically selected and labeled by hand, it is time consuming and inefficient to include many similar shapes—it is better to choose a variety of different shapes which cover the whole range of variations one is likely to observe (where such are available). It is at this stage that the expert knowledge of a human can play a part.

5.1.2. *Multipart Models.* The heart example illustrates an important fact—that the points used to construct a PDM and its derived ASM do not need to belong to a single object or shape. The connectivity of the points is not relevant to the construction of the PDM and is only used by the ASM to determine the direction of the local normal at each point during image search. The shapes of multiple subparts of a complex assembly and the spatial relationships between them can thus be represented by a single PDM. A significant advantage of handling shape and spatial relationships in a unified way is that correlations between the positions and shapes of subcomponents can be modeled; this is important, for example, in assemblies of interlinked mechanical components or in medical images where several organs are "packed" into the same cavity.

5.1.3. *Modeling Shape Variation.* We showed in Section 3.3 that each aligned shape can be considered as a single point in $2n$ dimensional space, and the whole training set as a cloud of points in this space. We attempt to model this cloud using the idea of an Allowable Shape Domain. For the search method to work effectively it is important that this domain be simply connected, and that we have a simple method of navigating around the domain. The assumption that the domain is an ellipsoid (or a box with the same axes) allows us to do this. However, under certain circumstances this is an inappropriate model. When there is a large degree of bending or relative rotation in the training set, nonlinear relationships between landmarks can give the cloud in the $2n$ dimensional space a "banana" shape or worse. Under these circumstances, as was demonstrated in Section 3.4.4, the ellipsoidal assumption gives a shape model which can generate shapes badly distorted when compared with those from the training set. The model is not as *specific* as one would like, and only a subset of the shapes it can generate would be considered "legal." In some situations this is not disastrous. For instance, the worm model given can be used successfully to locate examples of worms in images, but the models are more susceptible to being distorted by noise or clutter than a more specific model would be.

A more general model of the allowable shape domain could lead to more specific shape models. We have experimented using polynomials, rather than straight lines, for the axes of the domain with encouraging results. Instead of each mode defining straight line motion for each point, the points follow polynomial curves as the parameter varies. Results will be presented in a further paper.

5.1.4. *Dealing with Small Numbers of Examples.* If there are fewer training examples, N, than point coordinates ($2n$), as is often the case, particularly for complex models, there can be no more than $N - 1$ degrees of freedom in the model. The principal component analysis required for the method uses the eigenvectors of the $2n \times 2n$ matrix $\mathbf{S}$ (Eqs. 9, 10). When $N < 2n$ this matrix has no more than $N - 1$ nonzero eigenvalues. Calculating all $2n$ eigenvectors in this case is unnecessary. An efficient way of calculating the eigenvectors associated with nonzero eigenvalues is given in Appendix B.

5.1.5. *Extensions to the Model.* Rather than have one "flat" PDM representing a complex assembly, it is possible to build a hierarchical PDM in which the top layer controls the position, scale, orientation, and shape parameters of the layer below. The bottom layer can consist of a number of subcomponents, each represented by a "flat" PDM. Varying the parameters of the top layer varies the pose, scale, and shape of the various components below. This avoids problems with the PDM due to rotating subcomponents—their orientation relative to the rest of the assembly can be modeled explicitly, rather than implicitly in a single-layer linear PDM.

It is also easy to extend the Point Distribution Model to deal with three dimensional data, for example, 3D medical images. We have recently described a successful system for automated interpretation of 3D Magnetic Resonance images of the brain using a 3D PDM [35].

5.1.6. *The Chord Length Distribution.* Elsewhere we have described how to derive a shape model from a training set using the distances between pairs of points—a Chord Length Distribution Model [31]. The distance, R_{ij}, between every pair of points i, j in each example of the training set is calculated, and the way these chord lengths vary is modeled by calculating their mean and covariances and applying a Principal Component Analysis. A model with several parameters is obtained, which returns sets of interpoint distances, R_{ij}, from which a new shape can be constructed. Varying the parameters varies the distances, which causes the shape to change. Such a system is able to model the rigid parts of an object regardless of their orientation, since it relies only on internal distances. Though this technique is sometimes better than the linear PDM at representing objects which can bend (such as

the "worms"), the reconstruction of the shape from the distances between points is iterative and slow. A refinement technique using such a model would be complex, leading us to favour the PDM for practical applications.

5.1.7. *Use as a Classifier.* A PDM can also be used in a classifier. Given an example of a shape, an estimate can be made of how likely that example is to be a member of the class of shapes described by a model. If labeled points are placed on the example and the point set is aligned with the mean shape, we can calculate the model parameters required to generate the example. The distributions of the parameters can be estimated from the training set, allowing probabilities to be assigned. Alternatively, classification of unknown objects in images can be made by training a model for each class to be considered. Given a new image, the ASM technique is used to fit each model to the image data. The one which gives the best fit to the image is chosen as the result [36].

5.2. *Using PDMs in Image Search*

We have shown that ASMs are effective at locating known objects in images given initial estimates of position, scale, and orientation. How good an estimate is required will depend on how cluttered the image is and how well the model describes the objects in the image. For instance, unless some points in the model are close enough to the image example for their search profiles to overlap appropriate edges on the target object, the model instance cannot "see" the target, and will not be able to move towards it. We have found that techniques using Genetic Algorithms [26, 7, 27] provide a suitably close initial estimate to achieve successful refinement with ASMs.

5.2.1. *Occlusion and Clutter.* The hand example demonstrates that ASMs can deal successfully with occlusion and clutter. The heart example also shows that the method works with noisy images and missing data. As noise, clutter, and occlusion increase it becomes more likely that the models will latch onto the wrong edges, although the constraints applied by the PDM ensure that the shape of the final result is "sensible," even when it does not locate all the edges correctly. This is most likely to happen when the clutter has a similar structure to that of the objects being modeled and is thus most likely to be mistaken for some part of the object. Because the models we use are specific, we argue that we are applying the strongest possible shape-based constraints. In extreme cases, where the method fails, the fault lies in our approach to incorporating the image evidence—this is discussed below.

5.2.2. *Updating the Model Parameters.* How suggested adjustments are found for each point is important. Calculating the suggested movement by looking for strong nearby edges is simple and has proved effective in many cases. However, to search for more complex objects, where the model points do not necessarily lie on strong edges, more sophisticated algorithms are required. Potential maps can be derived, describing the likelihood for each point in an image that it is a particular model point. During search each model point attempts to move to more likely locations, climbing hills in the potential map. Alternatively, a model of the expected gray levels around each model point can be generated from the training examples; during image search each point is then moved toward the nearby area which best matches its local gray level model. Experiments using the latter technique have shown that it is more flexible and less sensitive to noise and clutter than simply searching for strong edges [32, 34].

By allowing the model to deform, but only in ways seen in the training set, we have a powerful technique for refinement. The constraints on the shape of the model are applied by the limits on the shape parameters. The $2n - t$ unrepresented modes of variation effectively have limits of zero on their parameters. Rather than fixed limits being used to enforce shape constraints, restoring forces in the parameter space could be applied, pulling the parameters back toward zero against the external "forces" from the image;

$$\mathbf{b} \rightarrow \mathbf{b} + \mathbf{W}_b d\mathbf{b} - k_b \mathbf{W}_b \mathbf{b} \quad (0 < k_b < 1). \tag{28}$$

This would give more weight to solutions closer to the mean shape, and require strong evidence for shapes which were considerably deformed. However, since this would be likely to lead to compromise solutions between image data and model we favor the fixed limit approach.

5.2.3. *Comparison with Other Work.* The work we present here can be thought of as a two dimensional application of Lowe's refinement technique [37]. Because of the linear nature of the Point Distribution Model, the mathematics is considerably simpler and leads to rapid execution.

Our Active Shape Models are superficially similar to Active Contour Models (Snakes) and Finite Element Models (FEMs). Each have their advantages. Snakes and FEMs can be created relatively easily and have the ability to locate partially occluded objects in noisy, cluttered scenes. The models are not, however, very specific, so under difficult conditions they can generate implausible interpretations. ASMs are harder to create because they require the user to annotate each of a set of training images with the correct interpretation. However, they model the allowed variability more specifically and are thus more robust to noise, clutter, and occlusion. A detailed experimental comparison between ASM and FEM methods is beyond the scope of this paper, but we hope to present results in the near future. We are also working actively

on methods which incorporate the advantages of both approaches.

5.2.4. *A Framework for Object Modeling and Recognition.* We have conducted experiments which suggest that our local optimization method can be fruitfully used in conjunction with a Genetic Algorithm (GA) search [26–28]. The GA can be run as a cue generator to produce a number of object hypotheses, which can be refined using the Active Shape Model. Alternatively, the ASM can be combined with the GA search, applying one iteration at each generation of the Genetic Algorithm. Both techniques have been used successfully to locate complex structures in a variety of images.

6. CONCLUSIONS

We have described Point Distribution Models (PDMs)—statistical models of shape which can be constructed from training sets of correctly labeled images. A PDM represents an object as a set of labeled points, giving their mean positions and a small set of modes of variation which describe how the object's shape can change. Applying limits to the parameters of the model enforces global shape constraints ensuring that any new examples generated are similar to those in the training set. Given a set of shape parameters, an instance of the model can be calculated rapidly. The models are compact and are well suited to generate-and-test image search strategies.

Active Shape Models (ASMs) exploit the linear formulation of PDMs in an iterative search procedure capable of rapidly locating the modeled structures in noisy, cluttered images—even if they are partially occluded. Object identification and location are robust because the models are specific in the sense that instances are constrained to be similar to those in the training set.

We have demonstrated the ability to create compact models of resistors, hearts (in echocardiograms), and hands. We have also shown that these models can be used successfully in image search. Using a conventional workstation a good interpretation can typically be obtained in seconds. We have described elsewhere various other applications in which the same methods have been exploited successfully, including examples where very complex structures (e.g., faces and automobile brake assemblies) are modeled. The important point to stress is that precisely the same software can be applied to a broad range of image interpretation problems—both medical and industrial—specialized only by training with suitable examples.

We believe that this approach holds considerable promise as a practical but generic technique for automated image interpretation.

APPENDIX A: ALIGNING A PAIR OF SHAPES

Given two similar shapes, $\mathbf{x}_1$ and $\mathbf{x}_2$, we would like to choose a rotation, θ, a scale, s, and a translation, (t_x, t_y), mapping $\mathbf{x}_2$ onto $M(\mathbf{x}_2) + \mathbf{t}$ so as to minimize the weighted sum

$$E = (\mathbf{x}_1 - M(s, \theta)[\mathbf{x}_2] - \mathbf{t})^{\mathrm{T}}\mathbf{W}(\mathbf{x}_1 - M(s, \theta)[\mathbf{x}_2] - \mathbf{t}), \tag{3}$$

where

$$M(s, \theta)\begin{bmatrix} x_{jk} \\ y_{jk} \end{bmatrix} = \begin{pmatrix} (s\cos\theta)x_{jk} - (s\sin\theta)y_{jk} \\ (s\sin\theta)x_{jk} + (s\cos\theta)y_{jk} \end{pmatrix}, \tag{4}$$

$$\mathbf{t} = (t_x, t_y, \ldots, t_x, t_y)^{\mathrm{T}}, \tag{29}$$

and $\mathbf{W}$ is a diagonal matrix of weights for each point. If we write

$$a_x = s\cos\theta \quad a_y = s\sin\theta,$$

a least-squares approach (differentiating with respect to each of the variables a_x, a_y, t_x, t_y) leads to a set of four linear equations,

$$\begin{pmatrix} X_2 & -Y_2 & W & 0 \\ Y_2 & X_2 & 0 & W \\ Z & 0 & X_2 & Y_2 \\ 0 & Z & -Y_2 & X_2 \end{pmatrix}\begin{pmatrix} a_x \\ a_y \\ t_x \\ t_y \end{pmatrix} = \begin{pmatrix} X_1 \\ Y_1 \\ C_1 \\ C_2 \end{pmatrix}, \tag{30}$$

where

$$X_i = \sum_{k=0}^{n-1} w_k x_{ik} \quad Y_i = \sum_{k=0}^{n-1} w_k y_{ik} \tag{31}$$

$$Z = \sum_{k=0}^{n=1} w_k(x_{2k}^2 + y_{2k}^2) \quad W = \sum_{k=0}^{n-1} w_k \tag{32}$$

$$C_1 = \sum_{k=0}^{n-1} w_k(x_{1k}x_{2k} + y_{1k}y_{2k}) \tag{33}$$

$$C_2 = \sum_{k=0}^{n-1} w_k(y_{1k}x_{2k} - x_{1k}y_{2k}). \tag{34}$$

These can be solved for a_x, a_y, t_x, and t_y using standard matrix methods.

APPENDIX B: CALCULATING THE EIGENVECTORS OF THE COVARIANCE MATRIX WHEN THERE ARE FEWER SAMPLES THAN CO-ORDINATES

When there are fewer training examples, N, than point co-ordinates, $2n$, the eigenvectors of the $2n \times 2n$ covariance matrix $\mathbf{S}$ can be calculated from the eigenvectors of

a smaller $N \times N$ matrix derived from the same data. Because the eigenvector calculation time goes as the cube of the size of the matrix, this can give substantial time savings.

We have N examples, $\mathbf{x}_i$ $(i = 1, \ldots, N_s)$. Let $\mathbf{D}$ be a $2n \times N$ matrix with these as its columns;

$$\mathbf{D} = (\mathbf{x}_1\ \mathbf{x}_2\ \ldots\ \mathbf{x}_N). \tag{35}$$

We can write the covariance matrix, $\mathbf{S}$, as

$$\mathbf{S} = \frac{1}{N}\mathbf{D}\mathbf{D}^{\mathrm{T}}. \tag{36}$$

Let $\mathbf{T}$ be the $N \times N$ matrix

$$\mathbf{T} = \frac{1}{N}\mathbf{D}^{\mathrm{T}}\mathbf{D}, \tag{37}$$

and let $\mathbf{e}_i$ $(i = 1, \ldots, N)$ be the unit, orthogonal eigenvectors of $\mathbf{T}$ with corresponding eigenvalues γ_i:

$$\mathbf{T}\mathbf{e}_i = \gamma_i\mathbf{e}_i \quad (i = 1, \ldots, N). \tag{38}$$

Then

$$\frac{1}{N}\mathbf{D}^{\mathrm{T}}\mathbf{D}\mathbf{e}_i = \gamma_i\mathbf{e}_i. \tag{39}$$

Premultiplying by $\mathbf{D}$,

$$\frac{1}{N}\mathbf{D}\mathbf{D}^{\mathrm{T}}\mathbf{D}\mathbf{e}_i = \gamma_i\mathbf{D}\mathbf{e}_i \tag{40}$$

$$\mathbf{S}(\mathbf{D}\mathbf{e}_i) = \gamma_i(\mathbf{D}\mathbf{e}_i). \tag{41}$$

Thus if $\mathbf{e}_i$ is an eigenvector of $\mathbf{T}$, then $\mathbf{D}\mathbf{e}_i$ is an eigenvector of $\mathbf{S}$ and has the same eigenvalue. The N unit orthogonal eigenvectors of $\mathbf{S}$ are then $\mathbf{p}_i$ $(i = 1, \ldots, N)$, where

$$\mathbf{p}_i = \frac{1}{\sqrt{\gamma_i N}}\mathbf{D}\mathbf{e}_i \tag{42}$$

with corresponding eigenvalues $\lambda_i = \gamma_i$. The scaling factor in Eq. 42 is required to give the eigenvectors unit length. Mutual orthogonality is easily shown:

$$\begin{aligned}\mathbf{p}_i^{\mathrm{T}}\mathbf{p}_j &= \frac{1}{\gamma_i N}\mathbf{e}_i^{\mathrm{T}}\mathbf{D}^{\mathrm{T}}\mathbf{D}\mathbf{e}_j = \frac{1}{\gamma_i}\mathbf{e}_i^{\mathrm{T}}\mathbf{T}\mathbf{e}_j \\ &= \mathbf{e}_i^{\mathrm{T}}\mathbf{e}_j \quad \begin{matrix} = 1 & (i = j) \\ = 0 & (i \neq j) \end{matrix}.\end{aligned} \tag{43}$$

ACKNOWLEDGMENTS

This work was funded by the Science and Engineering Research Council under the Information Engineering Advanced Technology Programme (Project Number 3/2114). Tim Cootes is currently funded by an SERC Post-Doctoral Fellowship. The authors thank the other members of the Wolfson Image Analysis Unit particularly D. Bailes and A. Hill, for their help and advice, and the anonymous reviewers for their suggestions.

REFERENCES

1. R. T. Chin and C. R. Dyer, Model-based recognition in robot vision, *Comput. Surv.* **18,** 1986, 67–108.
2. W. E. L. Grimson, *Object Recognition by Computer: The Role of Geometric Constraints,* MIT Press, Cambridge, MA, 1990.
3. M. Kass, A. Witkin, and D. Terzopoulos, Snakes: Active contour models, in *Proceedings, First International Conference on Computer Vision,* pp. 259–268, IEEE Comput. Soc. Press, 1987.
4. G. E. Hinton, C. K. I. Williams, and M. D. Revow, Adaptive elastic models for hand-printed character recognition, in *Advances in Neural Information Processing Systems 4* (J. E. Moody, S. J. Hanson, and R. P. Lippmann, Eds.), Morgan Kauffmann, San Mateo, CA, 1992.
5. A. L. Yuille, D. S. Cohen, and P. Hallinan, Feature extraction from faces using deformable templates, *Int. J. Comput. Vision* **8,** 1992, 99–112.
6. P. Lipson, A. L. Yuille, D. O'Keeffe, J. Cavanaugh, J. Taaffe, and D. Rosenthal, Deformable templates for feature extraction from medical images, in *Proceedings of the First European Conference on Computer Vision* (O. Faugeras, Ed.), Lecture Notes in Computer Science, pp. 413–417, Springer–Verlag, Berlin/New York, 1990.
7. A. Hill and C. J. Taylor, Model based image interpretation using genetic algorithms, *Image Vision Comput.* **10,** 1992, 295–300.
8. A. Beinglass and H. J. Wolfson, Articulated object recognition, or: How to generalize the generalized Hough transform in *Proceedings, IEEE Computer Society Conference on Computer Vision and Pattern Recognition 1991,* pp. 461–466.
9. G. L. Scott, The alternative Snake—And other animals, in *Proceedings, 3rd Alvey Vision Conference, Cambridge, 1987,* pp. 341–347.
10. L. H. Staib and J. S. Duncan, Parametrically deformable contour models, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, 1989,* pp. 427–430.
11. H. I. Bozma and J. S. Duncan, Model-based recognition of multiple deformable objects using a game-theoretic framework, in *Information Processing in Medical Imaging—Proceedings of the 12th International Conference,* pp. 358–372, Springer-Verlag, Berlin/New York, 1991.
12. U. Grenander, Y. Chow, and D. M. Keenan, *Hands. A Pattern Theoretic Study of Biological Shapes,* Springer-Verlag, New York, 1991.
13. U. Grenander and M. I. Miller, Representations of knowledge in complex systems. *J. R. Stat. Soc. B,* in press.
14. C. Goodall, Procrustes methods in the statistical analysis of shape (with discussions), *J. R. Stat. Soc. B.* **53,** 1991, 285–339.
15. K. V. Mardia, J. T. Kent, and A. N. Walder, Statistical shape models in image analysis, in *Proceedings of the 23rd Symposium on the Interface, Seattle 1991,* pp. 550–557.
16. F. L. Bookstein, *Morphometric Tools for Landmark Data,* Cambridge Univ. Press, London/New York, 1991.
17. F. L. Bookstein, Principal warps: Thin-plate splines and the decomposition of deformations, IEEE Trans. Pattern Anal. Mach. Intell. **11,** 1989, 567–585.
18. A. Pentland, Automatic extraction of deformable part models, *Int. J. Comput. Vision* **13,** No. 2, 1990, 107–126.

19. A. Pentland and S. Sclaroff, Closed-form solutions for physically based modeling and recognition, *IEEE Trans. Pattern Anal. Mach. Intell.* **13,** 1991, 715–729.
20. D. Terzopoulos and D. Metaxas, Dynamic 3D models with local and global deformations: Deformable superquadrics, *IEEE Trans. Pattern Anal. Mach. Intell.* **13,** 1991, 703–714.
21. C. Nastar and N. Ayache, Fast segmentation, tracking and analysis of deformable objects, in *Proceedings, International Conference on Computer Vision, 1993,* pp. 275–279, IEEE Comput. Soc. Press, 1993.
22. P. Karaolani, G. D. Sullivan, K. D. Baker, and M. J. Baines, A finite element method for deformable models, in *Proceedings of the Fifth Alvey Vision Conference, Reading, 1989,* pp. 73–78.
23. P. Karaolani, G. D. Sullivan, and K. D. Baker, Active contours using finite elements to control local scale, in *Proceedings, British Machine Vision Conference 1992,* pp. 481–487, Springer-Verlag, Berlin/New York, 1992.
24. J. C. Gower, Generalized Procrustes analysis, *Psychometrika* **40,** 1975, 33–51.
25. R. A. Johnson and D. W. Wichern, *Multivariate Statistics, A Practical Approach,* Chapman & Hall, London/New York, 1988.
26. A. Hill, T. F. Cootes, and C. J. Taylor, A genetic system for image interpretation using flexible templates, in *British Machine Vision Conference,* Springer-Verlag, 1992.
27. A. Hill, C. J. Taylor, and T. Cootes, Object recognition by flexible template matching using genetic algorithms, in *Proceedings, European Conference on Computer Vision* (G. Sandini, Ed.), pp. 852–856, Springer-Verlag, Berlin/New York, 1992.
28. A. Hill, A. Thornham, and C. J. Taylor, Model-based interpretation of 3D medical images, in *Proceedings, British Machine Vision Conference 1993* (J. Illingworth, Ed.), Vol. 2, pp. 339–348, BMVA Press, 1993.
29. D. J. Burr, A dynamic model for image registration, *Comput. Graphics Image Process.* **15,** 1981, pp. 102–112.
30. S. Sclaroff and A. Pentland, A model framework for correspondence and description, in *Proceedings, International Conference on Computer Vision, 1993,* pp. 715–729, IEEE Comput. Soc. Press, 1993.
31. T. F. Cootes, D. H. Cooper, C. J. Taylor, and J. Graham, A trainable method of parametric shape description, *Image Vision Comput.* **10,** 1992, 289–294.
32. T. F. Cootes and C. J. Taylor, Active shape model search using local grey-level models: A quantitative evaluation, in *Proceedings, British Machine Vision Conference 1993* (J. Illingworth, Ed.), Vol. 2, pp. 639–648, BMVA Press, 1993.
33. T. F. Cootes, A. Hill, C. J. Taylor, and J. Haslam, The use of active shape models for locating structures in medical images, in *Proceedings, Information Processing for Medical Imaging* (H. H. Barrett and A. F. Gmitro, Ed.), pp. 33–47, Springer-Verlag, Berlin/New York, 1993.
34. T. F. Cootes, C. J. Taylor, A. Lanitis, D. H. Cooper, and J. Graham, Building and using flexible models incorporating grey-level information, in *Proceedings, International Conference on Computer Vision,* pp. 242–246, IEEE Comput. Soc. Press, 1993.
35. A. Hill, A. Thornham, and C. J. Taylor, Model-based interpretation of 3D medical images, in *Proceedings, British Machine Vision Conference, 1993* (J. Illingworth, Ed.), Vol. 1, pp. 339–348, BMVA Press, 1993.
36. A. Lanitis, C. J. Taylor, and T. F. Cootes, A generic system for classifying variable objects using flexible template matching, in *Proceedings, British Machine Vision Conference, 1993* (J. Illingworth, Ed.), Vol. 1, pp. 329–338, BMVA Press, 1993.
37. D. G. Lowe, Fitting parameterized three-dimensional models to images, *IEEE Trans. Pattern Anal. Mach. Intell.* **13,** 1991, 441–450.
38. J. A. Grogan, *Automated Analysis of Pedobarograph Images,* M. Sc. Thesis, Victoria University of Manchester, Oct. 1993.

25. **Building and using flexible models incorporating grey level information.** T.F. Cootes, C.J. Taylor, A. Lanitis, D.H. Cooper and J. Graham, *Proceedings of the fourth International Conference on Computer Vision, Berlin, 1993, pp 242-246.* doi:10.1109/ICCV.1993.378212

Building and Using Flexible Models Incorporating Grey-Level Information

T.F.Cootes, C.J.Taylor, A.Lanitis, D.H.Cooper and J.Graham

Department of Medical Biophysics, University of Manchester, England

Abstract

This paper describes a technique for building compact models of the shape and appearance of flexible objects seen in 2-D images. The models are derived from the statistics of sets of labelled images of example objects. Each model consists of a flexible shape template, describing how important points of the object can vary, and a statistical model of the expected grey levels in regions around each model point. Such models have proved useful in a wide variety of applications. We describe how the models can be used in local image search and give examples of their application.

1: Introduction

Image interpretation using rigid models is well established [1]. However, in many practical situations objects of the same class are not identical and rigid models are inappropriate. This is particularly true in medical applications, but many industrial applications also involve assemblies with moving parts, or components whose appearance can vary. In such cases flexible models, or deformable templates, can be used to allow for some degree of variability in the shape of the imaged objects.

In previous papers [2,3] we have shown how one can build models of the shape of deformable objects, and described a local search technique which allows initial estimates of the pose and shape parameters to be iteratively refined. We have also described how these methods can be combined with Genetic Algorithm search to achieve fully automatic image interpretation [4,5]. The shape models rely on representing objects by sets of labelled points; each point is placed on a particular part of the object. By examining the statistics of the positions of the labelled points a 'Point Distribution Model' is derived. The model gives the average positions of the points, and has a number of parameters which control the main modes of variation found in the training set.

Given such a model and an image containing an example of the object modelled, interpretation involves choosing values for each of the parameters so as to best fit the model to the image. We previously described a technique which allows an initial guess for the best shape, orientation, scale and position to be refined by comparing the hypothesised model example with image data and using differences between model and image to deform the shape [3]. The method has similarities with the Active Contour Models (or snakes) of Kass *et al* [6], but differs in that global shape constraints are applied; to make this distinction clear we have adopted the term Active *Shape* Models. The key point is that an instance of a model can only deform in ways found in its training set.

In this paper we consolidate and extend out previous work, showing how it is possible to model the grey levels expected at each point of the shape model, and how this can be used in Active Shape Model search. Our results demonstrate that the combination of the method for constructing shape models with the active matching technique provides a systematic and effective way of interpreting complex images, and that the addition of grey-level models gives an improvement in performance when compared to the earlier methods.

1.1: The problem to be addressed

In this paper we will only be considering the problem of attempting to fit a 2-D model to new but similar images of similar objects. Suppose we have a set of images containing examples of a flexible object which we wish to model, for instance a person's face. We wish to model the variation in shape of the face, and to fit such a shape model to a new image. We have chosen the face as an example as it is clearly flexible and complex, but the methods can be applied to problems in both industrial and medical fields.

2: Point Distribution Models

In [2] we describe how to build flexible shape models called Point Distribution Models. These are generated from examples of shapes, where each shape is represented by a set of labelled points. A given point corresponds to a particular location on each shape or object to be modelled. The example shapes are all aligned into a standard co-ordinate frame, and a principal component analysis is applied to the co-ordinates of the points. This produces the mean position for each of the points and a description of the main ways in which the points tend to move together. The model can be used to generate new shapes using the equation

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{Pb} \qquad (1)$$

where $\mathbf{x} = (x_0, y_0, \ldots x_{n-1}, y_{n-1})^T$

(x_k, y_k) is the k^{th} model point

$\bar{\mathbf{x}}$ represents the mean shape

$\mathbf{P}$ is a $2n \times t$ matrix of t unit eigenvectors

$\mathbf{b} = (b_1 \ldots b_t)^T$ – a set of shape parameters b_i

0-8186-3870-2/93 $3.00 © 1993 IEEE

If the shape parameters b_i are chosen such that the square of the Mahalanobis distance is limited,

$$D_m^2 = \sum_{k=1}^{t}\left(\frac{b_k^2}{\lambda_k}\right) \le D_{\max}^2 \qquad (2)$$

λ_k is the variance of parameter b_k in the original training set.

for some given D_{max} (typically 2.5 - 3.0), then the shape generated by (1) will be similar to those given in the original training set.

By choosing a set of shape parameters **b** for a Point Distribution Model, we define the shape of a model object in an object centred co-ordinate frame. We can then create an instance, **X**, of the model in the image frame by defining the position, orientation and scale:

$$\mathbf{X} = M(s,\theta)[\mathbf{x}] + \mathbf{X}_c \qquad (3)$$

where $\mathbf{X}_c = (X_c,\ Y_c,\ \ldots,\ X_c,\ Y_c)^T$

$M(s,\theta)[\ ]$ performs a rotation by θ and a scaling by s.

and (X_c, Y_c) is the position of the centre of the model in the image frame.

3: Examples of shape models

11 images of the face of a subject were taken in a normal office environment, and 169 labelled points were planted on each, giving a set of 11 training shapes. The method described in [2] was applied to the sets of points to build a flexible 2-D model of the face. Figure 1 shows the effects of varying the first four shape parameters. The first mode of variation, obtained by varying b_1 while keeping the other shape parameters zero, gives the effect of shaking the head, whilst raising and lowering the eyebrows. The second mode shows the transition between smiling and frowning. The third mode gives a nodding motion and the fourth raises the eyebrows. The further modes are more subtle variations of sets of features.

The first and third modes of variation of the face model are interesting in that they model in two dimensions the effects of small changes in three dimensional viewpoint.

The modelling technique has been successfully applied to a wide variety of examples from both industry and medicine, including resistors, hands [2], brake assemblies, heart chambers in echocardiograms [5] and the ventricles of the brain in MR images [5].

4: Modelling grey level appearance

We wish to use our models for locating examples of objects in new images. For this purpose, not only shape, but also grey-level appearance is important. We account for this by examining the statistics of the grey levels in regions around each of the labelled model points. Since a given point corresponds to a particular part of the object, the grey-level patterns about that point in images of different examples will often be similar. The ability of a system to locate model points in images should be improved by incorporating each point's grey-level environment into the model.

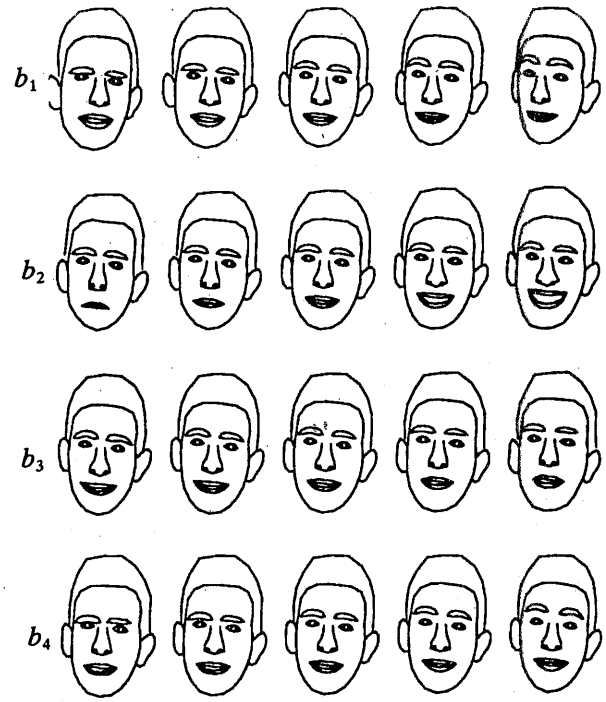


Figure 1 : Effects of varying the first four parameters of the face model individually.

We need to associate an orientation with each point of our shape model in order to align the region correctly. A convenient way to do this is to define an orientation with respect to nearby model points. For instance, if the points lie along a boundary, we can choose to align the region with the normal to the boundary.

Although in general we can consider a region of any shape around each point, we will concentrate on one-dimensional profiles normal to curves passing through each point.

For every point i in each image, j, we can extract a profile, g_{ij}', of length n_p pixels, centred at the point. We choose to sample the derivative of the grey levels along the profile in the image and normalise. This gives invariance to uniform scaling of the grey levels and the addition of a constant.

If the profile runs from p_{start} to p_{end} and is of length n_p pixels, the k^{th} element of the derivative profile is

$$g_{ijk}' = I_j(\mathbf{y}_{k+1}) - I_j(\mathbf{y}_{k-1}) \qquad (4)$$

where $\mathbf{y}_k$ is the k^{th} point along the profile :

$$\mathbf{y}_k = \mathbf{p}_{start} + \frac{k-1}{n_p - 1}(\mathbf{p}_{end} - \mathbf{p}_{start}) \qquad (5)$$

and $I_j(y_k)$ is the grey level in image j at that point. We then normalise this profile,

$$\mathbf{g}_{ij} = \frac{\mathbf{g}_{ij}'}{\sum_{k=1}^{n_p} |g_{ijk}'|} \qquad (6)$$

The normalised derivative profile tends to be more invariant to changes in the image caused by variations in lighting than a simple grey-level profile.

For each point, i, we can calculate a mean normalised derivative profile,

$$\bar{\mathbf{g}}_i = \frac{1}{N_s}\sum_{j=1}^{N_s}\mathbf{g}_{ij} \qquad (7)$$

We can then calculate an n_p x n_p covariance matrix, $\mathbf{S}_{\mathbf{g}i}$, giving us a statistical description of the expected profiles about the point. Alternatively a simpler model can be generated by calculating the standard deviation of each profile model pixel about the mean, $\mathbf{g}_{i\sigma}$.

$$(g_{i\sigma k})^2 = \frac{1}{N_s}\sum_{j=1}^{N_s}(g_{ijk}' - \bar{g}_{ik})^2 \qquad (8)$$

(This is the leading diagonal of $\mathbf{S}_{\mathbf{g}i}$.)

5: Using Point Distribution Models in image search – Active Shape Models

Having generated a flexible model and a description of the grey levels about each model point we would like to find new examples of the modelled object in images. In general, a procedure for achieving this has two stages :

- A number of hypotheses are made, giving approximate locations of the model points
- Each of the hypotheses is refined and the best chosen.

The initial hypotheses take the form of estimates for the position of the centre of the object, its orientation, scale and the shape parameters required to fit the model to the image. Such hypotheses can be obtained from cue generators or by using suitable search techniques. Hill *et al* [4,5] describe methods for finding flexible objects in images which use Genetic Algorithms to generate a set of hypotheses quite rapidly.

In this section we describe an iterative method for refining hypotheses so as to give a better match of the model example to the object in the image. The approach is as follows :

i) Examine a region of the image around each point to calculate the displacement of the point required to move it to a better location.
ii) From these proposed displacements calculate adjustments to the pose and scale and to the shape parameters of the Point Distribution Model.
iii) Update the model parameters; by enforcing limits on the shape parameters, global shape constraints can be applied ensuring the shape of the model instance remains similar to those of the training set.

The procedure is repeated until no significant changes result.

5.1: Calculating A Suggested Movement For Each Model Point

Given an initial estimate of the positions of a set of model points which we are attempting to fit to an image object we need to estimate a set of adjustments which will move each point toward a better position. In the case in which the model points represent the boundary of an object (Figure 2) the required adjustments will move them toward the edges of the image object. If we have profile models for each point, the search involves finding nearby regions which best match the profile models. At a particular model point we extract a derivative profile, $\mathbf{g}$, from the current image of some length, l ($>n_p$), centred at the point and aligned parallel to the orientation we have defined at that point (for instance, normal to the boundary). We then run the profile model along this sampled profile and find the point at which the model best matches.

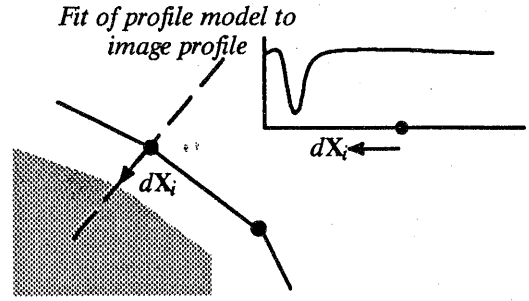


Figure 2 : Suggested movement of point is along normal to boundary, in a direction towards the point at which the profile model best fits the profile sampled from the image.

Given a sampled derivative profile the fit of the model at a point d pixels along it is calculated as follows;

$$f_{prof}(d) = (\mathbf{h}(d) - \bar{\mathbf{g}})^T \mathbf{S}_{\mathbf{g}}^{-1}(\mathbf{h}(d) - \bar{\mathbf{g}}) \qquad (9)$$

where $\mathbf{h}(d)$ is sub-interval of $\mathbf{g}$ of length n_p pixels centred at d, normalised using (6). This is the Mahalanobis distance of the sample from the mean grey model, and is proportional to the log of the probability of obtaining $\mathbf{h}(d)$ from the measured distribution of grey-levels.

If the model consists of only the mean $\bar{\mathbf{g}}$ and standard deviation $\mathbf{g}_\sigma$ of each pixel along the profile, then we can use

$$f_{prof}(d) = (\mathbf{h}(d) - \bar{\mathbf{g}})^T \mathbf{S}_\sigma^{-2}(\mathbf{h}(d) - \bar{\mathbf{g}}) \qquad (10)$$

where $\mathbf{S}_\sigma$ is a diagonal matrix whose leading diagonal is formed from the elements of $\mathbf{g}_\sigma$.

In both cases the value of f_{prof} decreases as the fit improves. The point of best fit is thus the point at which $f_{prof}(d)$ is a minimum.

Suppose d_{best} is the distance along the sampled profile from the model point to the point of best fit. We choose a displacement for the model point of $d\mathbf{X}$ which is parallel to the profile, in the direction of the point of best fit, with magnitude

$$\begin{aligned} |d\mathbf{X}| &= 0 && \text{if } |d_{best}| \le \delta \\ |d\mathbf{X}| &= 0.5d_{best} && \text{if } \delta < |d_{best}| < d_{\max} \\ |d\mathbf{X}| &= 0.5d_{\max} && \text{if } d_{\max} \le |d_{best}| \end{aligned} \qquad (11)$$

For some δ and d_{max}. (We use $\delta = 0.5$ pixels, d_{max} = 8 pixels in the following experiments).

An alternative approach is to generate potential images such as to those described by Kass *et al* [6], possibly one for each model point, describing how likely each point in the image is to be the model point. Adjustments

to the position of each point can then be derived from the gradient of the potential image at the current estimate of the point's position.

5.2: Calculating the adjustments to the pose and shape parameters

A set of adjustments can be calculated, one for each point of the shape. We denote such a set as a vector $d\mathbf{X}$;

$$d\mathbf{X} = (dX_0,\ dY_0,\ \ldots,\ dX_{n-1},\ dY_{n-1})^T$$

Suppose the current estimate of the model is centred at (X_c, Y_c) with orientation θ and scale s. We aim to adjust the pose and shape parameters to move the points from their current locations, $\mathbf{X}$, in the image frame to be as close to the suggested new locations $(\mathbf{X} + d\mathbf{X})$ as can be arranged whilst still satisfying the shape constraints of the model. We calculate the pose adjustment by finding the transform $(dX_c\ ,\ dY_c),\ d\theta\ ,\ (1+ds)$ which best maps the current set of points, $\mathbf{X}$, onto the set of points given by $(\mathbf{X} + d\mathbf{X})$. This can be done by a weighted least squares fit [2,3].

Having adjusted the pose variables there remain residual adjustments which can only be satisfied by deforming the shape of the model. We wish to calculate the adjustments, $d\mathbf{x}$, to the original model points in the local co-ordinate frame required to cause the scaled, rotated and translated points $\mathbf{X}$ to move by $d\mathbf{X}$ when combined with the new scale, rotation and translation parameters.

In [3] we show that

$$d\mathbf{x} = M((s(1+ds))^{-1}, -(\theta + d\theta))[d\mathbf{X}_r] - \mathbf{x} \qquad (12)$$

$$(\ d\mathbf{X}_r = M(s,\theta)[\mathbf{x}] + d\mathbf{X} - d\mathbf{X}_c\ )$$

To calculate the adjustments to the shape parameters, $d\mathbf{b}$, required to move the points by $d\mathbf{x}$ we solve

$$(\mathbf{P}^T\mathbf{W}_s)d\mathbf{x} = (\mathbf{P}^T\mathbf{W}_s\mathbf{P})d\mathbf{b} \qquad (13)$$

where $\mathbf{W}_s$ is a diagonal matrix of weights, one for each co-ordinate of each point.

This is a set of t linear equations in the t variables of $d\mathbf{b}$, and can be solved using standard matrix algebra.

We have found a useful weighting scheme to be

$$w_i = (2 + |d\mathbf{X}_i|^2)^{-1} \qquad (14)$$

In the special case in which all weights are set to unity, $\mathbf{W}_s = \mathbf{I}$, Eq.(13) simplifies to

$$d\mathbf{b} = \mathbf{P}^T d\mathbf{x} \qquad (15)$$

5.3: Updating the Pose and Shape Parameters

The equations above allow us to calculate changes to the pose variables, $dX_c\ ,\ dY_c,\ d\theta$ and ds, and adjustments to the shape parameters $d\mathbf{b}$ required to improve the match between an object model and image evidence. When we apply these to update the parameters in an iterative scheme we can ensure that the model only deforms into shapes consistent with the training set by placing limits on the values of b_k. As mentioned above, a shape can be considered acceptable if the Mahalanobis distance D_m is less than a suitable constant, D_{max}, for instance 3.0 (See Eq. 2). That is to say, the vector $\mathbf{b}$ should lie within a hyper-ellipsoid about the origin. If updating $\mathbf{b}$ leads to an implausible shape, ie $D_m > D_{\max}$ and the point lies outside the ellipsoid, $\mathbf{b}$ can be re-scaled to lie on the closest point of the allowed volume using

$$b_k \rightarrow b_k\ D_{\max}/(D_m) \quad (k = 1..t) \qquad (16)$$

Note that we have already applied implicit limits of zero to the weights of the eigenvectors truncated from our representation (ie $b_i = 0\ \forall\ i > t$). Once the parameters have been updated, and limits applied where necessary, a new example can be calculated, and new suggested movements derived for each point. The procedure is repeated until no significant change results.

6: Examples of ASMs in Action

To date we have implemented the above techniques using the simpler grey level profile models (mean and standard deviation only). We present some examples of using the flexible models in image search.

6.1: Face Images

Having built the face shape model (Section 3) we collected profiles of length 7 pixels for each of the 169 points of the model from each of our small sample set of 11 images. From these we calculated 169 profile models. Figure 3 shows a new image of the same person taken several weeks after the training set, with a different background. The initial estimate of the model point positions is overlayed. This was obtained by choosing a set of pose parameters and setting all the shape parameters to zero (corresponding to the mean model shape). In practice such an initial estimate can be obtained from cues or a global search [5]. Figure 4 shows the model after 100 iterations of the Active Shape Model. As can be seen, a good fit was obtained to most features except the outer edges of the hair, which had grown since the training images were taken. In this example the weights used during the calculation of pose and shape adjustment were calculated using (14). The profiles sampled from the new image at each point were of length 30 pixels. Each iteration took about 0.6 seconds on a Sun Sparc 2 workstation. When the the same experiment was run using no weightings a slightly poorer result was obtained, but each iteration took only 0.3 seconds.

6.2: Industrial Inspection

Figure 6 shows an image of a car brake assembly overlayed with a model used to locate its components. The model contains 308 points and has 14 degrees of freedom representing the relative motions of the components of the assembly. The model's position has been found using a coarse exhaustive search of the first two shape parameters followed by 12 iterations of an ASM.

7: Improvements Given by Grey-Profiles

In previous work [3] we described ASMs which moved their model points towards strong edges in the image. We wished to assess what improvements are gained by using the more detailed grey-level profile models. To do this we took a set of face images with known labelled points and ran ASMs starting at a set of systematically displaced positions. We measured the mean distance of the model points from the known image points after each iteration for ASMs using grey-level profiles and for ASMs whose points were attracted to strong nearby edges. Figure 5 shows the averaged results, showing that the profile based system converges faster and to a better

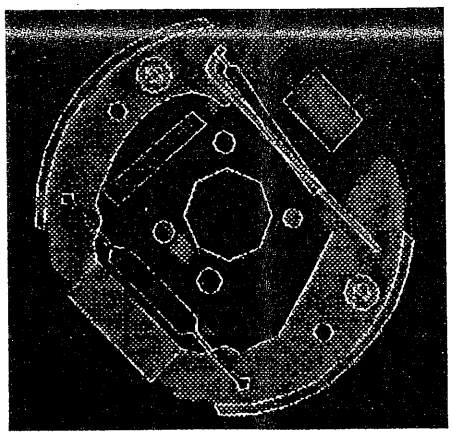


Figure 6 : Brake image with components located.

overall result. The grey-level profiles give better results as they are less likely to be confused by noise or clutter than simply searching for strong edges.

8: Conclusions

The methods we describe allow us to take a set of example images of one or more objects whose shape can vary, build a flexible shape model and use that model to locate new examples in new images given an initial approximate location. Now that the programs are in place the majority of the effort goes into capturing the example images and marking labelled points on each of them.

By varying the parameters of each model we can generate new examples similar to those in the training set, making them ideal for generate-and-test search strategies. By enforcing limits on the shape parameters during the local search we enforce global shape constraints on the model.

Profile models are able to represent the grey level environment around each point, reducing the deleterious effects of clutter and noise during image search. Results can be further improved by using suitable weighting during the parameter adjustment calculations of the Active Shape Models.

The models can represent the varying shapes of a wide variety of different objects in both industrial and medical applications. The techniques have great potential for use in many image analysis domains.

Acknowledgements

This work is funded by the SERC under the IEAT Programme (Project Number 3/2114). The authors would like to thank the other members of the Wolfson Image Analysis Unit for their help and advice.

9: References

[1] W.E.L. Grimson, *Object Recognition by Computer : The Role of Geometric Constraints*. The MIT Press, Cambridge Massachusetts, 1990.

[2] T.F.Cootes, C.J.Taylor, D.H.Cooper and J.Graham, Training Models of Shape from Sets of Examples. in *Proc. British Machine Vision Conference*. Springer-Verlag, 1992, pp.9-18.

[3] T.F.Cootes, C.J.Taylor, Active Shape Models - 'Smart Snakes'. in *Proc. British Machine Vision Conference*. Springer-Verlag, 1992, pp.266-275.

[4] A. Hill,C.J. Taylor and T.Cootes, Object Recognition by Flexible Template Matching using Genetic Algorithms, in *Proc. European Conference on Computer Vision* (G.Sandini. Ed.). pp. 852-856, Springer-Verlag, 1992.

[5] A. Hill, T.F. Cootes and C.J. Taylor, A Generic System for Image Interpretation Using Flexible Templates, in *Proc. British Machine Vision Conference*. Springer-Verlag, 1992, 276-285.

[6] M. Kass, A. Witkin and D. Terzopoulos , Snakes: Active Contour Models, in *Proc. First International Conference on Computer Vision*, pp 259-268 IEEE Computer Society Press, 1987.

Figure 3 : Initial estimate of model point positions overlaying new face image.

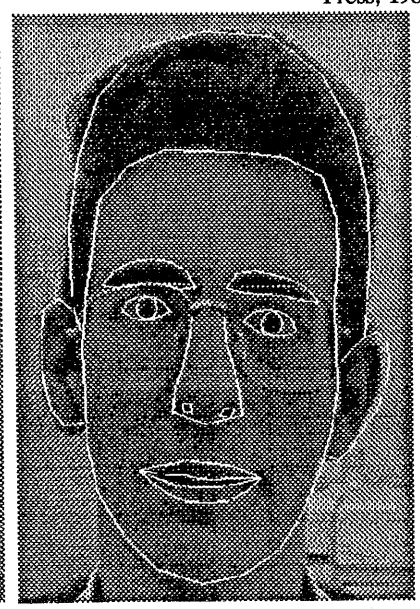


Figure 4 : Model point positions after 100 iterations.

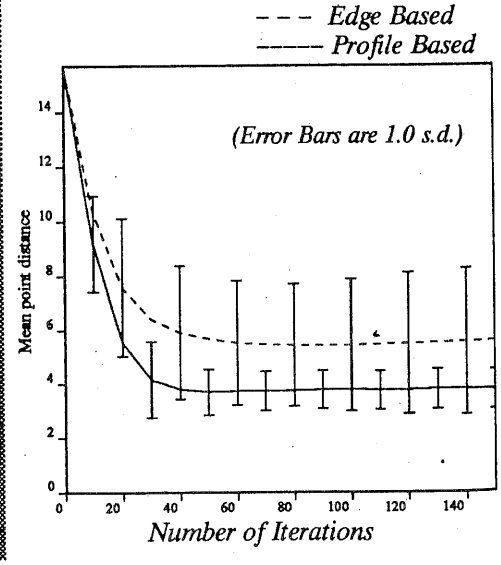


Figure 5 : Comparison between edge based and profile model based ASM.

26. **Image search using trained flexible shape models.** T.F. Cootes, D.H.Cooper, C.J. Taylor and J. Graham, *Journal of Applied Statistics, 21 (1-2) 111-139, 1994. doi:* 10.1080/757582971

CHAPTER 7

Image search using trained flexible shape models

T. F. COOTES, C. J. TAYLOR, D. H. COOPER & J. GRAHAM, *Department of Medical Biophysics, University of Manchester*

SUMMARY *This paper describes a technique for building compact models of the shape and appearance of flexible objects seen in two-dimensional images. The models are derived from the statistics of sets of images of example objects with 'landmark' points labelled on each object. Each model consists of a flexible shape template, describing how the landmark points can vary, and a statistical model of the expected grey levels in regions around each point. Such models have proved useful in a wide variety of applications. We describe how the models can be used in local image search and give examples of their application.*

1 Introduction

Image interpretation using rigid models is well established (Chin & Deyer, 1986; Grimson, 1990). However, in many practical situations, objects of the same class are not identical and rigid models are inappropriate. This is particularly true in medical applications, but many industrial applications also involve assemblies with moving parts, or components whose appearance can vary. In such cases, flexible models, or deformable templates, can be used to allow for some degree of variability in the shape of the imaged objects.

This paper contains a review of our work on a method of building and using models of objects whose shape can vary (Cootes *et al.*, 1992a, b, 1993a, b). The models are able to capture the natural variability within a class of shapes and can be used in image search to find examples of the objects modelled. Each object is represented by a set of points. The points can represent the boundary, internal features or even external features, such as the centre of curvature of a concave

section of boundary. Points are placed systematically on each of a training set of examples of the object, and the sets of points are aligned so as to minimize the variance in distance between equivalent points. By examining the statistics of the positions of the labelled points, a 'Point Distribution Model' is derived. The model gives the average positions of the points and has a number of parameters which control the main modes of variation found in the training set.

Given such a model and an image containing an example of the object modelled, interpretation involves choosing values for each of the model parameters so as to best fit the model to the image. We previously described a technique which allows an initial guess for the best shape, orientation, scale and position to be refined by comparing the hypothesized model example with image data and using differences between model and image to deform the shape (Cootes *et al.*, 1992b). The method has similarities to the active contour models (or snakes) of Kass *et al.* (1987) but differs in that global shape constraints are applied; to make this distinction clear, we have adopted the term 'Active Shape Models'. The key point is that an instance of a model can only deform in ways found in its training set.

In the remainder of the paper, we present a review of some of the relevant literature, describe the modelling method and show examples of trained models. We describe how it is possible to model the grey levels expected at each point of the shape model, and ow this can be used in active shape model search. We also show how one can associate weights with each point, specifying how important it is that the method should fit the model to the image data at that point. Our results demonstrate that the combination of the method for constructing shape models with the active matching technique provides a systematic and effective way of interpreting complex images.

2 Background

2.1 Image analysis literature

Many people have used flexible models or deformable templates to aid image interpretation. Such models usually have a number of parameters to control the shape and orientation of all or part of the model. Yuille *et al.* (1989, 1992) and Lipson *et al.* (1990) used 'hand-crafted' models of faces and vertebrae. These have to be individually tailored for each application. Kass *et al.* (1987) described flexible contour models which are attracted to image features. They are usually free to take almost any smooth boundary with few constraints on their overall shape. The method of fitting, that of using image evidence to apply forces and then minimizing an energy function, is an effective one. Hinton *et al.* (1992) described a type of spline snake governed by a number of control points which have preferred 'home' locations to give the snake a particular default shape. Deformations are caused by moving the control points away from their 'home' locations. Although the average shape of an object is represented, the modes of shape variation are only coarsely defined by the number and position of the control points.

Scott (1987), Staib and Duncan (1989) and Bozma and Duncan (1991) all used closed contour shape models based on expansions of trigonometric functions for the interpretation of medical images. However, the basis functions used are unlikely to give the most compact representation of shape and shape variability. Although recording the distributions of each parameter over a set of examples leads to a broad description of a class of shapes, without knowledge of how the parameters tend to

vary together over the training set, the models lack specificity; i.e. many examples can be generated which are not 'legal' generalizations of the class of shapes.

Karaolani *et al.* (1989) and Pentland and Sclaroff (1991) worked with finite element models of flexible objects. Like the trigonometric models, the basis functions used in finite element models are not necessarily the most effective at describing the variability occurring in a particular class of shapes, so may not give the most specific or compact model.

2.2 Statistical shape analysis

Statistical shape analysis is concerned with the analysis and classification of sets of shapes, often represented by sets of 'landmark' points marked systematically on each shape. Goodall (1991) discussed the registration of shapes in arbitrary dimensions, and the use of Procrustes Analysis for estimating the mean shape, the covariances between landmark points and for assessing the differences between sets of shapes. Bookstein (1989, 1991) has applied statistical techniques to determine relationships between shape and other variables for morphometric analysis. He showed how a set of landmarks in two dimensions, represented as a set of complex numbers $(z_1, \ldots, z_n)$, can be converted into Bookstein coordinates using

$$u_i = \frac{z_{i+2} - z_1}{z_2 - z_1}$$

Having registered sets of landmark points from different samples in this space, Bookstein then applied multivariate analysis, including principal component analysis, on the coordinates to examine the data and study shape change. He also discussed the use of thin-plate splines as interpolation functions to study deformations and to aid in the extraction of features.

Kent (1994) described the use of the complex Bingham distribution for shape analysis in two dimensions. Any set of n points can be scaled and represented by a point z on the unit complex sphere in n dimensions. The complex Bingham distribution is given by

$$f(z) = c(\mathbf{A})^{-1} \exp(z^{\star} \mathbf{A} z)$$

where $z^{\star} = \bar{z}^{\mathrm{T}}$ is the complex conjugate of the transpose of z, $\mathbf{A}$ is $(k \times k)$ Hermitian and $c(\mathbf{A})$ is a normalizing constant. He demonstrated how the parameters of the distribution can be estimated from a set of sample shapes, how it can be used to analyze shape change, and discussed the use of principal component analysis on data in such a coordinate system. Mardia and Dryden (1989, 1991) and Goodall and Mardia (1991) described various shape distributions which can be used to analyze sets of shapes.

Grenander *et al.* (1991) and Mardia *et al.* (1991) described statistical models of closed boundary shapes. These are represented as sets of points in the complex plane. The points are able to vary about a mean template in a $2n$-dimensional normal distribution with covariance $\mathbf{S}$. They suggested possible approximations for S using a first-order conditional auto-regressive (CAR) model or Toeplitz covariance matrices, and showed how such models can be used to reconstruct an outline from a degraded image. In the work which follows, we use a similar underlying model. However, we use principal component analysis to simplify the structure of the covariance matrix, which allows us to represent shapes using small numbers of parameters.

3 Modelling object shape: point distribution models

3.1 The problem to be addressed

Suppose we wish to derive a model to represent the shapes of resistors as they appear on printed circuit boards, as shown in Fig. 1 Different examples of resistors have sufficiently different shapes that a rigid model would not be appropriate. Figure 2 shows some examples of resistor boundaries which were obtained from back-lit images of individual resistors. Our aim is to build a model which describes both typical shape and allowed variability, using the examples in Fig. 2 as a training set.

3.2 Labelling the training set

To model the shape, we represent it by a set of points. For the resistors, we have chosen to place points around the boundary, as shown in Fig. 3. This must be done for each shape in the training set. The labelling of the points is important. Each labelled point represents a particular part of the object or its boundary. For instance, in the resistor model, points 0 and 31 always represent the ends of a wire, points 3, 4 and 5 represent one end of the body of the resistor, and so on. The method works by modelling how different labelled points tend to move together as the shape varies. If the labelling is incorrect, with a particular point placed at different sites on each training shape, the method will fail to capture the shape variability.

Bookstein (1989, 1991) labelled the significant points in images of biological and medical specimens to examine and measure shape changes and to correlate these with other factors. We do this here to capture the shape constraints which may be used to construct plausible new examples of the shape from a model. These examples could be used in a search of an image. Bookstein called these representative points "landmark points" and described them in terms of their usefulness. For our purposes, these can be reduced to three different types:

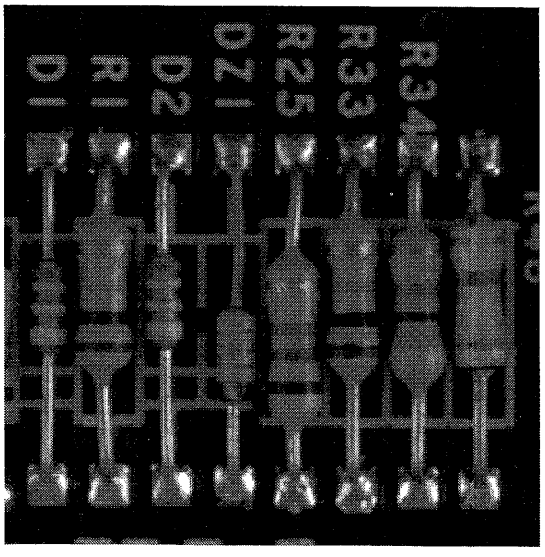


FIG. 1 Image of printed circuit board, showing examples of resistors.

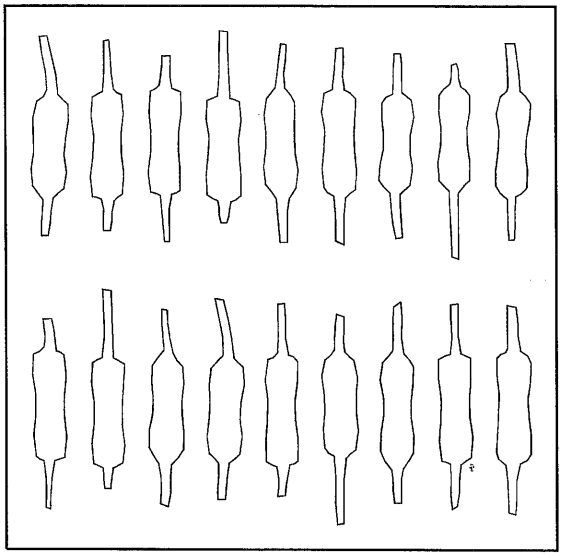


FIG. 2. Examples of resistor shapes from a training set.

(1) points marking parts of the object with particular application-dependent significance, such as the centre of an eye in the model of a face or sharp corners of a boundary;
(2) points marking application-independent locations, such as the highest point on an object in a particular orientation or curvature extrema;
(3) other points which can be interpolated from points of type (1) and (2), such as points marked at equal distances round a boundary between two type (1) landmarks.

On the resistor shown in Fig. 3, points 0, 3, 5, 10, etc. mark easily identified features, so ar points of type (1). The other points are equally spaced along the boundaries between, so are of type (3).

Landmark points of type (1) are preferable to those of type (2), since they are generally easier to identify precisely. However, points of types (2) and (3) are almost always necessary to define the boundary of a flexible shape in sufficient detail.

Landmark points can be used to describe single objects or sets of spatially related objects. We handle this by symbolically associating boundary segments with appropriate pairs of landmark points. Although we have implemented and investigated the use of interpolating splines to generate boundaries using a minimal set of

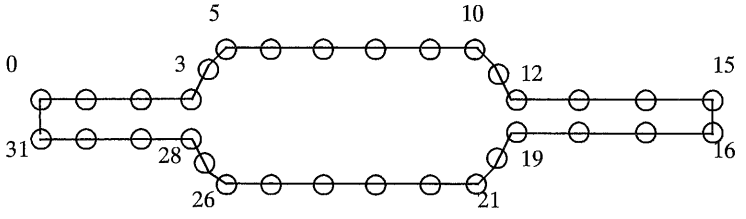


FIG. 3. 32-point model of the boundary of a resistor.

landmarks, we find that simply using enough type (3) points to describe the curve with sufficient accuracy is as effective and is computationally more efficient.

3.3 Aligning the set of training shapes

Our modelling method works by examining the statistics of the coordinates of the labelled points over the training set. To be able to compare equivalent points from different shapes, they must be aligned in the same way with respect to a set of axes. If they are not, we would not be comparing like with like, and any statistics derived would be meaningless. We achieve the required alignment by scaling, rotating and translating the training shapes so that they correspond as closely as possible. The measure of closeness is a weighted sum of squares of the distances betwen equivalent points on different shapes. This is a weighted version of the Procrustes method (see Gower, 1975; Goodall, 1991).

We first consider aligning a pair of shapes. Let $\mathbf{x}_i$ be a vector describing the n points of the ith shape in the set, such that

$$\mathbf{x}_i = (x_{i0}, y_{i0}, x_{i1}, y_{i1}, \ldots, x_{ik}, y_{ik}, \ldots, x_{in-1}, y_{in-1})^{\mathrm{T}}$$

Let $M(s, \theta)$ be a rotation by θ and a scaling by s. Given two similar shapes $\mathbf{x}_i$ and $\mathbf{x}_j$, we can choose θ_j, s_j and a translation (t_{xj}, t_{yj}) mapping $\mathbf{x}_j$ on to $M(s_j, \theta_j)[\mathbf{x}_j] + \mathbf{t}_j$ so as to minimize the weighted sum

$$E_j = \{\mathbf{x}_i - M(s_j, \theta_j)[\mathbf{x}_j] - \mathbf{t}_j\}^{\mathrm{T}} \mathbf{W} \{\mathbf{x}_i - M(s_j, \theta_j)[\mathbf{x}_j] - \mathbf{t}_j\}$$

where

$$M(s, \theta)\begin{bmatrix} x_{jk} \\ y_{jk} \end{bmatrix} = \begin{pmatrix} (s\cos\theta)x_{jk} - (s\sin\theta)y_{jk} \\ (s\sin\theta)x_{jk} + (s\cos\theta)y_{jk} \end{pmatrix}$$

$$\mathbf{t}_j = (t_{xj}, t_{yj}, \ldots, t_{xj}, t_{yj})^{\mathrm{T}}$$

and $\mathbf{W}$ is a diagonal matrix of weights for each point. Details are given in the appendix.

The weights can be chosen to give more significance to those points which tend to be most 'stable' over the set—i.e. those which move about least with respect to the other points ina shape. We have used a weight matrix defined as follows: let R_{kl} be the distance between points k and l in a shape; let $V_{R_{kl}}$ be the variance in this distanced over the set of shapes; we can choose a weight w_k for the kth point using

$$w_k = \left(\sum_{l=0}^{n-1} V_{R_{kl}} \right)^{-1}$$

If a point tends to move around a lot with respect to the other points in the shape, the sum of the variances will be large and a low weight will be given. However, if a point tends to remain fixed with respect to the others, the sum of the variances will be small, a large weight will be given, and matching such points in different shapes will be a priority.

We use the following algorithm to align all the shapes:

- rotate, scale and translate each shape to align with the first shape in the set;
- repeat

—calcualte the mean shape from the aligned shapes;
—normalize the orientation, scale and origin of the current mean to suitable defaults;
—realign every shape with the adjusted version of the current mean;

- until the process converges.

Normalizing the mean to a default scale and orientation during each iteration is required to ensure that the algorithm converges. Without this, there are in effect $4(N_s-1)$ constraints on $4N_s$ variables (θ, s, $\mathbf{t}_x$ and $\mathbf{t}_y$ for each of the N_s shapes) and the algorithm is ill-conditioned—the mean will shrink, rotate or slide off to infinity.

Constraints on the orientation and scale of the mean allow the equations to have a unique solution. Either the mean is scaled, rotated and translated, so it matches the first shape, or an arbitrary default setting can be used, such as choosing an origin at its centre of gravity, an orientation so that a particular part of the shape is at the top, and a scale so that the distance between two selected points is one unit.

It should be noted, however, that normalizing the current mean shape and then aligning the shapes to match is not the same as normalizing each individual shape. If every shape were normalized in scale by setting the distance between two particular points to be one unit (as with Bookstein coordinates), artificial correlations may be forced upon the set, distorting the model. However, if the mean shape is scaled in this way, the other aligned shapes will have a similar scale, but their sizes and point positions will be chosen so as to best match the mean, rather than rigidly imposed. This leads to better models.

The convergence condition in the alignment procedure can be tested by examining the average difference between the transformations required to align each shape to the recalculated mean and the identity transformation. Here, Kent (1991, 1994) showed that the same alignment can also be found using the complex Bingham distribution. In its unweighted form, the argument proceeds as follows. First, given a shape k, translate it and scale it so that its centre of gravity is at the origin and the sum of squared distances from any point to the origin is one unit. Thus, the shape can be represented as a vector of n complex numbers, $\mathbf{z}_k=(z_{k1}\cdots z_{kn})^{\mathrm{T}}$, with

$$\sum_{j=1}^{n} z_{kj}=0, \qquad \sum_{j=1}^{n} |z_{kj}|^2=1$$

It can then be shown that the mean shape μ is the dominant eigenvector of $\Sigma\mathbf{z}_k\mathbf{z}_k^*$, and that any shape $\mathbf{z}_k$ can be optimally aligned with the mean by multiplying by $r_k e^{i\theta_k}$, where $\theta_k=\arg(\mathbf{z}_k^*\mu)$ and, $r=|\mathbf{z}_k^*\mu|$. Bookstein (1989, 1991), Goodall (1991), Grenander *et al.* (1991) and Mardia *et al.* (1991) also discussed methods for constructing the average shape. All these methods will produce similar results when deviations from the mean are small.

3.4 Capturing the statistics of a set of aligned shapes

In Fig. 4, the coordinates of some of the vertices of the aligned resistor shapes are plotted, with the mean shape overlayed. It can be seen that some of the vertices show little variability over the training set, while others form more diffuse 'clouds'. The point distribution model (PDM) seeks to model the variation of the coordinates within these clouds.

Once a set of N_s aligned shapes is available, the mean shape and variability can be found. The mean shape $\bar{\mathbf{x}}$ is calculated using

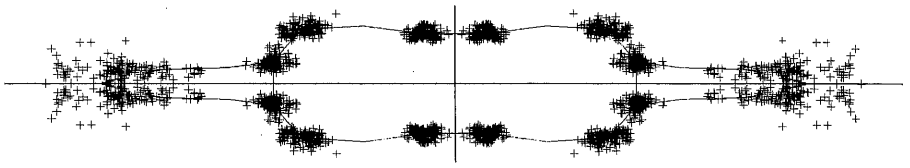


FIG. 4. Scatter of some points from aligned set of resistor shapes, with the mean shape overlayed.

$$\bar{\mathbf{x}}=\frac{1}{N_s}\sum_{i=1}^{N_s}\mathbf{x}_i$$

The modes of variation, i.e. the way in which the points of the shape tend to move together, can be found by applying principal component analysis to the deviations from the mean as follows (Johnson & Wichern, 1988). For each shape in the training set, we calculate its deviation from the mean $\mathrm{d}\mathbf{x}_i$, where

$$\mathrm{d}\mathbf{x}_i=\mathbf{x}_i-\bar{\mathbf{x}}$$

We can then calculate the $2n\times 2n$ covariance matrix $\mathbf{S}$ using

$$\mathbf{S}=\frac{1}{N_s}\sum_{i=1}^{N_s}\mathrm{d}\mathbf{x}_i\mathrm{d}\mathbf{x}_i^{\mathrm{T}}$$

The modes of variation of the points of the shape are described by $\mathbf{p}_k(k=1,\ldots,2n)$, the unit eigenvectors of $\mathbf{S}$, such that

$$\mathbf{S}\mathbf{p}_k=\lambda_k\mathbf{p}_k$$

where λ_k is the kth eigenvalue of $\mathbf{S}$, with $\lambda_k\geqslant\lambda_{k+1}$, and

$$\mathbf{p}_k^{\mathrm{T}}\mathbf{p}_k=1$$

It can be shown that the eigenvectors of the covariance matrix corresponding to the largest eigenvalues describe the most significant modes of variation in the variables used to derive the covariance matrix, and that the proportion of the total variance explained by each eigenvector is equal to the corresponding eigenvalue (Johnson & Wichern, 1988).

Most of the variation can usually be explained by a small number of $t(<2n)$ of modes. One method for calculating t is to choose the smallest number of modes such that the sum of their variances explains a sufficiently large proportion of λ_{T}, the total variance of all the variables, where

$$\lambda_{\mathrm{T}}=\sum_{k=1}^{2n}\lambda_k$$

The kth eigenvector affects point l in the model by moving it along a vector parallel to $(\mathrm{d}x_{kl},\mathrm{d}y_{kl})$, which is obtained from the lth pair of elements in $\mathbf{p}_k$, such that

$$(\mathrm{d}x_{k0},\mathrm{d}y_{k0},\ldots,\underline{\mathrm{d}x_{kl},\mathrm{d}y_{kl}},\ldots,\mathrm{d}x_{kn-1},\mathrm{d}y_{kn-1})$$

Combinations of vectors—one for each mode—move the modelled landmark points around in the regions of the 'clouds' of scattered points from the aligned training set. Any shape in the training set can be approximated using the mean shape and a weighted sum of these deviations obtained from the first t modes

$$\mathbf{x}=\bar{\mathbf{x}}+\mathbf{Pb}$$

where $\mathbf{P}=(\mathbf{p}_1\,\mathbf{p}_2\ldots\mathbf{p}_t)$ is the matrix of the first t eigenvectors and $\mathbf{b}=(b_1\,b_2\ldots b_t)^{\mathrm{T}}$ is a vector of weights for each eigenvector. Also, the eigenvectors are orthogonal, i.e. $\mathbf{P}^{\mathrm{T}}\mathbf{P}=\mathbf{I}$, so

$$\mathbf{b}=\mathbf{P}^{\mathrm{T}}(\mathbf{x}-\bar{\mathbf{x}}).$$

These equations allow us to generate new examples of the shapes by varying the parameters ($\mathbf{b}$) within suitable limits, so the new shape will be similar to those in the training set. The parameters are linearly independent, though non-linear dependencies may still be present. The limits for each element b_k of $\mathbf{b}$ are derived by examining the distributions of the parameter values required to generate the training set. Since the variance of b_k over the training set can be shown to be λ_k, suitable limits are likely to be of the order of

$$-3\lambda_k^{1/2}\leqslant b_k\leqslant 3\lambda_k^{1/2}$$

since most of the population lies within three standard deviations of the mean.

Alternatively, one can choose sets of parameters $\{b_1\ldots b_t\}$ such that the Mahalanobis distance (D_{m}) from the mean is less than a suitable value $D_{\max}$, i.e.

$$D_{\mathrm{m}}^2=\sum_{k=1}^{t}\left(\frac{b_k^2}{\lambda_k}\right)\leqslant D_{\max}^2 \tag{1}$$

4 Examples of shape models

The techniques described above have been used to generate shape models for both manufactured and biological objects. We present results for the set of resistor shapes shown in Fig. 2, a set of heart ventricle shapes and a set of hand shapes.

4.1 Resistor example

The resistor shapes were aligned using the method described above, with the mean shape scaled so the average distance of each point from its centre of gravity is one unit. Figure 4 shows the mean shape. The most significant eigenvalues of the covariance matrix derived are shown in Table 1.

TABLE 1. Eigenvalues of the covariance matrix derived from a set of resistor shapes

Eigenvalue	λ_i	$\lambda_i/\lambda_{\mathrm{T}}(\times 100\%)$	$\lambda_i^{1/2}$
λ_1	0.207	66	0.46
λ_2	0.026	8	0.16
λ_3	0.017	5	0.13
λ_4	0.013	4	0.11
λ_5	0.010	3	0.10
λ_6	0.008	3	0.09

We calculated the shape parameters $\mathbf{b}$ required to generate each example of the training set. Figure 5 shows the plot of b_1 against b_2 (normalized by their standard

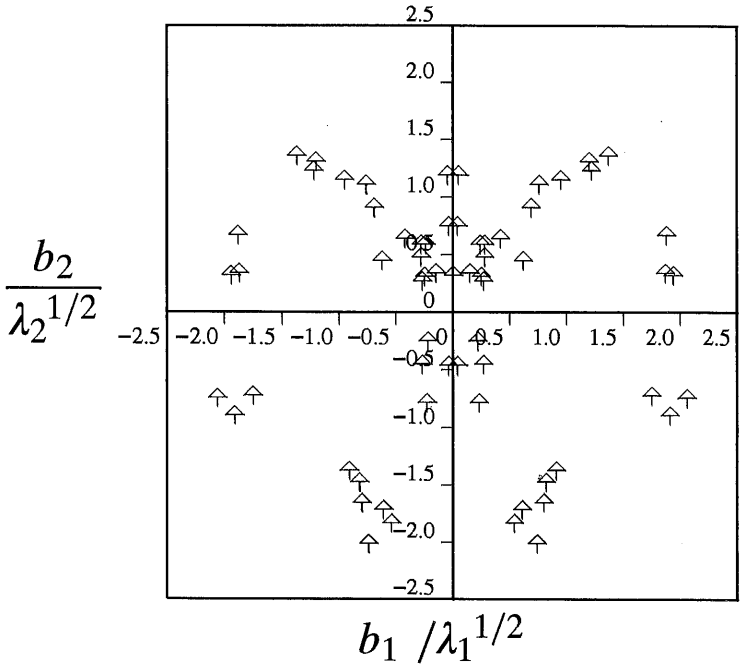


FIG. 5. Plot of b_1 versus b_2 for a training set of resistor shapes.

deviations) for the training set. The symmetry comes from the symmetry of the examples used in the training set. The lack of other structure in the scatter plot suggests that the parameters can be treated as independent. We are currently working on deriving more formal tests of independence. Any dependencies between the parameters would imply non-linear relationships between the original point positions and would result in some combinations of parameters generating 'illegal' shapes.

By varying the first three parameters separately, we can generate examples of the shape, as shown in Figs 6–8. Each parameter represents a mode of variation of the shape, which can frequently be associated with an intuitive description of the deformation (cf. Figs 6–8 with Fig. 2). Varying the first parameter (b_1) adjusts the position of the body of the resistor up and down the wire. The second parameter varies the shape of the ends of the main body of the resistor, between tapered and square. The third parameter affects the curvature of the wires at either end. Subsequent parameters have smaller effects, including the wires bending in opposite directions. These modes of variation effectively capture the variability present in the training set.

4.2 Heart example

Figure 9 shows examples from a set of 66 heart ventricle boundaries obtained by drawing around ventricles in echocardiograms. Each is represented by 96 points. The closed boundary is the left ventricle. To its left is the septum wall, separating it from the right ventricle, and below it is the top of the left atrium chamber. This example demonstrates how one model can represent both shapes and spatial

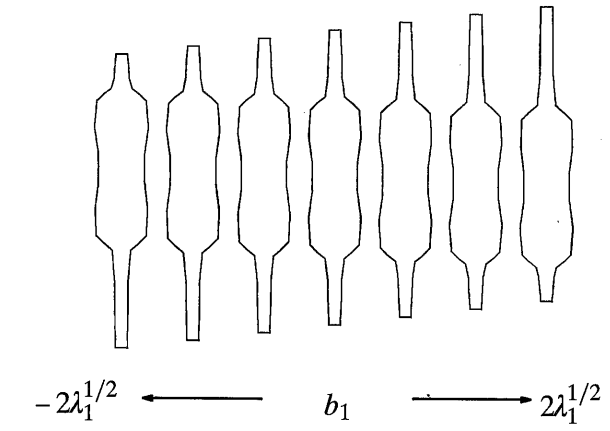


FIG. 6. Effects of varying the first parameter of the resistor model.

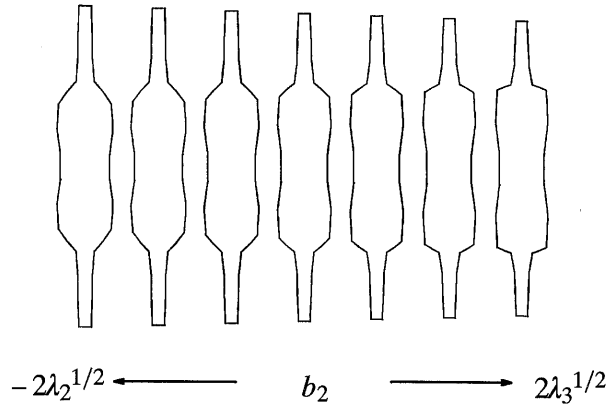


FIG. 7. Effects of varying the second parameter of the resistor model.

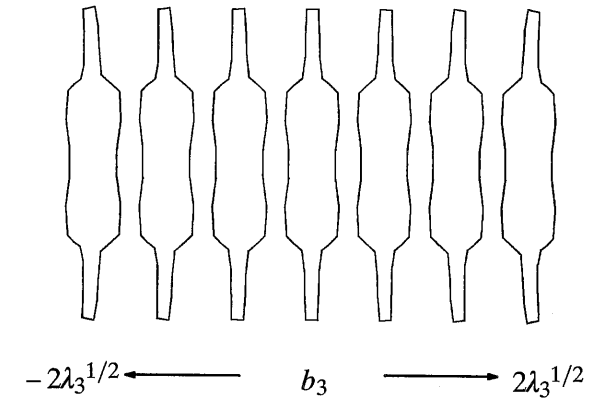


FIG. 8. Effects of varying the third parameter of the resistor model.

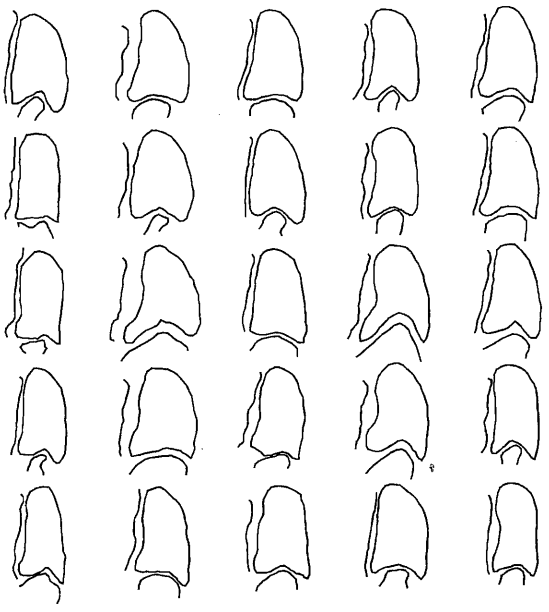


FIG. 9. Examples of heart ventrile shapes, each contaiing 96 points.

relationships between several objects. The points represent the boundary of the left ventricle, part of the boundary of the right ventricle and part of the boundary of the left atrium (below the ventricle in the examples). Table 2 shows the eigenvalues of the covariance matrix obtained from a PDM of the left ventricle of the heart, as described above.

Figure 10 suggests that b_1 and b_2 are acceptably independent, and Fig. 11 shows the reconstructed shapes using the first four model parameters in turn. The first parameter varies the width of the shape; the second parameter varies the shape of the atrium (below the ventricle); the third and fourth parameters vary the shape of the left ventricle and the modelled part of the atrium.

4.3 Hand example

A set of 18 hand shapes was generated from images of the right hand of one of the authors (Fig. 12). Each was represented by 72 points around the boundary. These were placed on the examples by locating 12 control points at the ends and joints of

TABLE 2. Eigenvalues of the covariance matrix derived from a set of heart ventricle shapes

Eigenvalue	λ_i	$\lambda_i/\lambda_T(\times 100\%)$	$\lambda_i^{1/2}$
λ_1	0.483	37	0.70
λ_2	0.225	17	0.47
λ_3	0.171	13	0.41
λ_4	0.097	7	0.31
λ_5	0.083	6	0.29
λ_6	0.048	4	0.22

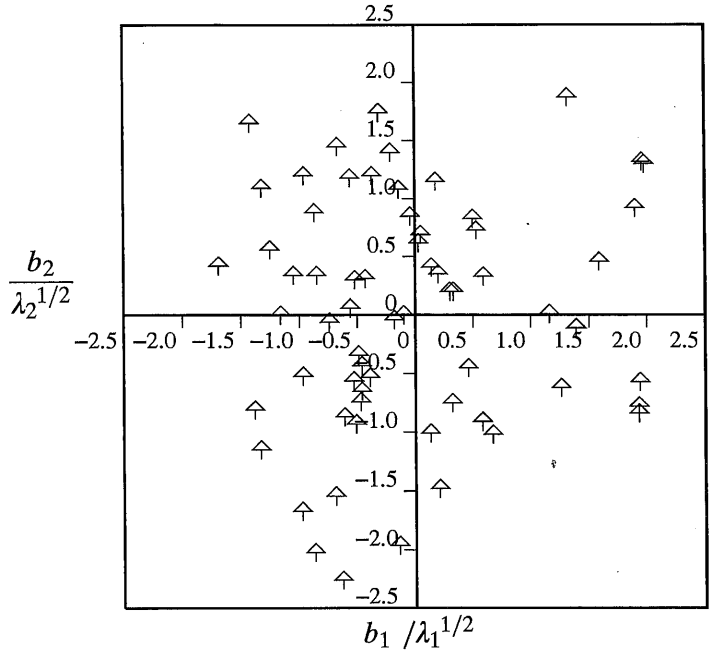


FIG. 10. b_1 versus b_2 for the training set of heart venticle examples.

the fingers and filling in the rest equally along the connecting boundaries. A model was trained on the data and it was found that 96% of the variance could be explained by the first six modes of variation. The first three modes are shown in Fig. 13 and consist of combinations of movements of the fingers. Again, a compact parameterized model has been generated.

5 Modelling grey level appearance

We wish to use our models to locate examples of objects in new images. For this purpose, not only shape but also grey level appearance is important. We account for this by examining the statistics of the grey levels in regions around each of the labelled model points. Since a given point corresponds to a particular part of the object, the grey level patterns above that point in images of different examples will often be similar. The work of Bailes and Taylor (1992) suggested that the location of model points in images can be improved by incorporating each point's grey level environment into the model.

We need to associate an orientation with each point of our shape model, in order to align the region to be modelled correctly. A convenient way to do this is to define an orientation with respect to nearby model points. For instance, if the points lie along a boundary, we can choose to align the region with the normal to the boundary. This can be easily calculated from the positions of the points on either side of the chosen point.

Although we can generally consider a region of any shape around each point, we will concentrate on one-dimensional lines normal to curves passing through each point (Fig. 14). (The set of grey levels along such a line is known as a 'grey level profile'.) This requires that we define the connectivity of the model points. In many

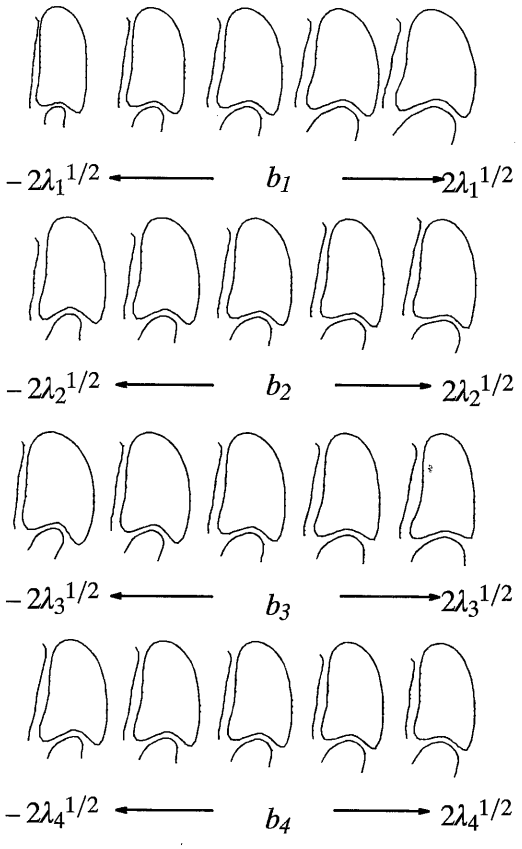


Fig. 11. Effects of individually varying each of the first four parameters of the heart ventricle model.

cases, this is straightforward, particularly when the points lie around a boundary. For every point i in each image j, we can extract a profile g'_{ij} of length n_p pixels, centred at the point. Following Bailes and Taylor (1992), we choose to sample the derivative of the grey levels along the profile in the image and normalize. This gives invariance to uniform scaling of the grey levels and the addition of a constant.

If the profile runs from p_{start} to p_{end} and is of length n_p pixels, the kth element of the derivative profile is defined to be

$$g'_{ijk} = I_j(y_{k+1}) - I_j(y_{k-1})$$

where y_k is the kth point along the profile, such that

$$y_k = p_{\text{start}} + \frac{k-1}{n_p - 1}(p_{\text{end}} - p_{\text{start}})$$

and $I_j(y_k)$ is the grey level in image j at that point. We then normalize this profile to obtain

$$g_{ij} = \frac{g'_{ij}}{\sum_{k=1}^{n_p} |g'_{ijk}|}$$

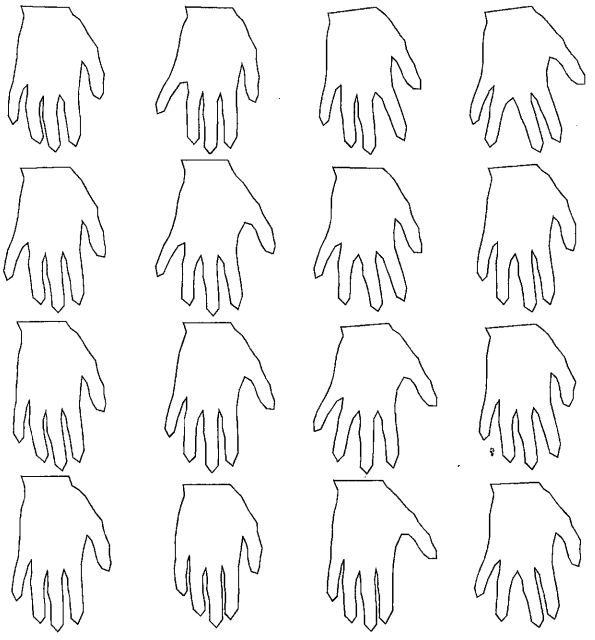


FIG. 12. Training set of hand shapes, each defined by 72 points.

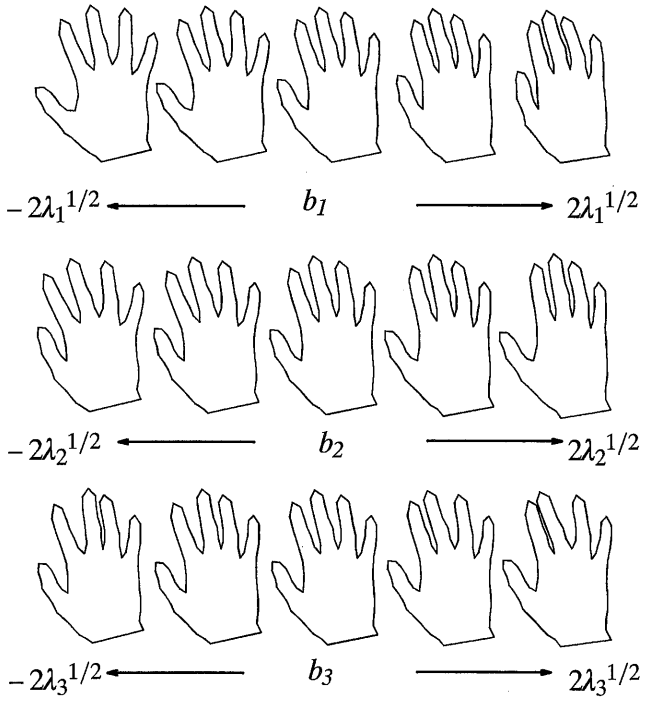


FIG. 13. Effects of individually varying each of the first three parameters of the hand model.

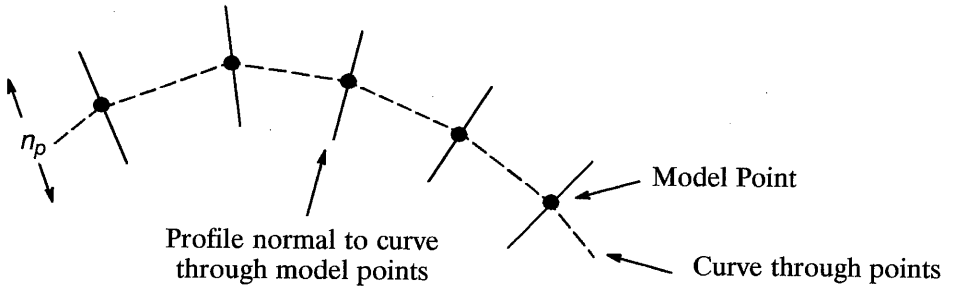


FIG. 14. Profiles of length n_p are constructed normal to the curve through each model point.

The normalized derivative profile tends to be more invariant to changes in the image caused by variations in lighting than simply using the raw set of grey levels.

For each point i, we can calculate a mean normalized derivative profile to be

$$\bar{g}_i = \frac{1}{N_s} \sum_{j=1}^{N_s} g_{ij} \tag{2}$$

We can then calculate an $n_p \times n_p$ covariance matrix $\mathbf{S}_{gi}$, giving us a statistical description of the expected profiles about the point. Alternatively, a simpler model can be generated by calculating the standard deviation of each profile model pixel about the mean $g_{i\sigma}$:

$$(g_{i\sigma k})^2 = \frac{1}{N_s} \sum_{j=1}^{N_s} (g_{ijk} - \bar{g}_{ik})^2$$

(This is the leading diagonal of $\mathbf{S}_{gi}$.)

6 Using PDMs in image search: active shape models

Having generated a flexible model and a description of the grey levels about each model point, we would like to find new examples of the modelled object in images. In general, a procedure for achieving this has two stages:

(1) a number of hypotheses are made, giving approximate locations of the model points;
(2) each of the hypotheses is refined and evaluated, and the best is chosen.

The initial hypotheses take the form of estimates for the position of the centre of the object, its orientation, scale and the shape parameters required to fit the model to the image. Such hypotheses can be obtained from cue generators or by using suitable search techniques. Hill and co-workers (1992a, b) described methods for finding flexible objects in images which use genetic algorithms to generate set of hypotheses quite rapidly.

In this section, we describe an iterative method for refining hypotheses to give a better match of the model example to the object in the image. The approach is as follows:

(1) examine a region of the image around each point to calculate the displacement of the point required to move it to a better location;

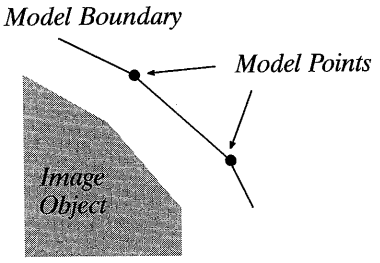


FIG. 15. Part of a model boundary approximating to the edge of an image object.

(2) from these, calculate adjustments to the orientation and to the shape parameters of the PDM;
(3) update the model parameters; by enforcing limits on the shape parameters, global shape constraints can be applied, ensuring the shape of the model instance remains similar to those of the training set.

The procedure is repeated until no significant changes result, suggesting that a local optimum has been found. Because the models are deformed to give a better fit of the data, but only in ways which are consistent with the shapes found in the training set, we call the active shape models (ASMs).

By choosing a set of shape parameters **b** for a PDM, we define the shape of a model object in an object-centred coordinate frame. We can then create an instance **X** of the model in the image frame by defining the position, orientation and scale:

$$\mathbf{X} = \mathbf{M}(s, \theta)[\mathbf{x}] + \mathbf{X}_c$$

where

$$\mathbf{X}_c = (X_c, Y_c, X_c, Y_c, \ldots, X_c, Y_c)^T$$

$M(s, \theta)[\ ]$ denotes performing a rotation of θ and with scaling by s, and (X_c, Y_c) is the position of the centre of the model in the image frame.

6.1 Calculating a suggested movement for each model point

Given an initial estimate of the positions of a set of model points which we are attempting to fit to an image object, we need to estimate a set of adjustments which will move each point toward a better position. In the case in which the model points represent the boundary of an object (Fig. 15), the required adjustments will move them towards the edges of the image object. If we have profile models for each point, the search involves finding nearby regions which best match the profile models (Fig. 16). At a particular model point, we extract a derivative profile **g**, from the current image of some length $l(>n_p)$ centred at the point and aligned parallel to the orientation we have defined at that point (for instance, normal to the boundary). We then search this sampled profile to find the point at which the profile model best matches (Fig. 17).

Give a sampled derivative profile, the fit of the model at a point d pixels along it is calculated according to

$$f_{\text{prof}}(d) = (h(d) - \bar{g})^T \mathbf{S}_g^{-1} (h(d) - \bar{g})$$

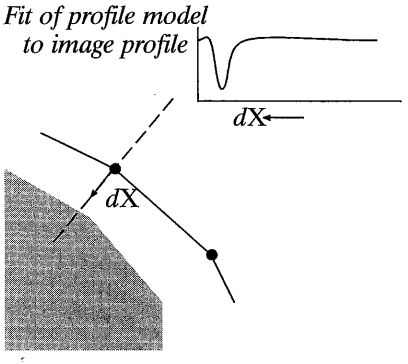


FIG. 16. Suggested movement of point is along normal to boundary, in a direction towards the point at which the profile model best fits the profile sampled from the image.

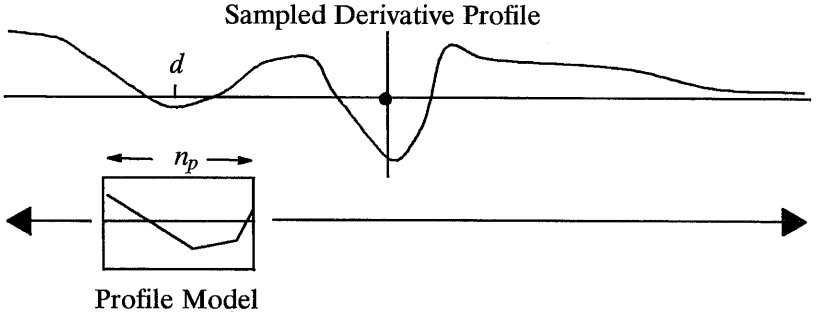


FIG. 17. Sampled profile is searched for the best fit of the profile model.

where $h(d)$ is a sub interval of $\mathbf{g}$ of length n_p pixels centred at d, normalized using equation (2). This is the Mahalanobis distance of the sample from the mean grey model and is proportional to the logarithm of the probability of obtaining $h(d)$ from the measured distribution of grey levels.

If the model consists of only the mean $\bar{\mathbf{g}}$ and standard deviation $\mathbf{g}_\sigma$ of each pixel along the profile, when we can use

$$f_{\text{prof}}(d) = (h(d) - \bar{\mathbf{g}})^{\mathrm{T}} \mathbf{S}_\sigma^{-2} (h(d) - \bar{\mathbf{g}})$$

where $\mathbf{S}_\sigma$ is a diagonal matrix whose leading diagonal is formed from the elements of $\mathbf{g}_\sigma$. In both cases, the value of f_{prof} decreases as the fit improves. The point of best fit is thus the point at which $f_{\text{prof}}(d)$ is a minimum.

Let us suppose that d_{best} is the distance along the sampled profile from the model point to the point of best fit. We choose a displacement for the model point of $\mathrm{d}\mathbf{X}$ which is parallel to the profile, in the direction of the point of best fit, with magnitude

$$\begin{aligned} |\mathrm{d}\mathbf{X}| &= 0 && \text{if } |d_{\text{best}}| \leqslant \delta \\ |\mathrm{d}\mathbf{X}| &= 0{\cdot}5 d_{\text{best}} && \text{if } \delta < |d_{\text{best}}| < d_{\max} \\ |\mathrm{d}\mathbf{X}| &= 0{\cdot}5 d_{\max} && \text{if } d_{\max} \leqslant |d_{\text{best}}| \end{aligned}$$

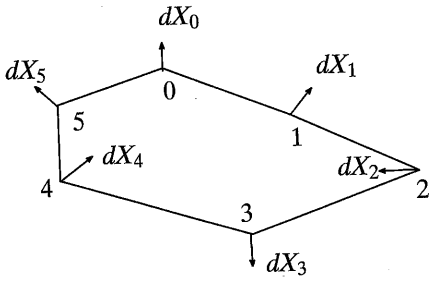


FIG. 18. Adjustments to a set of points.

for some δ and $d_{\max}$. (We use $\delta = 0.5$ pixels, $d_{\max} = 8$ pixels in the following experiments. Also, we choose not to set $|\mathrm{d}\mathbf{X}| = d_{\text{best}}$ without applying limits, because this may cause occasional noise matches a long way off to distort the model shape and position too much.)

An alternative approach is to generate potential images, such as to those described by Kass *et al.* (1987), possibly one for each model point, describing how likely each point in the image is to be the model point. Adjustments to the position of each point can then be derived from the gradient of the potential image at the current estimate of the point's position.

6.2 Calculating the adjustments to the orientation and shape parameters

A set of adjustments can be calculated, one for each point of the shape (Fig. 18). We denote such as set as a vector $\mathrm{d}\mathbf{X}$, where

$$\mathrm{d}\mathbf{X} = (\mathrm{d}X_0, \mathrm{d}Y_0, \ldots, \mathrm{d}X_{n-1}, \mathrm{d}Y_{n-1})^{\mathrm{T}}$$

We aim to adjust the orientation and shape parameters to move the points from their current locations in the image frame $\mathbf{X}$, to be as close to the suggested new locations $(\mathbf{X} + \mathrm{d}\mathbf{X})$ as can be arranged while still satisfying the shape constraints of the model.

If the current estimate of the model is centred at (X_c, Y_c) with orientation θ and scale s, we would like first to calculate how to update these parameters to give a better fit for the image. This is achieved by finding the translation $(\mathrm{d}X_c, \mathrm{d}Y_c)$, rotation $\mathrm{d}\theta$ and scaling factor $(1 + \mathrm{d}s)$ which best map the current set of points, $\mathbf{X}$ on to the set of points given by $(\mathbf{X} + \mathrm{d}\mathbf{X})$. This can be achieved by a weighted least-squares fit (see appendix). The choice of possible weights is discussed below.

Having adjusted the orientation variables, there remain residual adjustments which can only be satisfied by deforming the shape of the model. We wish to calculate the adjustments $\mathrm{d}\mathbf{x}$ to the original model points in the local coordinate frames, as required to cause the scaled, rotated and translated points $\mathbf{X}$ to move by $\mathrm{d}\mathbf{X}$ when combined with the new scale, rotation and translation parameters.

We showed previously (Cootes *et al.*, 1992b) that

$$\mathrm{d}\mathbf{x} = M((s(1 + \mathrm{d}s))^{-1}, -(\theta + \mathrm{d}\theta))[M(s, \theta)[\mathbf{x}] + \mathrm{d}\mathbf{X} - \mathrm{d}\mathbf{X}_c] - \mathbf{x} \qquad (3)$$

This lets us calculate the suggested movements to the points $\mathbf{x}$ in the local model coordinate frame. Since there are only $t(< 2n)$ modes of variation available and $\mathrm{d}\mathbf{x}$ can move the points in $2n$ different degrees of freedom, we can only achieve an approximation to the deformation required. The movements are not generally consistent with our shape model.

We wish here to calculate the adjustments $\mathbf{db}$ to the shape parameters which will give the best match of the model to the suggested new positions. This can be thought of as minimizing a (possibly weighted) sum of squares of differences between model points and desired points. We wish to minimize

$$(\mathbf{dx}')^{\mathrm{T}}\mathbf{W}_s(\mathbf{dx}') \tag{4}$$

where $\mathbf{W}_s$ is a diagonal matrix of weights, one for each coordinate of each point, and

$$\begin{aligned}\mathbf{dx}' &= ((\mathbf{x}+\mathbf{dx})-(\bar{\mathbf{x}}+\mathbf{P}(\mathbf{b}+\mathbf{db}))\\ &= (\mathbf{dx}-\mathbf{P}\mathbf{db})\end{aligned}$$

It can be shown that equation (4) is minimized when

$$(\mathbf{P}^{\mathrm{T}}\mathbf{W}_s)\mathbf{dx} = (\mathbf{P}^{\mathrm{T}}\mathbf{W}_s\mathbf{P})\mathbf{db} \tag{5}$$

This is a set of t linear equations in the t variables of $\mathbf{db}$, and can be solved using standard matrix algebra.

In the special case in which all weights are set to unity, i.e. $\mathbf{W}_s = \mathbf{I}$, equation (5) simplifies to

$$\mathbf{db} = \mathbf{P}^{\mathrm{T}}\mathbf{dx} \tag{6}$$

6.3 Choice of weights for orientation and shape parameter adjustment calculation

In the calculation of the adjustments to both the orientation and shape parameters, we can assign a weight to each point, to indicate the confidence we have in the suggested new position for the point. If all the weights are set to unity, simplifications can be made which reduce the complexity of the calculations required (equation (6)). The weights could be based on the fit which the grey level model achieved to the image.

Here, we use a method which penalised points which are found to be further away from the current model points than the average, and may be the result of outliers:

$$w_i = \frac{1}{2+|\mathbf{dX}_i|^2} \tag{7}$$

The '2' in the denominator is to prevent division by zero errors and to ensure that the weightings lie in the range $(0, 0.5)$.

6.4 Updating the orientation and shape parameters

The equations above allow us to calculate changes to the pose variables—$\mathbf{dX}_c$, $\mathbf{dY}_c$, $d\theta$ and ds—and adjustments to the shape parameters ($\mathbf{db}$) required to improve the match between an object model and image evidence. We apply these to update the parameters in an iterative scheme as follows:

$$\begin{aligned}\mathbf{X}_c &\rightarrow \mathbf{X}_c + w_t\mathbf{dX}_c\\ \mathbf{Y}_c &\rightarrow \mathbf{Y}_c + w_t\mathbf{dY}_c\\ \theta &\rightarrow \theta + w_\theta d\theta\\ s &\rightarrow s(1+w_s ds)\\ \mathbf{b} &\rightarrow \mathbf{b} + \mathbf{W}_b\mathbf{db}\end{aligned} \tag{8}$$

where w_t, w_s and w_θ are scalar weights, and W_b is a diagonal matrix of weights for each mode. This can either be the identity matrix or one in which each weight can

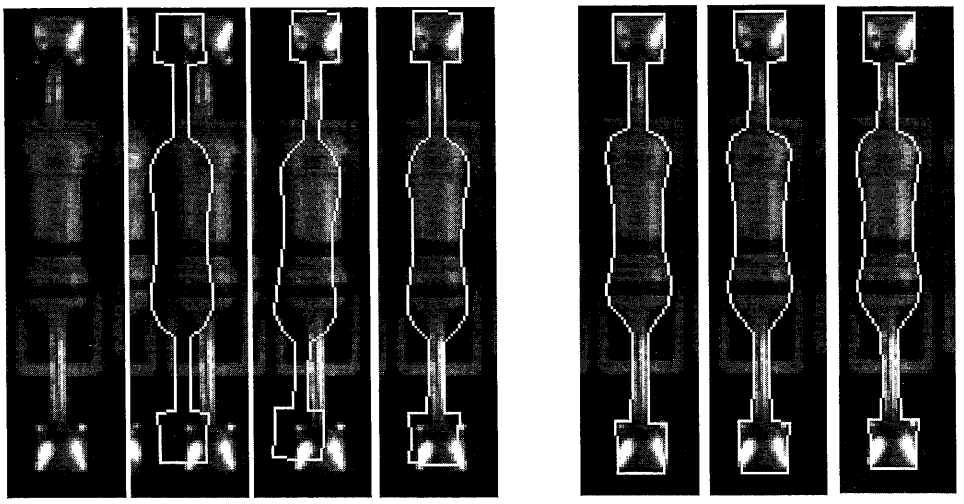


FIG. 19. Example of a resistor ASM iterating on an image.

be proportional to the standard deviation of the corresponding shape parameter over the training set. The second possibility allows more rapid movement in modes which represent larger shape variations.

We can ensure that the model only deforms into shapes consistent with the training set by placing limits on the values of b_k. As mentioned above, a shape can be considered acceptable if the Mahalanobis distance D_m is less than a suitable constant D_{max}, for instance 3.0 (see equation (1)). In other words, the vector **b** should lie within a hyper ellipsoid above the origin. If updating **b** using equation (8) leads to an implausible shape, i.e. $D_m > D_{max}$ and the point lies outside the ellipsoid, **b** can be rescaled to lie on the closest point of the allowed volume using

$$b_k \rightarrow b_k \left(\frac{D_{max}}{D_m} \right)^{1/2}, \qquad (k = 1, \ldots, t)$$

It should be noted that we have already applied implicit limits of zero to the weights of the eigenvectors truncated from our representation (i.e. $b_i = 0 \forall i > t$). Once the parameters have been updated and limits applied, where necessary, a new example can be calculated and new suggested movements derived for each point. The procedure is repeated until no significant change results.

7 Examples of ASMs in action

To date, we have implemented the above techniques using the simpler grey level profile models (mean and standard deviation only). We now present some examples of using the flexible models in image search.

7.1 Resistor images

A model was trained on a set of resistors. Figure 19 shows it fitting a new example on an image of a printed circuit board. The model is more detailed than the one

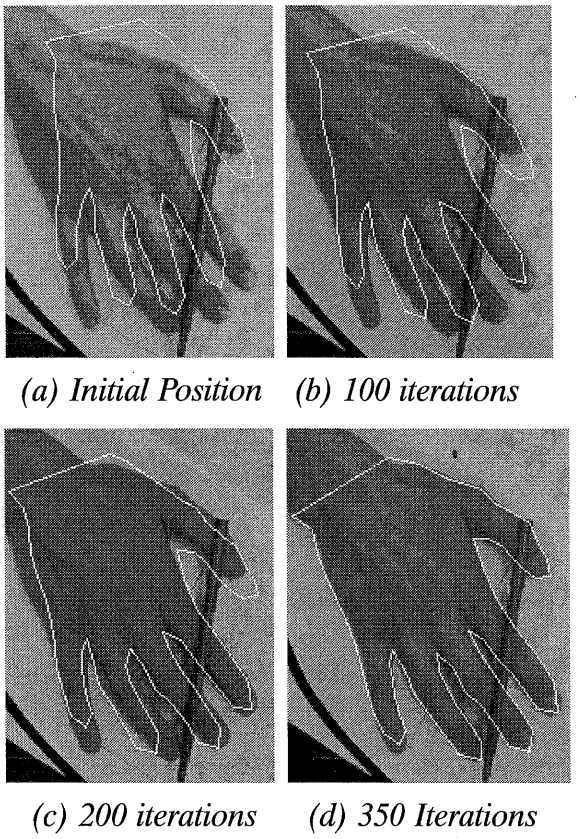


FIG. 20. Images of author's hand with hand model superimposed, showing (a) its initial position and its location after (b) 100, (c) 200 and (d) 350 iterations.

described above, because it includes the solder pads as well as the resistor body and wire. The shape was represented using 48 points placed around the boundaries of the body, wire and pads. It was trained on a set of 20 examples from a different circuit board. Its modes of variation include the body sliding up and down the wire; the wires moving about on the pads; and the body changing shape. Profile models seven pixels long were used at each point. It should be noted that there was little improvement in the fit after the first 50 iterations, and that the model became stable. Each iteration took about 0.18 s on a Sun SPARC 2 workstation.

7.2 Hand images

We have constructed a PDM of a hand, representing the boundary using 72 points (Section 4.3). Figure 20 shows an image of the hand of one of the authors and an example of the model iterating towards it. In this example, we calculate adjustments to each point simply by finding the strongest edge on a profile 35 pixels long centred on the point, rather than using grey level profile models. The shape model has eight degrees of freedom and each iteration takes about 0.03 s on a Sun SPARC 2 workstation. The result demonstrates that the method can deal with the limited occlusion caused by the pen laying across the fingers. The model points on the

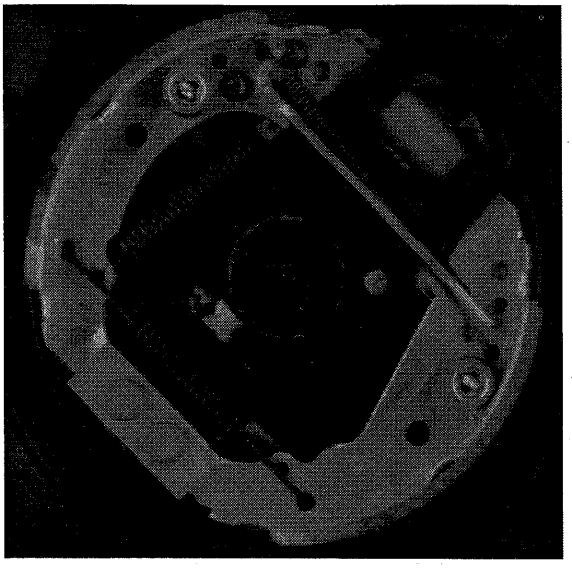


FIG. 21. Example of brake assembly.

hand's boundary under the pen are unable to find the correct edge but are constrained to the correct positions by the shape model and the positions of all the other points.

7.3 Brakes

Figure 21 shows an image of a car brake drum assembly. Although most of the components are rigid, the shape of the whole assembly can vary as the components move about relative to one another. Figure 22 shows the best fit of an ASM

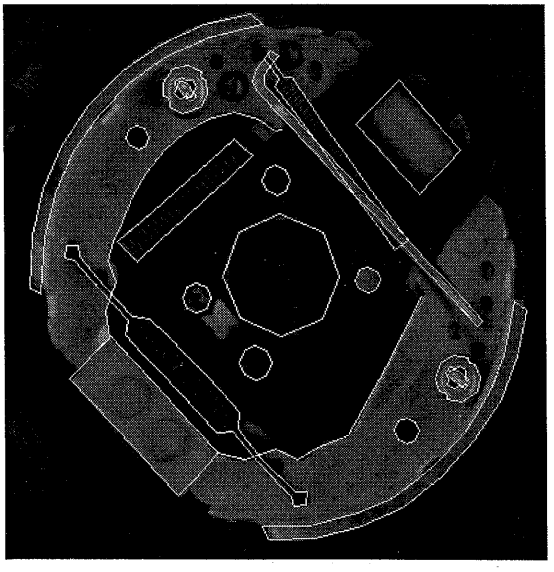


FIG. 22. Result of fitting brake model.

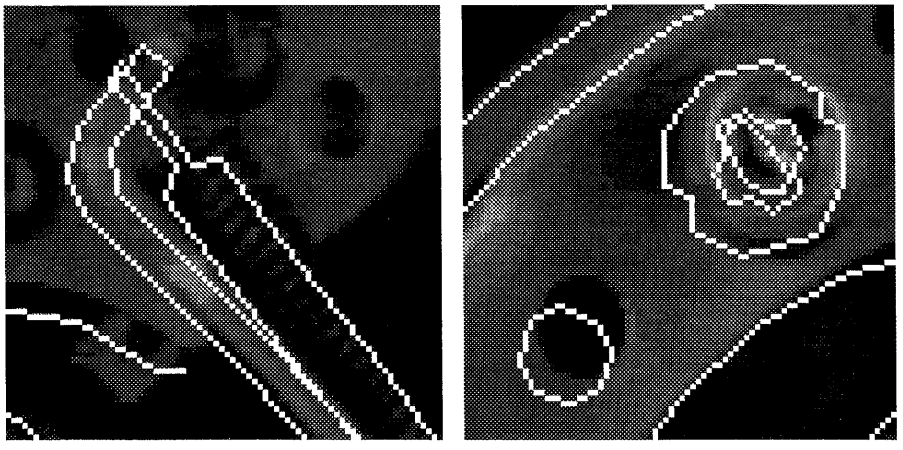


FIG. 23. Detail of model fit to image before ASM iterations applied.

superimposed. The model represents the components using 308 landmark points on 25 curves. It was trained on 15 examples and its modes of variation allow various components to slide the pivot relative to the rest of the assembly. To fit the model to the image, a coarse exhaustive search was carried out by varying the first two shape parameters and choosing the best fit.

The quality of the fit can be found by averaging the quality of the fit of each grey level profile model to the image data at each point. This resulted in an approximate fit, details of which are shown in Fig. 23. Then, 12 iterations of the ASM were applied, using profile models seven pixels long. Figure 22 shows the result. Details are shown in Fig. 24, demonstrating that the fit has improve significantly.

7.4 Other examples

We have applied the techniques described above to a number of different problems in both the industrial and medical fields. Other examples where the techniques have

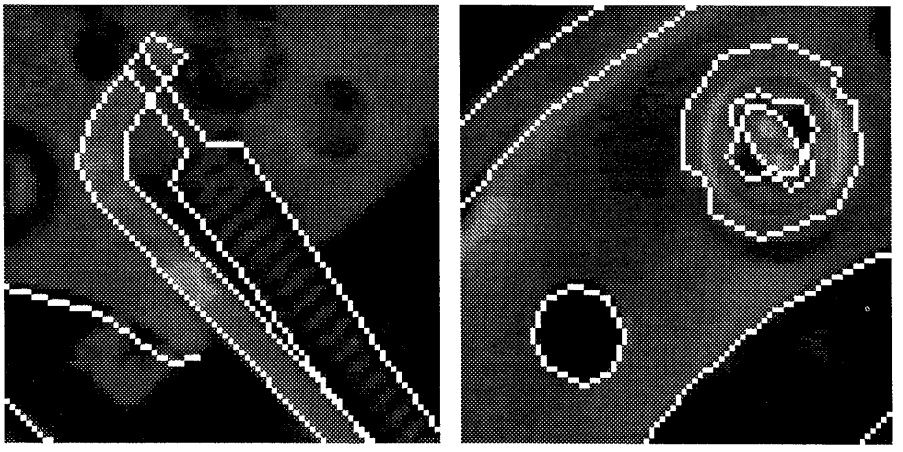


FIG. 24. Details of model fit to image after 12 ASM iterations applied.

worked successfully are the left ventricle of the heart, as seen in echocardiograms, the ventricles of the brain in magnetic resonance (MR) images (Hill *et al.*, 1992a), vertebrae in lateral X-rays of the spine (Lindley, 1992), and the outline of the abdomen in MR images of the torso. In each case, the same procedures were used to build the models and search new images.

8 Discussion

8.1 The shape model

8.1.1 Choice of model points. As can be seen in some of the examples, model points do not have to lie only on the boundaries of objects. They can represent internal features, and even subcomponents of a complex assembly. In the subcomponent case, the model describes both the variations in the shapes of the subcomponents and the geometric relationships between the components.

It is important that the points are placed on the training images as accurately as possible and the shapes aligned similarly. If not, the model will be unable to represent the position of the point correctly—it will include terms describing the noise caused by errors in point location and shape orientation.

8.1.2 Extensions to the model. A PDM is a linear model of shape variation. Each model point can be thought of as moving along a straight line as each shape parameter is varied. As a result, the method is unable to deal efficiently with the effects of large rotations of subparts or bending, which would require that points followed curves as each shape parameter varied. We are experimenting with non-linear models to deal with such cases and have obtained encouraging initial results.

It is also possible to extend the PDM to deal with volume data (e.g. three-dimensional medical images). The problems of devising ways of choosing suitable model points and placing them consistently on sets of examples are currently being investigated.

8.2 The ASM

The iterative approach described above, using image evidence to deform a PDM, is effective at locating objects, given an initial estimate of their position, scale and orientation. How good an estimate is required depends on how cluttered the image is and how well the model describes the object in the image.

8.2.1 Deforming the model examples. By allowing the model to deform, though only in ways seen in the set of examples used as a training set, we have a powerful technique for refinement. The constraints on the shape of the model are applied by the limits on the shape parameters.

We have found that the addition of weighting to the calculation of orientation and shape parameter adjustments can slow convergence but leads to better overall results, partly because spurious profile model matches away from the majority of the model points can be discriminated against.

8.2.2 Grey level profile models. The inclusion of information about the grey level environment around each point given in the profile models gives a significant improvement in the final results over methods which move each point to strong

edges. This is because the models are less confused by clutter and spurious noise edges than is the case with the edge-based methods. It is hoped that the more detailed profile models using full covariance matrices will improve performance further.

8.3 Comparison with other work

The work presented here can be thought of as a two-dimensional application of Lowe's (1991) refinement technique. Because of the linear nature of the PDM, the mathematics is considerably simpler and can lead to rapid execution.

Our ASMs have similarities with the snakes of Kass *et al.* (1987) in that they move around in potential fields, attempting to minimize some function, subject to certain constraints on their form. Snakes are usually free to take up a wide variety of shapes; when it is necessary to ensure they retain a certain shape, this is usually achieved by adding terms to the objective function favouring the desired shape (Yuille *et al.*, 1989).

ASMs explicitly model the shape of an object and the allowed variations, giving a more powerful decscription. Each model can be generated from a set of examples—the only 'hand-crafting' is in the choice of training examples and model points. As the point planting is performed in a consistent way from application to application, the modelling method is highly generic.

8.4 Framework for object modelling and recognition

We have conducted experiments which show that the local optimization method described can be fruitfully used in conjunction with a genetic algorithm search (Hill *et al.*, 1992a, b). The genetic algorithm can be run as a cue genertor to produce a number of object hypotheses, which can be refined using the ASM. Alternatively, the ASM can be combined with the genetic algorithm search, applying one iteration at each generation of the algorithm (Hill *et al.*, 1992a). Both techniques give good results.

9 Conclusions

The methods we describe allow us to take a set of example images of one or more objects whose shape can vary, build a flexible shape model, and use that model to locate new examples in new images, given an initial approximate location. Most of the effort of model generation goes into capturing the example images and marking labelled points on each of them.

By varying the parameters of each model, we can generate new examples similar to those in the training set, making them ideal for generate-and-test search strategies. By enforcing limits on the shape parameters during the local search, we enforce global shape constraints on the model.

Profile models are able to represent the grey level environment around each point, reducing the deleterious effects of clutter and noise during image search. Results can be further improved by using suitable weighting during the parameter adjustment calculations of the ASMs.

The models can represent the varying shapes of a wide variety of different objects in both industrial and medical applications. The techniques have great potential for use in many image analysis domains.

Acknowledgements

This work was funded by the Science and Engineering Research Council under the Information Engineering Advanced Technology Programme (Project 3/2114). The authors would like to thank the other members of the Wolfson Image Analysis Unit for their help and advice, and John Kent (Leeds University) for his helpful comments on the manuscript.

Correspondence: T. F. Cootes, Department of Medical Biophysics, University of Manchester, Oxford Road, Manchester M13 9PT, UK.

REFERENCES

BAILES, D. R. & TAYLOR, C. J. (1992) The use of symmetry chords for expressing grey level constraints, *Proceedings of the British Machine Vision Conference* (Berlin, Springer), pp. 296–305.

BOOKSTEIN, F. L. (1989) Principal warps: thin-plate splines and the decomposition of deformations, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11, pp. 567–585.

BOOKSTEIN, F. L. (1991) *Morphometric Tools for Landmark Data* (Cambridge, Cambridge University Press).

BOZMA, H. I. & DUNCAN, J. S. (1991) Model-based recognition of multiple deformable objects using a game-theoretic framework, *Information Processing Medical Imaging—Proceedings of the 12th International Conference* (Berlin, Springer), pp. 358–372.

CHIN, R. T. & DYER, C. R. (1986) Model-based recognition in robot vision, *Computing Surveys*, 18, pp. 67–108.

COOTES, T. F., TAYLOR, C. J., COOPER, D. H. & GRAHAN, J. (1992a) Training models of shape from sets of examples, *Proceedings of the British Machine Vision Conference* (Berlin, Springer), pp. 9–18.

COOTES, T. F., TAYLOR, C. J., COOPER, D. H. & GRAHAM, J. (1992b) Active shape models—'smart snakes', *Proceedings of the British Machine Vision Conference* (Berlin, Springer), pp. 266–275.

COOTES, T. F., TAYLOR, C. J., LANITIS, A., COOPER, D. H. & GRAHAM, J. (1993a). Building and using flexible models incorporating grey-level information, *Proceedings of the International Conference on Computer Vision, Berlin, May.*

COOTES, T. F., HILL, A., TAYLOR, C. J. & HASLAM, J. (1993) The use of active shape models for locating structures in medical images, *Proceedings of the Conference on Image Processing for Medical Imaging, Arizona, June.*

DRYDEN, I. L. & MARDIA, K. V. (1991) General shape distributions in a plane, *Advances in Applied Probability*, 23, pp. 259–276.

GRENANDER, U., CHOW, Y. & KEENAN, D. M. (1991) *Hands. A Pattern Theoretic Study of Biological Shapes* (New York, Springer).

GOODALL, C. (1991) Procrustes methods in the statistical analysis of shape (with discussions), *Journal of the Royal Statistical Society, Series B*, 53, pp. 285–339.

GOODALL, C. R. & MARDIA, K. V. (1991) A geometric derivation of the shape density, *Advances in Applied Probability*, 23, pp. 496–514.

GOWER, J. C. (1975) Generalized Procrustes analysis, *Psychoimetrika*, 40, pp. 33–51.

GRIMSON, W. E. L. (1990) *Object Recognition by Computer: The Role of Geometric Constraints* (Cambridge MA,, MIT Press).

HILL, A., COOTES, T. F. & TAYLOR, C. J. (1992a) A generic system for image interpretation using flexible templates, *Proceedings of the British Machine Vision Conference* (Berlin, Springer), pp. 276–285.

HILL, A., TAYLOR, C. J. & COOTES, T. (1992b) Object recognition by flexible template matching using genetic algorithms. *In:* G. SANDINI (Ed.), *Proceedings of the European Conference on Computer Vision* (Berlin, Springer), pp. 852–856.

HINTON, G. E., WILLIAMS, C. K. I. & REVOW, M. D. (1992) Adaptive elastic models for hand-printed character recognition. *In:* J. E. MOODY, S. J. HANSON & R. P. LIPPMANN (Eds), *Advances in Neural Information Processing Systems 4 (*San Mateo, CA, Morgan Kauffmann).

JOHNSON, R. A. & WICHERN, D. W. (1988) *Multivariate Statistics, A Practical Approach* (London, Chapman & Hall).

KARAOLANI, P., SULLIVAN, G. D., BAKER, K. D. & BAINED, M. J. (1989) A finite element method for deformable models, *Proceedings of the Fifth Alvey Vision Conference, Reading*, pp. 73–78.

KASS, M., WITKIN, A. & TERZOPOULOS, D. (1987) Snakes: active contour models, *Proceedings of the First International Conference on Computer Vision* (London, IEEE Computer Society Press), pp. 259–268.

KENT, J. T. (1991) Discussion to a paper by C. Goodall, *Journal of the Royal Statistical Society, Series B*, 53, pp. 324–325.

KENT, J. T. (1994) The complex Bingham distribution and shape analysis, *Journal of the Royal Statistical Society, Series B*, 56, pp. 285–299.

LINDLEY, K. (1992) Model based interpretation of lumbar spine radiographs, *MSc Thesis*, University of Manchester.

LIPSON, P., YUILLE, A. L., O'KEEFFE, D., CAVANAUGH, J., TAAFFE, J. & ROSENTHAL, D. (1990). Deformable templates for feature extraction from medical images. *In:* O. FAUGELAS (Ed.), *Proceedings of the First European Conference on Computar Vision (Lecture Notes in Computer Science)* (Berlin, Springer), pp. 413–417.

LOWE, D. G. (1991) Fitting parametrized three-dimensional models to images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13, pp. 441–450.

MARDIA, K. V., KENT, J. T. & WALDER, A. N. (991) Statistical shape models in image analysis, *Proceedings of the 23rd Symposium on the Interface, Seattle, WA*, pp. 550–557.

MARDIA, K. V. & DRYDEN, I. L. (1989) The statistical analysis of shape data, *Biometrika*, 76, pp. 271–281.

PENTLAND, A. & SCLAROFF, S. (1991) Closed-form solutions for physically based modelling and recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13, pp. 715–729.

SCOTT, G. L. (1987) The alternative snake—and other animals, *Proceedings of the 3rd Alvey Vision Conference, Cambridge*, pp. 341–347.

STAIB, L. H. & DUNCAN, J. S. (1989) Parametrically deformable contour models, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA*, pp. 427–430.

YUILLE, A. L., COHEN, D. S. & HALLINAN, P. (1989) Feature extraction from faces using deformable templates, *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 104–109.

YUILLE, A. L., COHEN, D. S. & HALLINAN, P. (1992) Feature extraction from faces using deformable templates, *International Journal of Computer Vision*, 8, pp. 99–112.

Appendix: aligning a pair of shapes

Given two similar shapes $\mathbf{x}_1$ and $\mathbf{x}_2$, we would like to choose a rotation θ, a scale s and a translation (t_x, t_y) mapping $\mathbf{x}_2$ on to $M(\mathbf{x})+\mathbf{t}$ so as to minimize the weighted sum

$$E=(\mathbf{x}_1-M(s,\theta)[\mathbf{x}_2]-\mathbf{t})^{\mathrm{T}}\mathbf{W}(\mathbf{x}_1-M(s,\theta)[\mathbf{x}_2]-\mathbf{t})$$

where

$$M(s,\theta)\begin{bmatrix}x_{jk}\\ y_{jk}\end{bmatrix}=\begin{pmatrix}(s\cos\theta)x_{jk}-(s\sin\theta)y_{jk}\\ (s\sin\theta)x_{jk}+(s\cos\theta)y_{jk}\end{pmatrix}$$

$$\mathbf{t}=(t_x,t_y,\ldots,t_x,t_y)^{\mathrm{T}}$$

and $\mathbf{W}$ is a diagonal matrix of weights for each point.

If we write

$$a_x=s\cos\theta, \qquad a_y=s\sin\theta$$

a least-squares approach (differentiating with respect to each of the variables a_x, a_y, t_x, t_y) leads to a set of four linear equations, such that

$$\begin{bmatrix} X_2 & -Y_2 & W & 0 \\ Y_2 & X_2 & 0 & W \\ Z & 0 & X_2 & Y_2 \\ 0 & Z & -Y_2 & X_2 \end{bmatrix} \begin{bmatrix} a_x \\ a_y \\ t_x \\ t_y \end{bmatrix} = \begin{bmatrix} X_1 \\ Y_1 \\ C_1 \\ C_2 \end{bmatrix}$$

where

$$X_i = \sum_{k=0}^{n-1} w_k x_{ik} \qquad Y_i = \sum_{k=0}^{n-1} w_k y_{ik}$$

$$Z = \sum_{k=0}^{n-1} w_k (x_{2k}^2 + y_{2k}^2) \qquad W = \sum_{k=0}^{n-1} w_{k_s}$$

$$C_1 = \sum_{k=0}^{n-1} w_k (x_{1k} x_{2k} + y_{1k} y_{2k})$$

$$C_2 = \sum_{k=0}^{n-1} w_k (y_{1k} x_{2k} - x_{1k} y_{2k})$$

These can be solved for a_x, a_y, t_x and t_y using standard matrix methods.

27. **Structured point distribution models: modelling intermittently present features.** M. Rogers and J. Graham, *Proceedings of the British Machine Vision Conference, University of Manchester, 2001. T.F.Cootes and C.J. Taylor (eds.) BMVA Press, pp 33-41.* doi:10.5244/C.15.5

Structured Point Distribution Models: Modelling Intermittently Present Features

Mike Rogers and Jim Graham
Imaging Science and Biomedical Engineering
University of Manchester

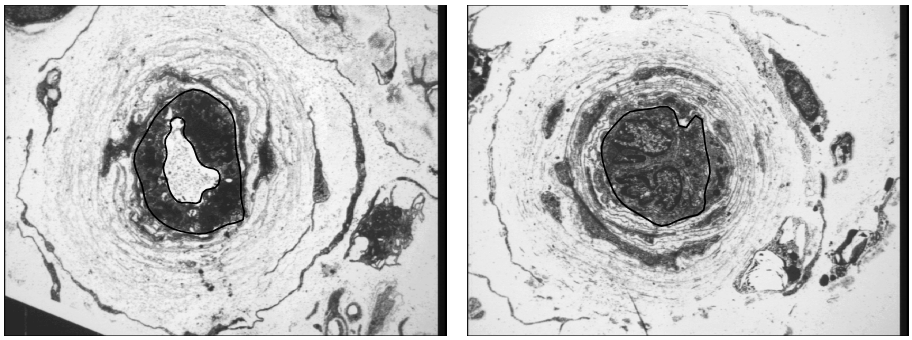
Abstract

Point distribution models have been successful in describing the shape constraints on two dimensional objects for shape description and image search. It is often the case that a class of objects to be modelled contains certain features which may be wholly present or absent in different instances. Moustaches on faces are a common example. Here we describe a method of coding the presence or absence of a feature within the PDM framework. We show that the method captures the intermittent nature of the feature as one of the modes of variation, and demonstrate that, where features are intermittently present, greater model specificity is achieved.

1 Intermittently Present Features

There are classes of images that exhibit features which are only found in some instances and not others. Examples include face images which may or may not show moustaches and/or glasses and histological sections, in which structures may appear in a proportion of contiguous slices in a stack. The particular example that led to the approach described here is the study of electron microscope images of nerve capillaries. There are several concentric layers of structures in capillary cross-sections (figure 1). The central region is the lumen: the space through which blood cells pass; this is surrounded by a layer of endothelial cells, and then the basement membrane. In disease condition, such as diabetic neuropathy, changes occur in the normal structure of the capillaries, including constriction of the lumen. In some cases the lumen can become so constricted as to be unidentifiable (figure 1(b)). Finding the boundaries between these structures is important in quantifying disease status and we have approached this task using Active Shape Models and Genetic Search for the Basement Membrane / Endothelial Cell (BMEC) boundary [6]. The lumen boundary is potentially easier to locate due to the clearer contrast, but in modelling it we need to take account of the fact that it is often missing.

To use Active Shape Model search we need to build Point Distribution Models (PDMs) of all the structures in the capillaries, including the lumen, when it is present. We have considered three possibilities for dealing with the intermittent presence of the lumen: *Separate models* for capillaries with and without a visible



(a) (b)

Figure 1: Examples from the set of diabetic nerve capillary data

lumen; a *Segmented model* in which each separate boundary has a flag for presence or absence; or a *Single model* in which each point is flagged for inclusion or exclusion individually. We prefer the third option as it allows us more flexibility in admitting arbitrary patterns of inclusion, and is more likely to capture the relationships between different components, for example, the gradual inclusion of a new feature across a stack of histological slices. The difficulty presented by this approach is in training a PDM with arbitrarily missing data points.

2 Data Imputation

Our approach to building models with arbitrarily missing data points is to include in the PDM the coordinates of those boundary points that are present, and to estimate the positions of the points that are not represented in some examples. This problem of *data imputation* - estimating missing data values - is a fairly common one in statistical applications, and a number of methods have been proposed (Rubin [5]). In adopting a method, our goal is to end up with a PDM (means and eigenvectors of the point positions) as close as possible to those we would have obtained had all the data been available. In this section we describe our own, novel, method of data imputation and evaluate its performance in comparison with three other well-founded methods with a view to their suitability for PDM building.

2.1 Imputation Methods

Replacement with mean: The simplest method is to replace each missing value with the mean of the values that are present. This clearly underestimates the variance in the data – a serious disadvantage for building PDMs.
Principal Component Analysis: Dear [4] proposed an imputation technique in which initial imputation with the means is then re-estimated using the first principal component of the imputed data. In this way, gross trends in the data are preserved.

Maximum Likelihood: Beale and Little [1] present an iterative method to produce a maximum likelihood estimate of the missing values using a form of the *Expectation Maximisation* (EM) algorithm. Before this algorithm can begin, an initial estimate of the missing data must be generated. A sensible initial value is the mean over all available data. The algorithm can be extremely sensitive to the quality of this initial estimate as is shown in the evaluation in section 2.2.
Iterated PCA: We have developed a further method of imputation, designed to retain data characteristics required by a subsequent PCA carried out on the imputed data. Specifically, we wish to impute values in such a way as to retain relationships found in the original data and do this without reducing the total variance. The algorithm is based on an iterative version of Dear's [4] PCA imputation with several modifications and can be described with the following equations:

$$(\boldsymbol{P_{xm}}, \boldsymbol{\mu_x}, \boldsymbol{\sigma^2_{xm}}, \boldsymbol{b_{xm}}) = \texttt{pca}(\boldsymbol{x}, m) \tag{1}$$

$$\hat{\boldsymbol{x}} = \boldsymbol{\mu_x} + \boldsymbol{b_{xm}}\boldsymbol{P_{xm}}^T, x_{i.M_i} = \hat{x}_{i.M_i} \tag{2}$$

where $\boldsymbol{x}$ is the original data, x_{ij} is the j^{th} observed value in example i, M_i is the set of variables missing in example i, $x_{i.M_i}$ is the set of estimated missing values from x_i and `pca` is a function that computes the first m principle components ($\boldsymbol{P_{xm}}$), the variance each mode represents ($\boldsymbol{\sigma^2_{xm}}$) and the mean ($\boldsymbol{\mu_x}$) of $\boldsymbol{x}$, together with the associated reconstruction parameters ($\boldsymbol{b_{xm}}$) for each example.

We begin by initialising $\boldsymbol{x}$, for which we use mean value imputation, and cycle through equations 1-2 until convergence. Choosing the value of m is crucial to the well-mannered convergence of the algorithm. We use the following scheme: m is set to 1 and the algorithm is run to convergence. The imputed data is now consistent with data patterns represented by the first mode of variation, but no others. To include relationships represented by other modes we increase m by 1 and repeat the convergence, starting at the result of the previous iteration. At each stage of the iteration we are including effects of higher modes in the imputed data, and matching it more closely to the original data patterns. However, the imputed data itself also has some influence on the modes produced by PCA. As we continue to include higher modes we will eventually reach one which is mainly influenced by the effect of the imputed data, after which the algorithm will not converge. Rather, the imputed data would be updated to reinforce the effects of earlier imputed data. We therefore need a stopping criterion. In our experiments we continue iterating until :

$$\frac{\sum \boldsymbol{\sigma^2_{xm}}}{\sum \boldsymbol{\sigma^2_x}} > p \tag{3}$$

where $\boldsymbol{\sigma^2_x}$ is the variance of all modes of x and p is the proportion of complete data examples. This stopping criterion is somewhat heuristic, and has not been shown to be optimal. However, it leads to satisfactory performance in the evaluation experiments.

2.2 Evaluation of Imputation Methods

Each of the imputation methods described in section 2.1 was evaluated using synthesised data and some real shape data from annotated capillary boundaries.

Synthetic data: fifty vectors, each with ten elements, were constructed using the following algorithm:

`for` $i = 1$ `to 50`
$\boldsymbol{x_i} = (i, 2i, 3i, ..., 10i)$`;` $\boldsymbol{x_i} = \boldsymbol{x_i} + i\boldsymbol{r}$`;` $\boldsymbol{x} =$ `concat` $\boldsymbol{x_i}$ `with` $\boldsymbol{x}$

The intention of this data is to evaluate the ability of an imputation method to retain the underlying relationships in the data. There is one consistent relationship for each vector, namely the increment in successive values, proportional to the first element. The relationship is not perfect, being perturbed by the random factor r (between -0.5 and 0.5), also scaled by the first element in each vector, i. These vectors do not represent shapes, but give an insight into the effectiveness of the methods in reconstructing patterns in the data corrupted by noise and missing elements.

Nerve capillary landmark data: Here we use a subset of 30 examples of the marked-up BMEC boundaries from capillary images. We take the first 30 points in each case. This data gives an insight into the performance of the imputation methods on realistic data.

Evaluation tests: In each case we remove a proportion (varying between 1% and 50%) of the data points and replace them with imputed values according to each of the four schemes. To measure the effectiveness of the imputation we make two measures on the resulting data. Firstly we measure the Euclidean distance (in the vector space of the data) between each example and its imputed version, giving a measure of the *raw error* in the imputation process. Secondly, as we are interested in preserving the modes of variation of the original data we measure the Euclidean distance between the corresponding eigenvectors of the original and imputed data sets. In the case of the synthetic data, there is only one significant mode of variation, and only one eigenvector. In the case of the capillary data, we estimate this distance for the first three eigenvectors. For the synthetic data, there is a third measure we can make. In this case we know the underlying "ideal" relationship between the elements of each vector before corruption by randomisation. It is interesting to see how well the imputation process reconstructs this underlying relationship in the presence of the noise. We therefore measure the distance between the ideal vector and the imputed vector in each case.

Results of the evaluation are shown in figure 2.2. The iterated PCA method gives the closest imputation to the original data in both cases. For the synthetic example, the maximum likelihood (EM) method gives almost identical results (fig 2(a)). In the case of the capillary data (fig 2(b)), however, the PCA method comes closest to the performance of iterated PCA, though noticeably worse at higher proportions of imputed data. In calculating the distance between the raw and imputed eigenvectors, both iterated PCA and EM again perform equivalently, and much better than the other methods, and PCA and iterated PCA give similar performance on the capillary data. The maximum likelihood estimates for imputation are influenced strongly by the initial estimates of the missing data (in this case the mean values). This is a poor estimate in the case of the capillary data and

results in the poor performance in this case. The structure in the variation of Euclidean distance between imputed and original modes of variation, with increasing proportion of imputation, seen in figures 2(c) and 2(d), is due to the significant effect that small changes can have on an eigen-analysis of the data. Figure 2(e) shows the difference between the "ideal" synthetic data and the imputed values after randomisation. The distance between the "raw" randomised data and the underlying data is, of course, independent of the quantity of imputation being applied and therefore constant. Both the EM and iterated PCA methods retain a good estimate of this distance in the presence of up to 50% imputation, and therefore seem to be responding to underlying patterns in the data. The other methods, as might be expected from figures 2(c) and 2(e), do not. Figure 2(f) shows the difference in total variance between the original and imputed capillary data. The iterated PCA method retains the total variance of the data even in the presence of large amounts of missing data. The other methods all perform poorly on this measure.

The iterated PCA method appears to have the desired properties of an imputation scheme. Other methods also have these properties for one or other of the test cases, but not both. Mean imputation was always, of course, unlikely to meet our criteria, but has been included to give a yardstick for measuring inadequate performance.

3 Modelling Shape and Structure

Here we describe how we combine data imputation with a model of structural variation. As our models constitute a variant of PDMs we call them *Structured Point Distribution Models* (SPDMs).

3.1 Building the models

The modifications that need to be made to a standard PDM to deal with intermittent structures are the following. We build a model that assumes all points are represented (our capillary model would assume a lumen, a face model might assume the presence of a moustache). When a PDM landmark point is not represented in a particular image it is replaced by a placeholder (such as NaN - a computational representation of Not a Number). Once the training set has been assembled, the shapes are aligned using the data points that are available, and the missing data values imputed by some imputation scheme (we prefer iterated PCA, of course). So an initial training vector for a shape i represented by points $[(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4)]$ where (x_3, y_3) is unobserved, is represented as the data vector: $\boldsymbol{x_i} = (x_1, x_2, NaN, x_4, y_1, y_2, NaN, y_4)^T$. Following alignment and imputation of missing values we get a new shape vector (primed elements are aligned, hat elements are imputed): $\hat{\boldsymbol{x}}'_{\boldsymbol{i}} = (x'_1, x'_2, \hat{x}_3, x'_4, y'_1, y'_2, \hat{y}_3, y'_4)^T$. Shape parameters, $\boldsymbol{b_s}$ are then calculated by PCA in the usual way [3].

While this gives us a model of shape that represents as closely as possible the shape variation we observe in the entire structure, we have lost the structural information about which boundary points may or may not be missing. We therefore

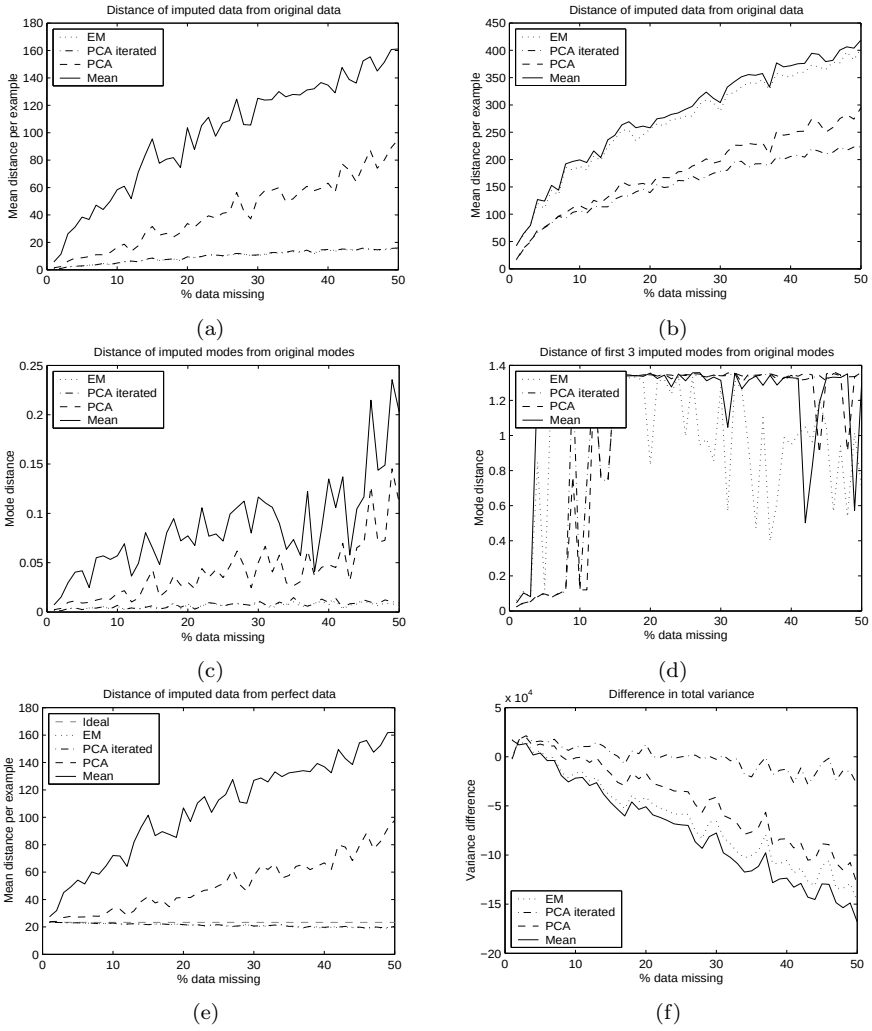


Figure 2: Imputation performance. (a),(c),(e) – synthetic data, (b),(d),(f) – capillary images. Error is shown as a function of increasing proportion of imputed data. (a),(b) – raw error. (c),(d) – error in principal components. (e) error in capturing the "ideal" data pattern. (f) difference in total variance (see text).

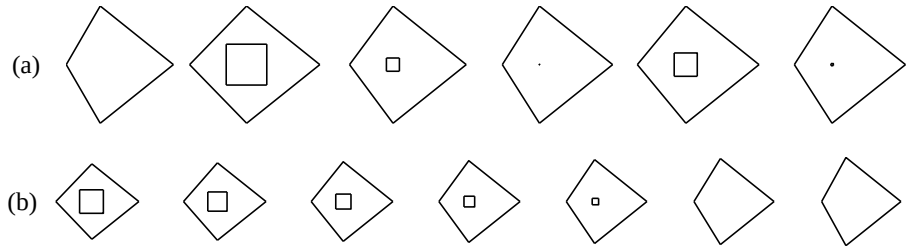


Figure 3: Synthetic shape. (a) examples from the set of synthetic training data. (b) the first mode of variation.

augment the shape vector with a binary structure vector. For our shape i above with point 3 missing the structure vector would be: $\boldsymbol{x_i^s} = (1, 1, 0, 1)^T$.

This gives us a representation of the structure containing significant redundancy, but which allows for arbitrary patterns of inclusion or exclusion of landmark points in the model. This redundancy can be reduced using PCA, just as in the case of classical PDMs. The modes of the PCA represent the relationships between the structures in the landmark data. In the case of the capillary boundaries the analysis results in a single mode, containing almost all the variance, representing the presence or absence of the lumen at its extremes. We have therefore reduced our structure vector to a parameter vector of length 1. The SPDM like the PDM is a generative model; that is, given a parameter vector we can recreate the structure vector for a particular instance. The disadvantage of this approach is that we are representing a binary process (presence or absence) by a linear model. To recover binary parameters in the reconstructed structure vector we threshold the individual elements. We use a threshold that represents the probability that a particular feature point will be present in the image.

The PCA of the structural data matrix $\boldsymbol{x^s}$ results in a matrix of continuous structure parameters $\boldsymbol{b_d}$, which can then be used, together with the shape parameters $\boldsymbol{b_s}$ to build the combined model of shape and structure. This is done in a way similar to the construction of Active Appearance Models [2].

For each training example we generate a concatenated vector:

$$\boldsymbol{b} = \begin{pmatrix} \hat{\boldsymbol{b}}_{\boldsymbol{s}} \\ \boldsymbol{b_d} \end{pmatrix} \tag{4}$$

where $\hat{\boldsymbol{b}}_{\boldsymbol{s}}$ is a matrix of shape parameters generated after first shifting and scaling the input (imputed) data to lie between 0 and 1. We perform this scaling to avoid problems associated with shape and structure being measured on different scales. In choosing to perform this transformation on the data, we are effectively treating shape and structure as equally important. A combined model of shape and structure is obtained by a further application of PCA.

$$\boldsymbol{c} \approx \boldsymbol{Q}\boldsymbol{b} \tag{5}$$

where $\boldsymbol{Q}$ is a matrix of t eigenvectors expressing the correlations between the shape and structure data in vector $\boldsymbol{b}$ and $\boldsymbol{c}$ is a vector of combined model parameters which controls both the shape and structure of the data. We can obtain $\boldsymbol{b}$ from $\boldsymbol{c}$:

$$\boldsymbol{b} = \boldsymbol{Q}^T\boldsymbol{c} \tag{6}$$

From these equations we can produce the shape and structure vector of any shape represented by the model.

3.2 Evaluation

We evaluate our approach to shape and structure modelling using a synthetic shape set, nerve capillary images and face images. Firstly we demonstrate that presence or absence of structure is represented in the model, and that correlations with

shape data are captured. Secondly we demonstrate that modelling the presence or absence of structures increases the specificity of the model.

Synthetic data: A set of synthetic shapes was generated using the following algorithm.

`generate a random number` r`, between 10 and 30`
`form a kite from the points` $[(r, 50), (50, 90), (100, 50), (50, 10)]$
`if` $(r < 25)$ `form a square, centred` $(50, 50)$`, with side length` $100 - 2r$
`otherwise put 4` NaN `values in the data vector`

This generates a set of structures consisting of squares within kites (see figure 2). The first coordinate of the kite and the size of the square are correlated. When the size of the inner square would be less than 0 the feature is not present in the image. The proportion of complete to incomplete structures is 5:1. The SPDM built from 50 training examples, retaining 99.5% of variation has only one mode of variation shown in figure 2(a). The shape model has captured the correlation between the size of the square and the shape of the kite, and the thresholding of the structure vector has removed it at the relevant places.

Nerve Capillaries: An SPDM was calculated from 38 nerve capillary images, 15 of which contained lumens so constricted that they are practically undetectable, so that only the BMEC boundary was annotated. Examples of the shapes are shown in figure 4(a). The 99.5% of data retained produced 6 modes of variation, the first three of which are shown in figure 4(b). Note that all the structural information is contained in the first mode of the model. The second expresses lumen constriction and the third appears to be capturing the translation of the lumen within the capillary.

Faces: Figure 3.2(a) shows some examples from a set of 29 face images marked up with 33 landmarks on the face outline, eyes nostrils and moustache (present in nine out of the 29 faces). Figure 3.2(b) shows the first two modes of variation. Once again the first mode represents the structural variation and the others represent shape variation.

Model Specificity: The inclusion of the lumen structure into the model of nerve capillaries is intended to contribute additional constraints to the model during search, i.e. to increase its specificity. To measure the specificity of the models, we used the 38 training examples of capillaries and 29 face images to build SPDMs and PDMs retaining 99.5% of observed variability in each case. From each training set we created increasingly invalid shapes by randomly perturbing the point positions in the training examples using the following algorithm.

`for i=1 to 25`
`  for each training example` $\boldsymbol{x}$
$\boldsymbol{x_{ir}} = \boldsymbol{x} + \frac{i\bar{\boldsymbol{x}}\boldsymbol{r}}{100}$; $\boldsymbol{b} = \boldsymbol{Q}\boldsymbol{x_r}$; $\hat{\boldsymbol{x}}_{\boldsymbol{ir}} = \boldsymbol{Q}^T\boldsymbol{b}$

This creates, for each training example, twenty five increasingly invalid shapes obtained by adding a random shift to each point. If we try to fit the model to the invalid data, a highly specific (constrained) model will find the nearest valid shape, whereas a less- specific model will fit more closely to the invalid example. For our purposes, we measure closeness as the mean point to point distance between the model fit landmarks and the corresponding annotation landmark.

Figure 5 shows the fits of SPDM and PDM models of capillaries and faces to the unperturbed(valid) and perturbed(invalid) data, with increasing random perturbation. In each case the model shows some specificity by fitting more closely to the nearest valid example than the perturbed version. However, for both capillary and face shapes, the effect is more marked for the SPDM, indicating increased specificity of the structural model.

4 Conclusions and Discussion

We have presented an extension to Point Distribution Modelling to deal with circumstances in which features of the objects to be modelled may be wholly present or absent in a proportion of examples. Our method combines the use of a structure vector, which is subject to the same statistical analysis as the shape vector, and imputation of values for model points which are coded as absent in the structure vector. We have developed a straightforward method for imputation which causes minimal distortion to the distributions of shapes in the original data. Using experiments on synthesised data and data from real images we have shown that the Structured Point Distribution Models successfully capture the variation in shape and structure present in an image set and the correlations among these, and that the use of the structured models improves the specificity of the model over the classical PDM. Although not demonstrated in this short paper, the method can be applied to Appearance Models [2] also, and model the grey level appearance of intermittently present features.

References

[1] E M L Beale and R J A Little. Missing values in multivariate analysis. 1975.

[2] T F Cootes, G J Edwards, and Taylor. Active appearance models. In *Proceedings of the European Conference on Computer Vision*, volume 2, pages 484–498, 1998.

[3] T F Cootes, C J Taylor, D H Cooper, and J Graham. Active Shape Models - their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, January 1995.

[4] R E Dear. A principle-component missing data method for multiple regression models. Technical Report Report SP-86, 1959.

[5] R J A Little and D B Rubin. *Statistical Anaylsis with Missing Data*. Wiley, New York, USA, 1987.

[6] M Rogers, J Graham, and R A Malik. Exploiting weak shape constraints to segment capillary images in microangiopathy. In *Proceedings of the 3rd International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 717–726, Pittsburg, PA, USA, October 2000.

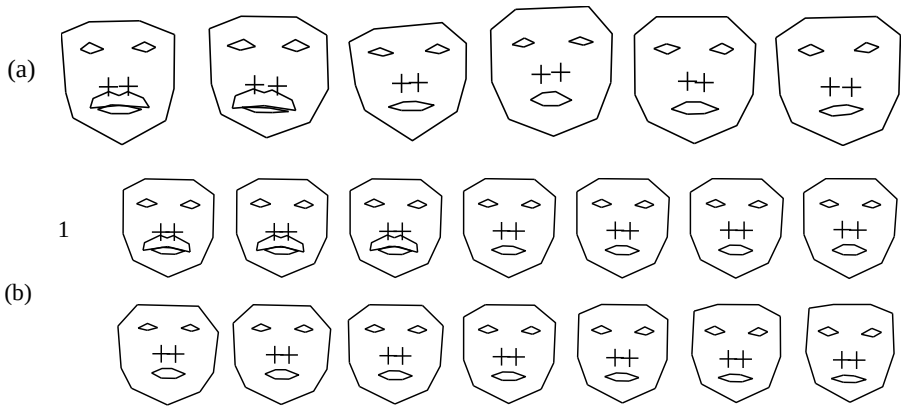


Figure 4: Faces. (a) examples from the set of face shape training data. (b) first two modes of variation. Note that the first mode encapsulates the structure of the missing data.

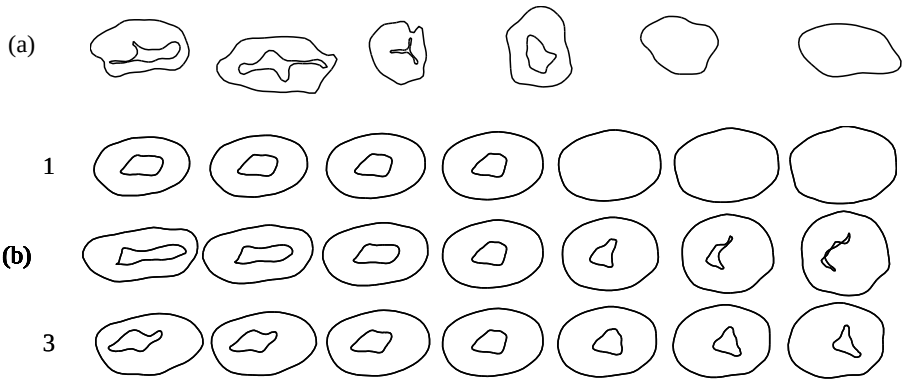
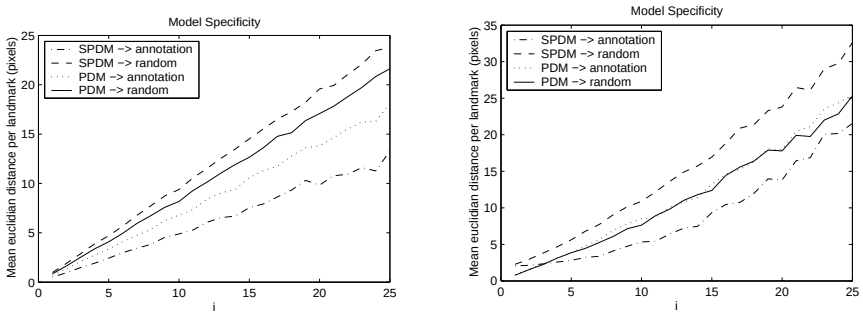


Figure 5: Capillaries. (a) examples from the set of nerve capillary training data. (b) the first three of six modes. Note that the first mode encapsulates the structure of the missing data.



(a) (b)

Figure 6: The curves show the mean point to point landmark distance for both PDM and SPDM model fits to the original shape (annotations) and perturbed examples (random). (a) capillaries, (b) faces.

28. **Robust Active Shape Model search**. M. Rogers and J. Graham, *Proceedings of the seventh European Conference on Computer Vision, Copenhagen, 2002 (vol. 4). A. Heyden, G. Sparr, M. Nielsen, P. Johansen (eds.). Lecture Notes in Computer Science 2353, Springer-Verlag, Berlin. pp 517-530.* doi: 10.1007/3-540-47979-1_35

Robust Active Shape Model Search

Mike Rogers and Jim Graham

Division of Imaging Science and Biomedical Engineering, University of Manchester,
mike.rogers@man.ac.uk,
http://www.isbe.man.ac.uk/~mdr/personal.html

Abstract. Active shape models (ASMs) have been shown to be a powerful tool to aid the interpretation of images. ASM model parameter estimation is based on the assumption that residuals between model fit and data have a Gaussian distribution. However, in many real applications, specifically those found in the area of medical image analysis, this assumption may be inaccurate. Robust parameter estimation methods have been used elsewhere in machine vision and provide a promising method of improving ASM search performance. This paper formulates M-estimator and random sampling approaches to robust parameter estimation in the context of ASM search. These methods have been applied to several sets of medical images where ASM search robustness problems have previously been encountered. Robust parameter estimation is shown to increase tolerance to outliers, which can lead to improved search robustness and accuracy.

Keywords. Medical Image Understanding, Shape, Active Shape Models, Robust Parameter Estimation, M-estimators, RANSAC, Weighted Least Squares.

1 Introduction

Statistical shape models have been shown to be a powerful tool to aid the interpretation of images. Models represent the shape and variation of object classes and can be used to impose *a-priori* constraints on image search. A frequently used formulation, on which we shall concentrate in this paper, is the Active Shape Model (ASM) [3], which also provides a method of fitting the model to image data. ASMs have been applied to many image analysis tasks, most successfully when the object class of interest is fairly consistent in shape and gray–level appearance [3][13][16]. The technique can suffer from a lack of robustness when image evidence is noisy or highly variable [2][10]. Many medical images display these types of characteristics.

To fit a model to data, parameters must be estimated in an optimal manner. Standard ASM parameter estimation minimises the sum of squares of residuals between the model and the data. It has been widely recognised that least squares minimisation only yields optimal results when the assumption of Gaussian distributed noise is met. Under real conditions a Gaussian model of residual distribution is seldom accurate. Least squares estimation is especially sensitive

A. Heyden et al. (Eds.): ECCV 2002, LNCS 2353, pp. 517–530, 2002.
© Springer-Verlag Berlin Heidelberg 2002

to the presence of gross errors, or outliers [5]. Medical images containing widely varying appearance and detailed structure potentially give rise to non-Gaussian residuals, including outliers, breaking down the assumptions upon which ASM parameter estimation is built. Robust methods of parameter estimation provide a promising method of increasing the accuracy and robustness of ASM search.

Many computer vision problems are related to estimating parameters from noisy data. Robust minimisation techniques have been applied to many areas of machine vision. Torr and Murray [17] compare methods for the calculation of the Fundamental Matrix, the calibration-free representation of camera motion. Robust techniques have also been used in conic fitting [19], cartography [4], tracking [1] and registration [20].

In this paper we investigate the case of robust parameter estimation techniques for shape-model fitting. We use ASM search as a paradigm for shape fitting, although the methods may be applied using any other parameterisation of shape. We formulate several robust parameter estimation schemes and present a quantitative comparison of the methods against standard least squares parameter estimation using several sets of medical images. The image sets have been chosen because ASM search has previously been found to be insufficiently robust in locating object boundaries.

1.1 Statistical Shape Models

Here we describe briefly the shape models used to represent deformable object classes. ASMs are built from a training set of annotated images, in which corresponding points have been marked. The points from each image are represented as a vector $\boldsymbol{x}$ after alignment to a common co-ordinate frame [3]. Eigen-analysis is applied to the aligned shape vectors, producing a set of *modes of variation* $\boldsymbol{P}$. The model has parameters $\boldsymbol{b}$ controlling the shape represented as:

$$\boldsymbol{x} = \bar{\boldsymbol{x}} + \boldsymbol{P}\boldsymbol{b} \tag{1}$$

where $\bar{\boldsymbol{x}}$ is the mean aligned shape. The shape $\boldsymbol{x}$ can be placed in the image frame by applying an appropriate pose transform.

Neglecting alignment, the process of estimating model parameters $\boldsymbol{b}$ for a given shape $\boldsymbol{x}$ proceeds by minimising the residuals:

$$\boldsymbol{r} = (\boldsymbol{x} - \boldsymbol{x}_0) \tag{2}$$

where $\boldsymbol{x}_0$ is the current reconstruction of the model. Least squares estimation therefore seeks to minimise $E_1(\boldsymbol{b}) = \boldsymbol{r}^T\boldsymbol{r}$, specifically we wish to find $\boldsymbol{\delta b}$ so as to minimise $E_1(\boldsymbol{b} + \boldsymbol{\delta b})$. This can be shown to have a solution of the form:

$$\boldsymbol{\delta b} = \left(\boldsymbol{P}^T\boldsymbol{P}\right)^{-1}\boldsymbol{P}^T\boldsymbol{\delta x} \tag{3}$$

which, as $\boldsymbol{P}$ is orthonormal, simplifies to:

$$\boldsymbol{\delta b} = \boldsymbol{P}^T\boldsymbol{\delta x}. \tag{4}$$

This is the standard ASM parameter update equation for iterative search [3] and as such is expected to operate in the presence of noise in the shape update vector $\boldsymbol{\delta x}$. The estimator is optimal with respect to a Gaussian noise model. In the presence of non–Gaussian noise the estimator is suboptimal, specifically a single outlying value can significantly distort model parameters by an arbitrary amount.

In ASM search, the update vector $\boldsymbol{\delta x}$ is obtained by searching the local image area around each landmark point. Models of local image appearance for each landmark point are built from the training set. These are used at each iteration of search to determine the best local landmark position.

2 Robust Parameter Estimation

There are many robust estimation techniques in the literature [4][5][7][12][15][18]. For the purposes of ASM parameter estimation, they can be divided into two categories: M-estimators and random sampling techniques, which we describe in the following sections.

2.1 M-Estimators

The aim of M-estimators is to alter the influence of outlying values to make the distribution conform to Gaussian assumptions. The estimators minimise the sum of a symmetric positive definite function:

$$\min \sum_i \rho(r_i). \tag{5}$$

where r_i is the ith element of the residual vector $\boldsymbol{r} = (r_1, r_2, \ldots, r_n)$.

The M-estimator of $\boldsymbol{b}$ based on the function $\rho(r_i)$ is the solution to the following m equations:

$$\sum_i \psi(r_i) \frac{\partial r_i}{\partial b_j} = 0, \text{for} j = 1, \ldots, m, \tag{6}$$

where the derivative $\psi(x) = d\rho(x)/dx$ is called the influence function, representing the the influence of a datum on the value of the parameter estimate. A weighting function is defined as:

$$\omega(x) = \frac{\psi(x)}{x} \tag{7}$$

and (6) becomes:

$$\sum_i \omega(r_i) r_i \frac{\partial r_i}{\partial b_j} = 0, \text{for} j = 1, \ldots, m. \tag{8}$$

which is a weighted least squares problem. In the ASM formulation $\partial \boldsymbol{r}/\partial \boldsymbol{b}$ is given by $\boldsymbol{P}$, resulting in a solution with the form:

$$\boldsymbol{\delta b} = \boldsymbol{K}^T \boldsymbol{\delta x} \tag{9}$$

where $\boldsymbol{K} = (\boldsymbol{P}^T \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{P})^{-1} \boldsymbol{P}^T \boldsymbol{W}^T \boldsymbol{W}$ and $\boldsymbol{W}$ is a diagonal matrix formed from the weights ω_i.

Weighting Strategies. There are many possible forms for the function $\rho(x)$, a summary is given in [19]. Each is designed to weight the influence of residuals under non–Gaussian conditions. One of the most consistent and widely used forms was devised by Huber [7], which results in a weighting function of:

$$\omega_i = \begin{cases} 1 & r_i < \sigma \\ \sigma/|r_i| & \sigma \leq r_i < 3\sigma \\ 0 & r_i \geq 3\sigma \end{cases} \tag{10}$$

where σ is an estimate of the standard deviation of the residuals. The standard deviation σ is not known, but can be robustly estimated from the median of the absolute values of the residuals [12]:

$$\sigma = 1.4826(1 + 5/(n - p))\text{median}|r_i| \tag{11}$$

where n is the number of data points and p is the length of the parameter vector.

Equations 10 and 11 allow the calculation of a set of weights, which can be applied in Eqn. 9 to calculate model parameter updates. This process must be iterated to convergence with re-weighting at each stage to form a final parameter estimate. We will refer to this weighted least squares method as WLS Huber.

In ASM search, profile models are used to generate the update shape vector $\boldsymbol{\delta x}$, by searching local image regions. The positions at which image data has the smallest Mahalanobis distance d from the profile models are chosen as the new landmarks for model parameter estimation. Bearing this in mind, an alternative weighting scheme can be devised for ASM model fitting that draws on the likelihoods of profile matches. Mahalanobis distance d is distributed as a χ^2 with $(p - 1)$ degrees of freedom, where p is the number of parameters of the profile model, and can therefore be used to generate a probability for each element of $\boldsymbol{x}$ given the model in the standard way. These probabilities $\boldsymbol{p}$, one for each element of $\boldsymbol{x}$, can be used directly in a weighted least squares estimator. This estimator reflects the quality of the image evidence from which the data resulted, rather than the spatial distribution of points from which the model parameters are to be estimated. We refer to this as WLS Image.

2.2 Random Sampling

Random sample based robust parameter estimation is, in some sense, the opposite approach to the smoothing effect of least squares and iterated weighted least squares. Rather than maximising the amount of data used to obtain an initial solution and then identifying outliers, as small a subset of the data as is feasible is used to estimate model parameters. This process is repeated enough times to

ensure that within a some level of probability at least one of the subsets will contain only good data.

One of the first robust estimation algorithms was random sample consensus (RANSAC), introduced by Fischler and Bolles [4]. RANSAC proceeds by selecting random subsets of data and evaluating them in terms of the amount of data that is consistent with the resulting model parameterisation. After a certain number of trails, the parameters with the largest *consensus set* is accepted as the parameter estimate. A threshold can be set to stop the algorithm when an acceptable consensus has been achieved. In order to determine the consensus set, a distance measure between model and data points must be defined, together with a threshold on this value.

In the case of ASMs, parameter estimation from random subsets of data can easily be achieved by a weighted least squares scheme with binary weights. The size of the consensus set can be determined by thresholding residual values after model reconstruction in the image co-ordinate system.

A later example of a random sampling algorithm was least median of squares (LMedS), proposed by Rousseeuw [12]. LMedS estimates model parameters by minimising the non-linear function:

$$\text{median} \boldsymbol{r}^T \boldsymbol{r}. \tag{12}$$

The algorithm is in fact extremely similar to RANSAC, the major differences being that LMedS does not require a consensus threshold, and unlike RANSAC no threshold is defined to end further random sampling.

In both algorithms we would ideally like to consider every possible subset of the data. This is usually infeasible, so methods are required to calculate the largest number of subsets required to guarantee a subset containing only good data. Assuming the proportion of outliers in the data is ϵ, the probability that at least one of m subsets consists of only good data is given by:

$$\gamma = 1 - (1 - (1 - \epsilon)^p)^m \tag{13}$$

where p is the size of each subset. If $\gamma \to 1$, then

$$m = \frac{\log(1 - \gamma)}{\log(1 - (1 - \epsilon)^p)} \tag{14}$$

In practice ϵ must be estimated by an educated worst guess of the proportion of outliers. Commonly $\gamma \geq 0.95$. We note that LMedS is computationally inefficient in the presence of Gaussian noise [12] as a fixed number of parameter estimations are always carried out. Both RANSAC and LMedS can be optimised for specific tasks [20].

We note that other random sampling algorithms exist in the literature, for example: least trimmed squares [11], MINPRAM [14] and MLESAC [18][1]. These techniques have not been evaluated here because they are closely related to

[1] For a review of random sampling techniques see [8].

RANSAC and LMedS. For application to ASM parameter estimation, it is expected that any suitable random sampling robust parameter estimation algorithm will produce similar results to RANSAC or LMedS.

3 Experiments

To test the effects of using a robust ASM parameter estimation technique on image interpretation, we have carried out a set of experiments on three difference image sets. The image sets we have chosen are: electron microscope images of diabetic nerve capillaries [10], Echocardiograms of the left ventricle [6] and magnetic resonance images (MRI) of the prostate [2]. Each set has been chosen because of the poor performance achieved using ASM interpretation in previous studies. Each type of image presents its own set of problems which must be addressed to achieve optimal interpretation results. Details of the training data available for each data set and the various modifications to the standard ASM algorithm are as follows:

- **Capillaries.** The training set consists of 33 electron microscope images digitised at 575×678 pixel 8-bit grey-scale. Each image has been annotated at the basement membrane/endothelial cell boundary up to 4 times by an expert on separate occasions, giving a set of 131 annotations. There is a large amount of ambiguity in the position of desired boundary, caused in part by locally consistent but confusing image evidence, resulting in considerable variability in expert-placed landmarks. A smoothed version of each annotated boundary has been represented by 50 evenly spaced landmark points. For this data, ASM profile models consist of a two cluster mixture model, where one cluster represents normal profile appearance and one for the locally confusing evidence. The two classes of profile appearance have been selected manually. The images are extremely complex and variable. Wavelet texture analysis has been applied to the images to remove some of this complexity. The ASM model built for this set of data has been found to impose only weak constraints due to the small amount of consistent structure in the capillary boundary shapes [9]. Figure 1 shows examples of these images.

- **Left Ventricle Echocardiograms.** The training set consists of 64 echocardiogram images digitised at 256×256 pixel 8-bit grey-scale. Each image has been expertly annotated with a plausible position of the left ventricle boundary. The boundaries have been represented by 100 evenly spaced landmarks. Echocardiogram images are inherently noisy and structural delineations are often poorly defined. ASMs have previously been applied to this data set using a genetic algorithm search technique [6] that identified many ambiguous possible model fit positions. Figure 2 shows some examples from this data.
- **Prostate.** This training set consists of 95 T2 weighted MRI images of the prostate and surrounding tissues at differing anterior depths, digitised at 256×256 pixel 8-bit gray scale. Annotations of the perimeter of the prostate

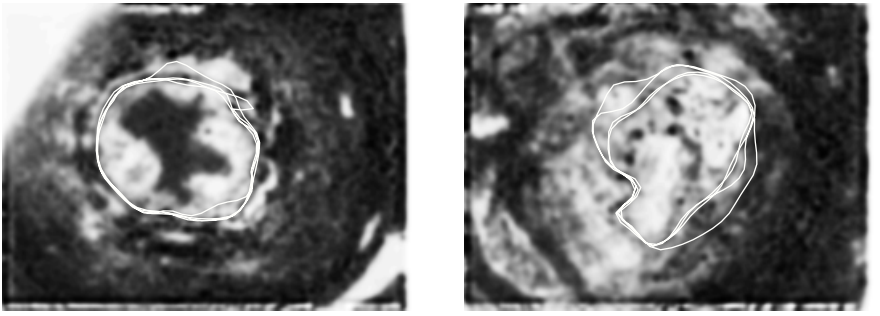


Fig. 1. Example capillary texture images with multiple ambiguous expert annotations.

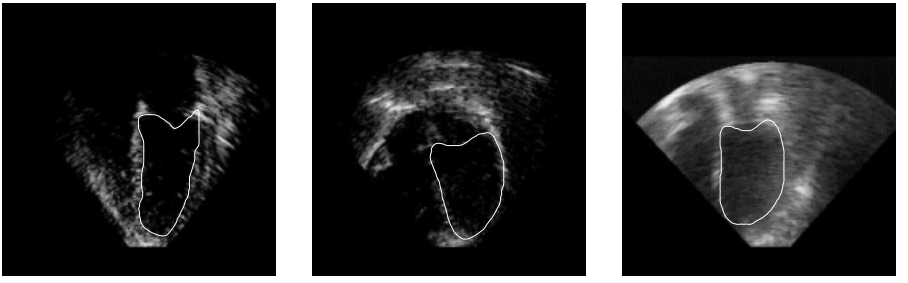


Fig. 2. Example echocardiogram images with left ventricle boundary marked.

gland, consisting of 27 manually positioned landmarks, were used to train the ASM. There is significant variation in the structure and appearance of the tissue surrounding the prostate as the depth of the image slice varies. This has been found to adversely affect ASM search [2]. Figure 3 shows some examples from this data set.

On each image set we have evaluated several robust parameter estimation methods in terms of robustness and accuracy: simple least squares (LS), weighted least squares using Huber's iterated scheme (WLS Huber), weighted least squares with weights obtained directly from the image data (WLS Image), RANSAC and LMedS.

3.1 Random Sampling Subset Size

Before we can apply a random sampling technique to ASM parameter estimation, we must determine the smallest subset size that is feasible to instantiate the model parameters. There is no precise fixed method of directly determining

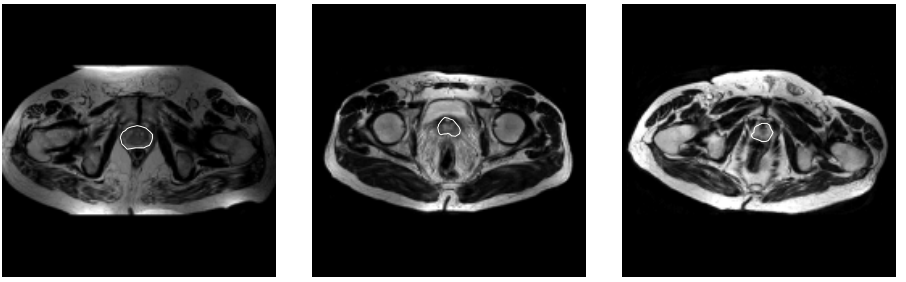


Fig. 3. Example prostate images with prostate perimeter marked.

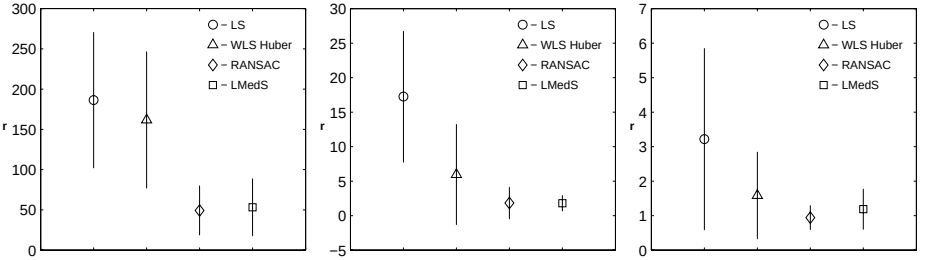
this value. Rather, we can tune the subset size to achieve a certain accuracy whilst avoiding degenerate matrices. When constructing an ASM, the number of modes kept can be chosen to ensure that the model's training set is represented to a given accuracy [3]. A similar approach can be taken to determine the random subset size. A model's training set can be reconstructed using a RANSAC or LMedS approach with varying subset size and the corresponding residuals recorded. The random subset size can then be chosen to give a desired reconstruction accuracy, bearing in mind that larger subsets require far more random trials to ensure the same probability of considering a good subset.

Table 1 shows RANSAC training set reconstruction errors for a model built with 133 capillary boundaries. The table also shows the number of trails required to obtain a 95% probability of the good subset under the assumption that 10% of data is outlying.

Table 1. RANSAC training set reconstruction error for varying subset size. p is the proportion of the total landmark points in each subset, $\bar{\boldsymbol{r}}$ is the mean reconstruction residual in pixels across the entire training set and m is the number of trials required to obtain a good subset with 95% certainty under the assumption of $\epsilon = 10\%$.

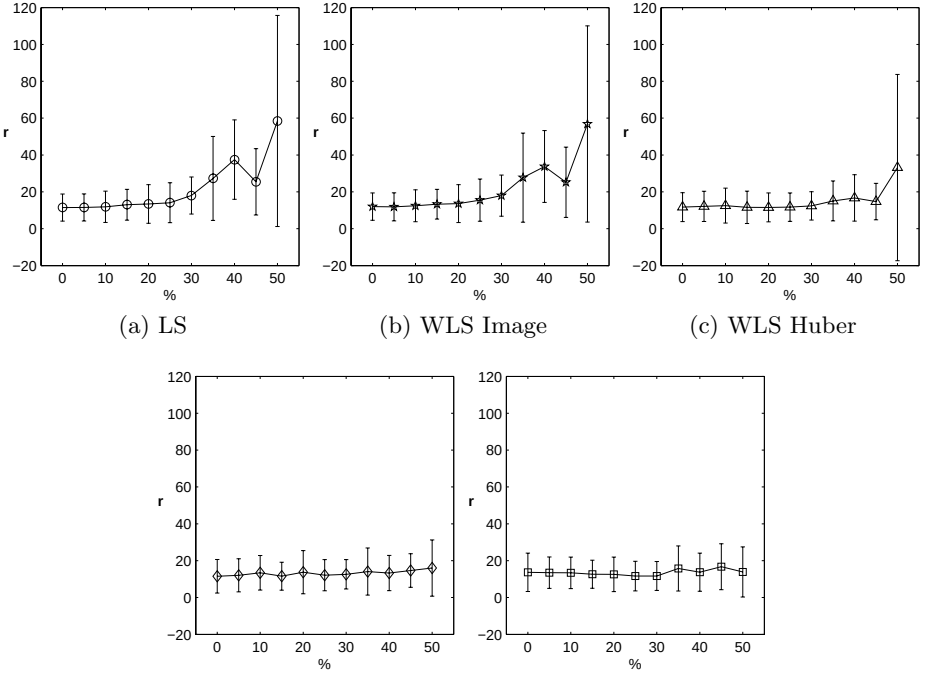
p	$\bar{\boldsymbol{r}}$	m
0.3	5.51	12
0.4	2.40	23
0.5	1.74	40
0.6	1.24	69
0.7	1.00	118
0.8	0.95	201
0.9	0.90	341

In the following evaluations we have chosen RANSAC and LMedS subset sizes of 0.4, 0.3 and 0.8 for capillary, left ventricle echocardiogram and prostate images respectively.



(a) Capillary (b) Ventricle (c) Prostate

Fig. 4. Outlier tolerance for each set of data. Mean residual $\boldsymbol{r}$ over model training set perturbed by non-Gaussian noise. Error bars show ± 1 std. dev.



(a) LS (b) WLS Image (c) WLS Huber

(d) RANSAC (e) LMedS

Fig. 5. Capillary Robustness. Mean residual $\boldsymbol{r}$ is plotted against initialisation outlier σ as a percentage of training set extent.

3.2 Outlier Tolerance

To investigate the effectiveness of the various parameter estimation methods under known conditions, models were fitted to data that had been perturbed by

non-Gaussian noise. Each annotation landmark in each data set was perturbed by Gaussian noise with standard deviation $\sigma = 0.5$ pixels. 30% of the landmarks in each annotation were selected at random and perturbed again by Gaussian noise with σ =50% of the maximum extent of the training set annotations. Model parameters were estimated using each method and the residuals calculated between the original data and the model representation. WLS Image was not considered in this case. Figure 4 shows the mean residual for each set of data and each method, together with error bars of ± 1 standard deviation. In each case, the random sampling methods outperform LS and WLS Huber. This is as expected as studies have reported that WLS methods are only robust when less than 20 – 25% of the data is outlying [17]. It is noteworthy that the worst performance for WLS Huber is exhibited when applied to the capillary model, which has been constructed from highly variable shapes with little consistent structure. The model contains only weak constraints on capillary shape [9] which contributes to the poor WLS Huber performance.

3.3 ASM Robustness

To investigate the effects of robust parameter estimation on ASM search, a set of leave-one-out searches were performed. A subset of 10 images were selected at random from each set of data and a single iteration of ASM search was carried out on each. Each search was initiated from the position and parameters of the correct annotated boundary, distorted by applying Gaussian noise to the landmark positions and randomly creating outliers, as in section 3.2. In this case 10% of points were made outliers. The size of σ used to create the outlier set was varied between 0 and 50% of the training set extent. The final residuals between the model fit position and the annotation(s) were recorded. In the case of the multiple capillary annotations, the smallest residual for each model point in each image was used to form $\bar{\boldsymbol{r}}$. Figures 5–7 show residual means and standard deviations for each set of data and method.

In each of the evaluations, the search gives comparable results when outlier distances are small. However, in each case the random sampling methods are much more robust in the presence of large outlying distances. WLS Huber estimation is more robust than the LS and WLS Image estimators, only breaking down when the outlier σ becomes large. LS and WLS Image estimators are approximately equivalent for each set of data. This suggests that there is no useful information in the quality of profile model matches to image data. In the case of capillary images it has previously been hypothesised that 'good' profile model matches are often found in inaccurate positions [9]. The similarity of LS and WLS Image supports this assumption.

3.4 ASM Accuracy

The utility of robust techniques in practical situations has been evaluated by performing a set of ASM searches. Models were fitted to the images used above

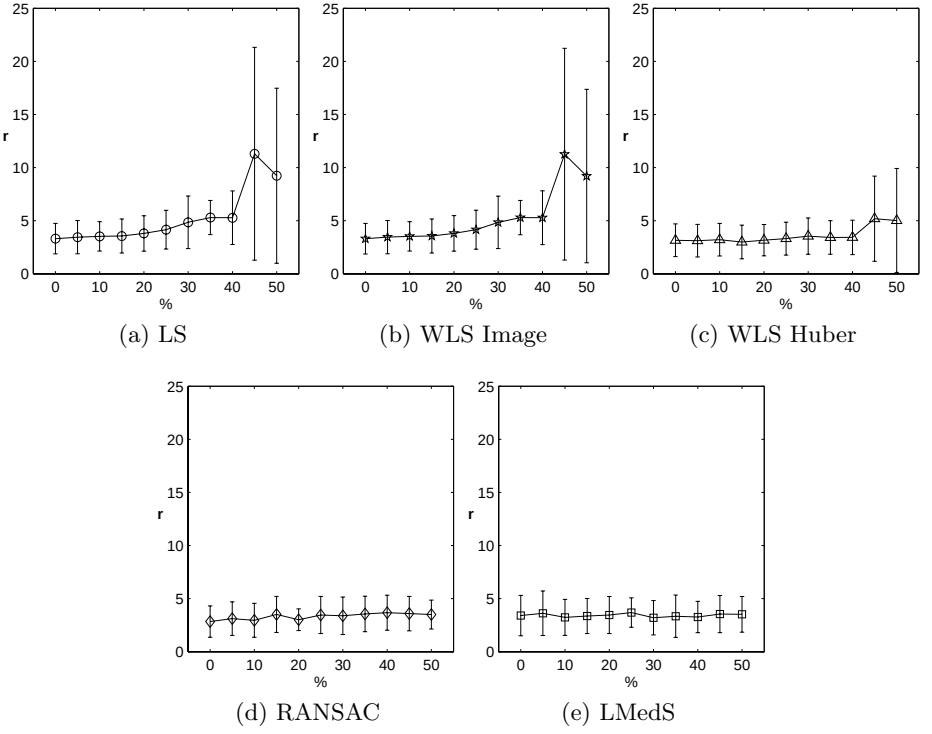


Fig. 6. Ventricle Robustness. Mean residual $\boldsymbol{r}$ is plotted against initialisation outlier σ as a percentage of training set extent.

using a single–resolution ASM search, initialised from the model mean shape and pose. Results from each set of data and each method are shown in Fig. 8.

In general, random sampling parameter estimation improves search accuracy and consistency compared to LS and WLS Image. This is shown well in prostate search results, where LMedS gives a reduction in average residual of 52%. Improvements in search accuracy using the weakly constrained capillary model are marginal. Capillary images contain many regions of locally confusing image evidence that consistently attract profile model matches. These areas can be see as the light image regions outside of the annotations of the capillary images in Fig. 1. The constraints imposed by the capillary model are not sufficient to identify these matches as inconsistent with the global trend of all profile model matches. Because of this, random sampling and M-estimator robust parameter estimators do not identify the data as outlying and performance does not improve. In general, WLS Image does not provide a reliable scheme to improve search accuracy. In each set of searches the other robust estimators result in better search accuracy than WLS Image. The approach actually degrades search accuracy in capillary images, an effect caused by the regions of locally misleading evidence.

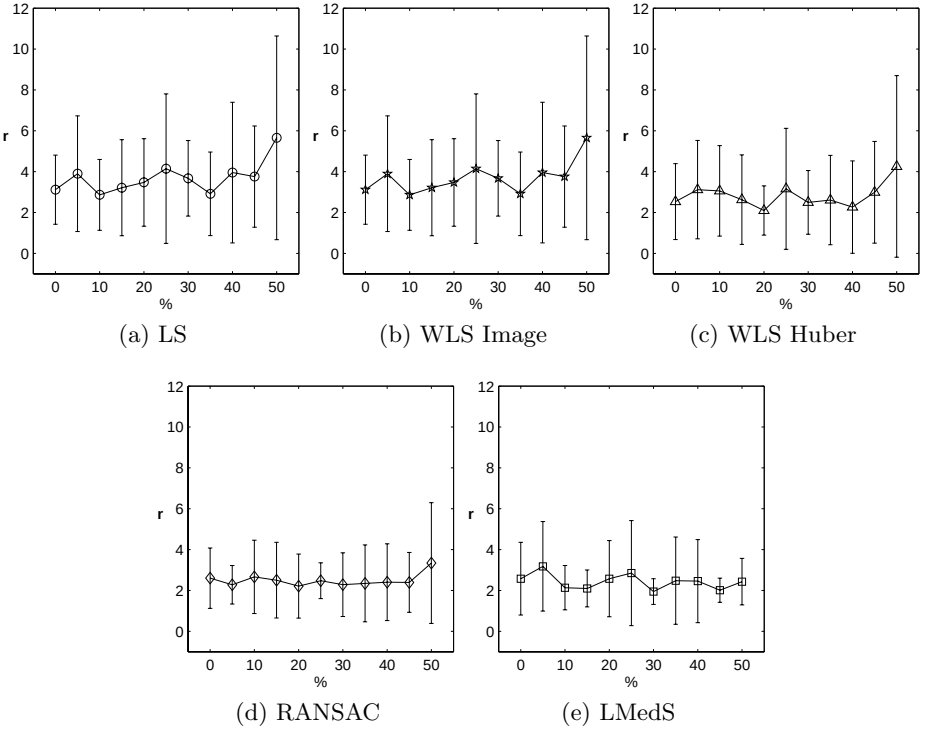


Fig. 7. Prostate Robustness. Mean residual $\boldsymbol{r}$ is plotted against initialisation outlier σ as a percentage of training set extent.

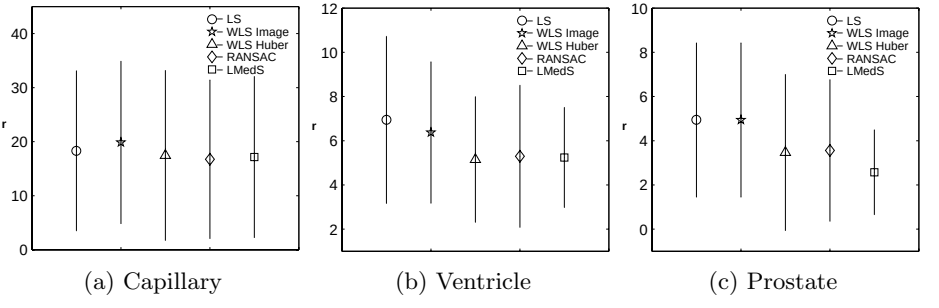


Fig. 8. ASM Search Accuracy. Mean residual $\boldsymbol{r}$ is plotted for each set of searches and each parameter estimation method. Error bars show ± 1 std. dev.

4 Conclusions

In many practical applications of shape model fitting, the assumption that residuals will have a Gaussian distribution will not be accurate. In particular, the presence of confusing evidence, due to noise, or highly variable image structure, produces "outliers" in the set of image points used to estimate the model parameters. We have evaluated four methods of parameter estimation intended to provide increased robustness to outliers in the fitting process, comparing these with the standard least squares approach. Two of the methods (RANSAC and LMedS) are examples of random sampling techniques. WLS Huber is a commonly used M-estimator. All three approaches gave improved robustness over the LS fit in the presence of synthetically generated outliers, the random sampling methods giving the best results. In general, the increased robustness results in increased search accuracy. This improvement is small in the case of capillary images because of the weakness of the constraints that can be imposed with models of capillary shape. The searches using WLS Image were intended to investigate the effect of reducing the influence in the fitting of points where the image data do not conform strongly to the local grey-level model. The hypothesis here is that most outliers occur because there is no strong image evidence in the local search area. Fitting should be improved if the influence of such data is reduced. The absence of any improvement using this technique indicates that outliers often occur because isolated points are found that correspond well to the model image patch.

The increased robustness of the random sampling and M-estimator methods is, of course, gained at the expense of increased computational cost.

Robust model parameterisation by itself is not a complete solution to the complex medical image analysis problems addressed in this paper. However, robust estimators have been shown to improve the robustness and accuracy of ASM search, and are potentially a useful modification to the standard ASM algorithm in many practical situations.

References

[1] M. J. Black and A. D. Jepson. EigenTracking: Robust matching and tracking og articulated objects using a view-based representation. In *Proceedings of the European Conference on Computer Vision*, pages 329–342. Springler-Verlag, 1996.

[2] A. Chauhan. The use of active shapes models for the segmentation of the prostate gland from magnetic resonance images. Master's thesis, University of Manchester, 2001.

[3] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active Shape Models - their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, Jan. 1995.

[4] M. A. Fischler and R. C. Bolles. Random Sample Consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commumications of the ACM*, 24(6):381–395, 1981.

[5] J. P. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics: An Approach Based on Influence Functions.* Wiley, New York, 1986.

[6] A. Hill, C. T. Taylor, and T. F. Cootes. Object recognition by flexible template matching using genetic algorithms. In *Proceedings of the 2nd European Conference on Computer Vision*, pages 852–856, Santa Margherita Ligure, Italy, May 1992.
[7] P. J. Huber. *Robust Statistics*. Wiley, New York, 1981.
[8] P. Meer, A. Mintz, and A. Rosenfeld. Robust regression methods for computer vision: A review. *International Journal of Computer Vision*, 6:59–70, 1991.
[9] M. Rogers. *Exploiting Weak Constraints on Object Shape and Structure for Segmentation of 2–D Images*. PhD thesis, University of Manchester, 2001.
[10] M. Rogers and J. Graham. Exploiting weak shape constraints to segment capillary images in microangiopathy. In *Proceedings of Medical Image Computing and Computer–Assisted Intervention*, pages 717–716, Pittsburg, USA, 2000.
[11] P. J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79:871–880, 1984.
[12] P. J. Rousseeuw. *Robust Regression and Outlier Detection*. Wiley, New York, 1987.
[13] S. Solloway, C. E. Hutchinson, J. C. Waterton, and C. J. Taylor. Quantification of articular cartilage from MR images using Active Shape Models. In B. Buxton and R. Cipolla, editors, *Proceedings of the 4th European Conference on Computer Vision*, volume 2, pages 400–411, Cambridge, England, April 1996. Springer-Verlag.
[14] C. V. Stewart. MINPRAM, a new robust estimator for computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-17(10):925–938, 1995.
[15] C. V. Stewart. Bias in robust estimation caused by discontinuities and multiple structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(8):818–833, Aug 1997.
[16] H. H. Thodberg and A. Rosholm. Application of the Active Shape Model in a commercial medical device for bone densitometry. In T. F. Cootes and C. J. Taylor, editors, *Proceedings of the 12th British Machine Vision Conference*, volume 1, pages 43–52, September 2001.
[17] P. H. S. Torr and D. W. Murray. The development and comparison of robust methods for estimating the fundamental matrix. *International Journal of Computer Vision*, 24(3):271–300, 1997.
[18] P. H. S. Torr and A. Zisserman. MLESAC: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding*, 78(1):138–156, April 2000.
[19] Z. Zhang. Parameter estimation techniques: A tutorial with application to conic fitting. *Image and Vision Computing*, 15:59–76, 1997.
[20] Z. Zhang, R. Deriche, O. Faugeras, and Q.-T. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence Journal*, 78:87–119, October 1995.

29. **Detecting asymmetries in hippocampal shape and receptor distribution using statistical appearance models and linear discriminant analysis.** D. Poxton, J. Graham and J.F.W. Deakin, *Proceedings of the British Machine Vision Conference, University of Southampton, 1998. P.H. Lewis and M.S. Nixon (eds.) BMVA Press, pp 525-534.* doi: 10.5244/C.12.

Detecting Asymmetries in Hippocampal Shape and Receptor Distribution using Statistical Appearance Models and Linear Discriminant Analysis

D. Poxton[a], J. Graham[a] and J.F.W. Deakin[b]
[a]Dept. Medical Biophysics and [b]Dept. Psychiatry, Manchester University, UK.
dp@sv1.smb.man.ac.uk

Abstract

Neurological studies are often concerned with identifying abnormalities in brain structure affecting asymmetry between left and right hemispheres. This paper presents techniques which allow measurement and characterisation of differences between neuroanatomic structures due to variation in both shape and receptor distribution. This provides a potentially powerful tool for identifying subtle pathological asymmetries. We propose a combination of appearance modelling and linear discriminant analysis and present preliminary results of the technique applied to 2D hippocampal autoradiographs. We also describe experiments testing the relative performance of variants of our method to test assumptions about the nature of the analysis and the nature of the data.

1 Introduction

Despite many studies, the anatomical characteristics of the major neuropsychiatric disorders are still poorly understood. Furthermore, few rapid and sensitive techniques exist for characterising morphological variation of neural structure with which pathology can be identified. Presently, studies depend upon fairly coarse and simplistic measurements such as anatomic volume or thickness, measures which are unable to register anything other than the most gross of structural and neurochemical abnormalities. This may be particularly inappropriate for complex 3D structures such as the hippocampus, a region often associated with schizophrenic patholologies[1].A specific area of investigation is concerned with the identification of pathological asymmetries between structures located in either hemisphere of the brain. For example, studies suggest that normal asymmetries of the brain are far less in schizophrenics, some imaging studies reporting loss or reversal [2, 6], although other studies conflict with these results [7]. This paper describes a method which can be used to accurately identify subtle asymmetry of neuroanatomy.

In order to confirm theories correlating psychological disorders with types of neurological pathology, it is required that both structure and neurochemical make-up of a region can be determined. Analysing the distribution of neurotransmitters can often reveal variations which are indicative of altered neuronal development. To this end, our technique is applicable to both shape **and** receptor distributions made visible using autoradiography.

BMVC 1998 doi:10.5244/C.12.53

In developing methods for identifying lateral asymmetries, a key issue is sensitivity. In structurally simple regions, such as the cortex, comparisons may be quite straightforward; methods such as the construction and averaging of depth profiles may suffice. However, more complex regions are not amenable to such simple approaches. The hippocampus, a highly concave and reentrant structure located in the temporal lobe, is an ideal test subject for any technique which seeks to identify complex or subtle asymmetry. Whilst a 3D analysis is the eventual aim of this project, we present preliminary results of our technique applied to 2D postmortem autoradiographic sections of the hippocampus.

2 Materials and Methods

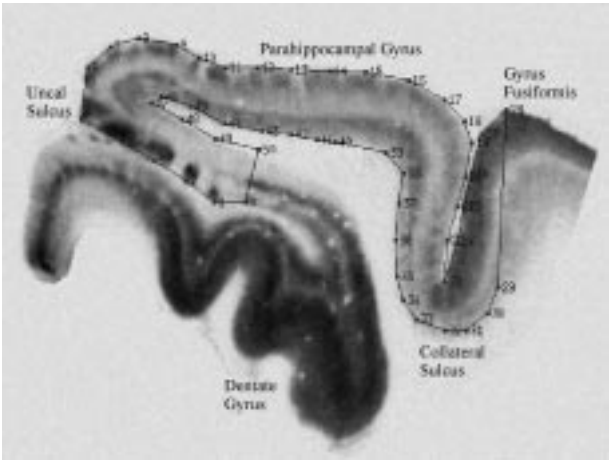


Figure 1: Hippocampal Autoradigraph

The hippocampal tissue used as test data comes from five normal brains: subjects free from a personal or family history of neurological or psychiatric disease. Both hemispheres of hippocampal tissue were cryosectioned, $20\mu m$ sections cut every $100\mu m$. Sections were stained with 8-hydroxy-2-(N,N-di-N-propyl-amino) tetralin ([3H]-8-OH-DPAT), selectively labelling 5-HT1A receptors, which are located in restricted classes of neuronal cells. The sections were then washed to remove unbound ligands, dried rapidly and exposed to high resolution tritium sensitive x-ray film for 8-12 weeks. In the resulting autoradiographs grey-level intensity represents receptor intensity. For the purpose of our 2D analysis, a single section located at a consistent anterior depth was selected from each hippocampal hemisphere. Analysis was centred on the relatively stable parahippocampal gyrus rather than the entire hippocampus, because of the intrinsic anterior-posterior variation of regions such as the dentate gyrus.

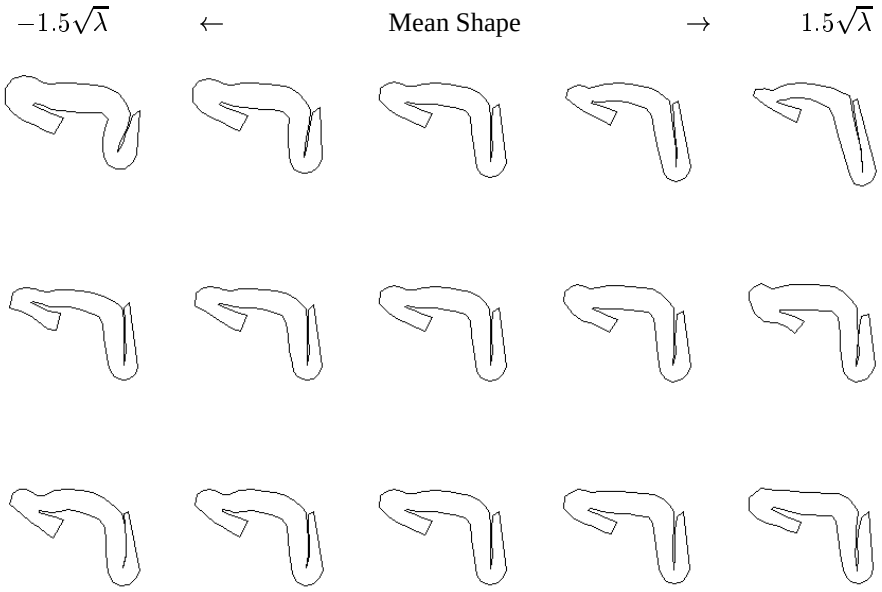


Figure 2: Shape Modes : Combined hemisphere model

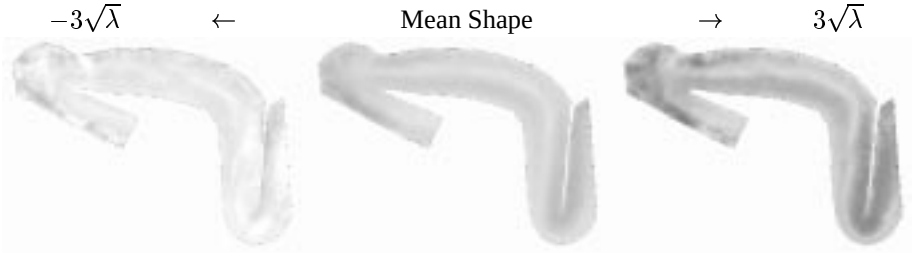


Figure 3: Grey-level Mode : Combined hemisphere model

2.1 Point Distribution Models

Our method of identifying shape and grey-level asymmetries employs the point distribution model and appearance modelling techniques presented by Cootes et al[3]. The shape information is captured by labelling the training images with consistent landmark points (See Figure 1). Our training set was labelled under the guidance of a neurologist and with the aid of a semi-automatic point planting software. Landmarks were typically curvature extrema or distinctive regions of receptor intensity, supplemented by uniformly spaced points between.

Each training image x_i labelled with p point coordinates can be described by its $2p$ shape vector $(x_{i0}, y_{i0}, x_{i1}, y_{i1} \ldots x_{ip-1}, y_{ip-1})^T$. It will be shown how the training set of shape vectors can be used to identify shape differences between left and right hemisphere hippocampi.

Grey-level information can also be expressed as a vector composed of the grey-level

intensity of pixels making up the hippocampus. However, before these can be constructed, variation due to shape must be eliminated. This is achieved by warping all training images to the mean shape calculated from the training shape vectors. For each training image we now have a grey-level vector $(x_{i0}, x_{i1}, \ldots x_{im-1},)^T$ where m is the number of pixels contained within the boundary of the mean shape. See related work by Lanitis et al[5].

Normal variation in the training set can be specified by performing principal components analysis on the shape and grey-level vectors. This generates a set of *modes of variation* : eigenvectors of the covariance matrix which span a shape or grey-level space of dimension considerably smaller than $2p$ (or m).

In addition to the data compaction, the modes *characterise* the principal ways in which the training set varies. Figure 2 shows the first three shape modes of a model built from hemispheres of *both* left and right hemisphere hippocampi. The most significant mode shows a lengthening of the collateral sulcus with an associated thinning of the parahippocampal gyrus. The second most significant mode shows some vertical movement of the right had side of the collateral sulcus. The third mode shows some bending and bowing of both the parahippocampal gyrus and the collateral sulcus.

Figure 3 shows the most significant grey-level mode superimposed onto a mean hippocampus shape. The variation described seems mainly to do with global increases in receptor intensity.

Whilst these modes may contain some of the variation between left and right hemisphere hippocampi, we cannot guarantee that they do so specifically and at the exclusion of other variations. Principal component analysis identifies the variation *within* a single training set. We require a technique which identifies variation *between* two training sets.

2.2 Linear Discriminant Analysis

We can think of each training example as a point in a space of high dimensionality. The task of identifying shape and grey-level differences between left and right hemisphere hippocampi can be viewed as the the task of separating two groups of points in this space. Linear discriminant analysis is a statistical technique which seeks to maximise the difference between the two groups.

In figure 6, we see how the discriminant vector, represented by the dashed line a provides an axis onto which the point distribution can be projected, maximally seperating the two groups. On this axis we can perform scalar measurements of separation between the groups. Furthermore, the discriminant vector characterises the group separation. Imagine a point resting on the vector at t: moving one way along the vector makes the point more like the first group, moving it the other way makes the point more like the other group.

Given a training set of points divided into two groups, how do we calculate the coefficients which ensure the discriminant function maximally separates the two groups?

A metric which describes the separation between two groups x_1 and x_2, subject to an arbitrary discriminant coefficient vector a, is :

$$V = \frac{a^T \overline{x_1} - a^T \overline{x_2}}{a^T W a} \tag{1}$$

where $\overline{x_1}$ and $\overline{x_2}$ are vectors of dimension $2p$, representing the means of groups 1 and 2 respectively, and W the *pooled within-class covariance matrix* given by

$$W = \frac{1}{n_{x_1} + n_{x_2} - 2} \sum_{i=1}^{2} \sum_{j=1}^{n} (x_{ij} - \overline{x_i})(x_{ij} - \overline{x_i})^T \quad (2)$$

nx_1 and n_{x_2} denote the number of members in groups 1 and 2 respectively.) W is simply the sum of the covariance matrices for groups 1 and 2.

So what does the metric described by equation 1 mean ? The term $a^T\overline{x_1} - a^T\overline{x_2}$ simply projects the means of both groups onto the discriminant vector formed by the coefficients a, and calculates their difference. This is intuitive : as the distance between the groups increases, so must the separation of their means, and so equation 1 is maximised. The term $a^T W a$ projects the pooled covariance matrix into a pooled variance value in the 1-D discriminant space. The smaller the variance of both groups (and hence the pooled variance value), the less likely they will be to overlap and hence their separation will increase. So as the variance decreases, so equation 1 is maximised.

Differentiating equation 1 with respect to a yields *Fishers Linear Discriminant Function* :

$$a = cW^{-1}(\overline{x_1} - \overline{x_2}) \quad (3)$$

where c is a scaling factor.

The discriminant coefficient vector a is a linear combination which maximally separates group x_1 from group x_2.(See ref[4]).

2.3 Paired Linear Discriminant Analysis

The definition of discriminant analysis provided above is phrased in terms of a separation between two groups. However, in the case of our hippocampal asymmetries we cannot be sure that such global distinctions between left and right hemispheres exist. In order to gain some feeling for how asymmetries may be expressed in the training set, the data was inspected in the following manner. Left and right hemisphere hippocampi were projected into the parameter space provided by the modes of a principal component analysis. With a reduced parameter space, it becomes possible to visualize the training set.

Figure 4 shows the hemispheres of the five brains projected onto the three most significant modes of shape variation (covering 85 per cent of all training set variation).The annotation of a point with the prefix *r* indicates a hippocampus from the right hemisphere, whilst *l* indicates a hippocampus from the left hemisphere. Points sharing the same symbol type indicate hippocampi from the same brain. As can be seen from Figure 4 the training set does not separate readily into distinct left and right hemisphere groups.

However, if we examine the training set purely on the basis of the most significant mode (representing 55 per cent of total variation) we can see that although the groups do not separate cleanly, the right hemisphere hippocampi have a *consistently* higher value than their left hemisphere partners (See Figure 5). So although there is no significant difference between the *group* of left hippocampi and the *group* of right hippocampi, there may be consistent differences between *pairs* of hippocampi from the same brain.

We propose a form of discriminant analysis which seeks to maximise separation between a *group of pairs* rather than a *pair of groups.* Figure 7 shows a distribution where a

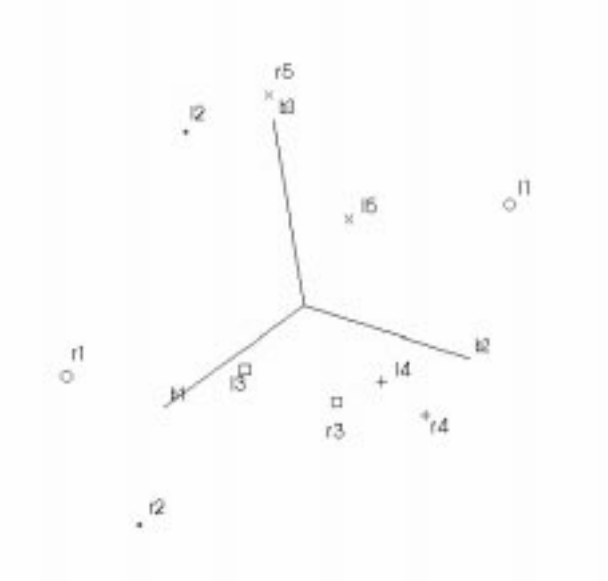


Figure 4: Training set projected into PCA space

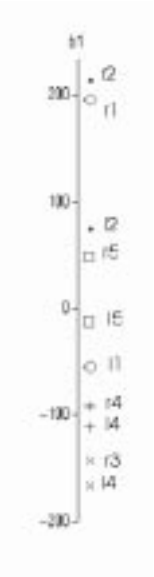


Figure 5: Training set projected onto most significant mode

set of pairs are maximised in their separation by a discriminant vector a. We modify the standard discriminant analysis scheme thus :

Let the ith pair of points in the distribution of n pairs be given by the m dimensional vectors:

$$x_{i1} = (x_1, x_2, ..., x_m) \qquad x_{i2} = (x_1, x_2, ..., x_m) \tag{4}$$

The difference between the ith pair is

$$d_i = x_{i1} - x_{i2} \tag{5}$$

and the mean difference is therefore

$$\overline{d} = \frac{1}{n} \sum_{i=1}^{n} d_i \tag{6}$$

We can define our *paired covariance matrix* as

$$P = \frac{1}{n-1} \sum_{i=1}^{n} (d_i - \overline{d})(d_i - \overline{d})^T \tag{7}$$

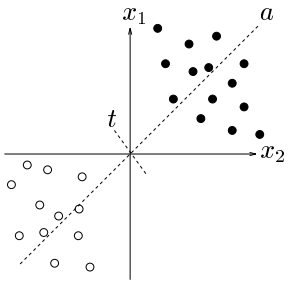


Figure 6: Discriminant mode for two populations

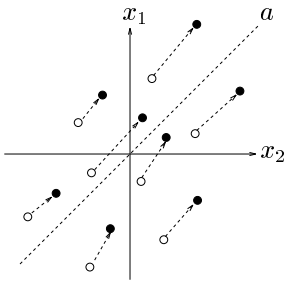


Figure 7: Discriminant mode for a paired population

where the variances are expressed in terms of differences between pairs. Using the same steps as in section 2.2, the set of coefficients which maximise paired separation are given as :

$$a = cP^{-1}\overline{d} \tag{8}$$

3 Experiment : Applying Discriminant Analysis to 2D Hippocampal Data

Although the theory behind standard and paired discriminant analysis is well founded, the assumption that hippocampal asymmetries are a paired rather than global phenomena is untested. The first task then, is to assess to what extent paired linear discriminant analysis produces better separations in hippocampal sections than standard discriminant analysis.

A second issue is what parameter space to perform the analysis on. Using principal component analysis, the dimensionality of the training vectors can be drastically reduced. With this in mind, comparisons need to be made to make clear whether the computational savings achieved by performing discriminant analysis on the reduced space are outweighed by any effects this may have on the detection of separations.

3.1 Experimental Procedure

The 10 hippocampal sections (5 left hemisphere and 5 right) were subject to discriminant analysis of shape and grey-level under the following conditions:

- **Paired Discriminant Analysis** : maximisation of separation between paired observations of data, *or* **Standard Discriminant Analysis** : maximisation of separation between two groups of data.
- **Reduced b-space Vectors** : training data composed of b-space vectors formed in construction of shape and grey-level models of combined hemisphere hippocampi (see section 2.1) *or* **Sample Space Vectors** : training data composed of vectors containing the coordinates of landmark points describing the hippocampal structure, or vectors of pixel grey-level values describing receptor distribution.

Each experiment will yield a set of discriminant coefficients, each of which allow the training set to be projected onto a one dimensional discriminant mode. Comparison of the separations provided by the different modes can then be performed. The metric proposed to allow quantitative comparison of separations is the t-test statistic. Although this test requires normality, which is certainly not guaranteed using our small data set, we only require a measure which gives an indication of the *relative* significance of separations over the different conditions.

4 Results

The t-test statistics and corresponding significance levels for the four different experimental conditions are presented in tables 1 and 2 . It is clear that paired discriminant analysis is providing a better description of the separation between the groups, particularly in the case of grey-level differences. Figures 8 and 9 provide visualisations of the shape and grey-level changes which occur along the axis of greatest separation between left and right hemispheres. In these visualisations the centre hippocampal section can be regarded as a section which is neutral of laterality, being an average of left and right hemispheres. Moving one way along the mode, makes the section more ”leftish” and the other way more “rightish”. The limit set for the variation in these visualisations is $l/2$, where l is the average separation between paired hemispheres.

$-\frac{l}{2}$ (left hemisphere) $\leftarrow$ Mean Shape $\rightarrow$ $+\frac{l}{2}$ (right hemisphere)

Figure 8: Paired discriminant mode for shape

$-\frac{l}{2}$ (left hemisphere) $\leftarrow$ Mean Grey-level $\rightarrow$ $+\frac{l}{2}$ (right hemisphere)

Figure 9: Paired discriminant mode for grey-level intensity

The visualisations demonstrate the the form of left-right hippocampal asymmetry. Left hemisphere hippocampi have longer and more vertically aligned collateral sulci than

Condition	Num. samples	Dimensions	t-value	Sig. level
Normal LDA (unpaired)				
Reduced b-space	10	9	1.22	74.3%
Full sample space	10	102	2.99	98.3%
Paired LDA				
Reduced b-space	5	9	2.35	92.2%
Full sample space	5	102	35.46	$> 99.9\%$

Table 1: Shape asymmetry significance levels over four experimental conditions

Condition	Num. samples	Dimensions	t-value	Sig. level
Normal LDA (unpaired)				
Reduced b-space	10	9	0.03	2.2%
Full sample space	10	40186	0.19	14.9%
Paired LDA				
Reduced b-space	5	9	3.56	97.6%
Full sample space	5	40186	4.39	98.8%

Table 2: Grey-level asymmetry significance sevels over four experimental conditions

right hemisphere hippocampi, whose collateral sulci are stumpy and often slanted in orientation. In addition, left hemisphere hippocampi have slightly straighter parahippocampal gyri than right hemisphere hippocampi, whose gyri are more bowed. (See Fig 1 for anatomical terms).

The grey-level discriminant mode is more difficult to interpret, although it can be said that most of the left/right asymmetry takes place in the top left hand corner, where the parahippocampal gyrus bends into the uncal sulcus. Although it is difficult to discern from these diagrams, animations show that right hemisphere hippocampi have a greater profusion of striations in the parahippocampal gyrus.

5 Discussion

The difference between left and right hemisphere populations is small in the context of natural variability amongst individuals. The paired discriminant analysis seeks to find a *consistent* mode of separation. The fact that a better separation is found by the paired analysis indicates that while the left and right populations might overlap in their shape and grey-level, the shifts between them are consistent. The paired discriminant analysis is clearly a better way of identifying a discriminant vector for groups which are paired.

The second issue is the performance of both discriminant techniques when applied to the model-space representation of the data set. The significance of the separations is not as great as that gained when using the full sample space. There are two points regarding this result. Firstly, the significance values for full parameter space seem suspiciously high. This is due to the fact that we are trying to locate a vector which separates only 10 pieces of data in a space of very high dimensionality: it is possible for many such vectors to be

located. Results must therefore be regarded cautiously. Secondly, the use of a reduced parameter space results in lower separations, possibly truncating some of the asymmetry we are hoping to identify. It is possible, and indeed even quite likely, that some of the asymmetries are quite small, and so are subsequently removed by the dimensional reduction taking place in the principal component analysis. However, the fact that significant separations are still detectable under such a reduction offers encouragement.

6 Summary

We have demonstrated that linear discriminant analysis, coupled with accurate landmarking of structure, provides a potentially powerful way of generating quantitative and specific descriptions of lateral asymmetries in hippocampal sections, both in shape and receptor distribution. We have presented a modified discriminant analysis scheme which detects paired asymmetries. The results suggest that whilst left-right shape asymmetries exist, and may be detected by considering the two hemispheres as groups; *paired* asymmetries due to shape *and* receptor distribution seem to be more pronounced on examining the *paired* differences.

At a particular level of the parahippocampal gyrus, we have identified specific lateral asymmetries. The significance of the measurements needs to be regarded with caution given the small data set available, but the initial result allows us to form the hypothesis that similar differences will be detected by a 3D study using the more substantial data set which is currently being collected for this project.

References

[1] L. Altshuler, M. Casanova, T. Goldberg, and J. Kleinman. The hippocampus and parahippocampus in schizophrenic, suicide and control brains. *Archives of General Psychiatry*, 47:36–42, 1990.

[2] P. Barta, G. Pearlson, I. McGilchrist, A. T. R.W. Lewis, A.Pulver, D. Vaughn, G. M.F. Casanova, and R. Powers. Reversal of asymmetry of the planum temporale in schizophrenia. *American Journal of Psychiatry*, 152:715–721, 1995.

[3] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models : their training and application. *Computer Vision and Image Understanding*, 61:38–59, 1995.

[4] W. Krzanowski and F. Marriot. *Kendall's Library of Statistics 2 : Multivariate Analysis*, page 5. Arnold, 1995.

[5] A. Lanitis, C. Taylor, and T. Cootes. Recognizing human faces using shape and grey-level information. In *Proceedings of 3rd International Conference on Automation, Robotics and Computer Vision*, volume 2, pages 1153–1157, 1994.

[6] R. Ross and P. Stratta. Planum temporale in schizophrenia: a magnetic resonance study. *Schizophrenia Research*, 152:19–22, 1992.

[7] K. Vladar, B. Fantae, D. Jones, K. J.J, and D. Weinberger. Normal asymmetry of the planum temporale in patients with schizophrenia - 3D cortical morphometry with MRI. *British Journal of Psychiatry*, 166:742–749, 1995.

30. **An Investigation of morphometric changes in the lateral ventricles of schizophrenic subjects.** K.O. Babalola, J. Graham, W. Honer, L. Kopala, D. Lang and R. Vandorpe, *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI 2003), Montreal, Canada, 2003,R.E. Ellis and T.M. Peters (eds.) (Lecture Notes in Computer Science 2879) Springer Berlin, pp 521-529.* doi: 10.1007/978-3-540-39903-2_64

An Investigation of Morphometric Changes in the Lateral Ventricles of Schizophrenic Subjects

Kolawole Babalola[1], Jim Graham[1], William Honer[2], Lili Kopala[3], Donna Lang[2], and Robert Vandorpe[3]

[1] University of Manchester, Imaging Science & Biomedical Engineering, Stopford Building, Oxford Road, Manchester, M13 9PT, United Kingdom
{Kola.Babalola, Jim.Graham}@man.ac.uk
[2] University of British Columbia, Department of Radiology and Center for Complex Disorders, Vancouver, BC, V5Z 1L8, Canada
[3] Dalhousie University, Department of Psychiatry and Department of Neuroimaging, Halifax, Nova Scotia, B2H 2G2, Canada

Abstract. We present results of morphometric analysis of the lateral ventricles of a group of schizophrenic and control subjects to investigate possible shape differences associated with schizophrenia. Our results show shape changes localised to three regions : the temporal horn (its tip near the amygdala, and along its body near the parahippocampal fissure), the central part of the lateral ventricles around the corpus callosum, and the tip of the anterior horn in the region of the frontal lobe. The differences in the temporal and anterior horns are in regions close to structures thought to be implicated in schizophrenia. The changes observed are the most significant changes ($p < 10^{-13}$) in shape parameters calculated using a 3D statistical shape descriptor (point distribution model). Corresponding points on the surface of the ventricles in the training set were obtained using an transportation-based method to match high curvature points.

1 Introduction

Schizophrenia is a serious brain disorder which is accompanied by altered brain structure. Interest in investigation of shape changes of the lateral ventricles due to schizophrenia can be attributed to the work of Johnstone et al. [7] who showed that schizophrenia is accompanied by an increase in the volume of the lateral ventricles. Several groups e.g. [5] [11], are currently developing methods to investigate whether specific localised shape changes occur in the lateral ventricles and other neuroanatomic structures due to schizophrenia and other brain diseases.

Because of the wide range of natural variability in the shape of structures in the human body, statistical approaches to measuring differences in shape are desirable. Statistical shape models (SSMs) use samples from control and/or disease populations, the training set, to learn the variability in the structures being modelled. They can therefore allow separation of shape changes due to disease in the presence of natural variation, and provide better characterisation

R.E. Ellis and T.M. Peters (Eds.): MICCAI 2003, LNCS 2879, pp. 521–529, 2003.
© Springer-Verlag Berlin Heidelberg 2003

of differences between populations than volumetric techniques. A diverse number of SSMs have been described. However, these all need a method of representing shape, establishing correspondence across the training set and obtaining shape differences qualitatively and/or quantitatively.

The particular SSM we use here is the point distribution model (PDM) [3], which characterises shape by a small number of "modes of shape variation", providing a compact parameterisation. We apply linear discriminant analysis (LDA - see e.g. [6]) to the shape parameters to characterise inter-group differences.

2 Related Work

Buckley et al. [2] use 48 manually defined landmarks corresponding to curvature extrema on the surface of the ventricles of 20 schizophrenic patients and 20 control subjects to investigate shape differences. They considered the whole ventricular system and reported no overall shape differences between the entire patient group and the entire schizophrenic group. However, when only the males of both groups were considered, significant shape differences were identified in the proximal juncture of the temporal horn and in the foramen of Monro.

Gerig et al. [5] performed shape analysis on the lateral ventricles of 5 pairs of monozygotic and 5 pairs of dizygotic twins. Ventricles were mapped to a unit sphere and decomposed into a summation of spherical harmonic functions. The first order harmonics were used to impose correspondence between points and the measure of shape differences was the mean squared distance between corresponding points on the surfaces. They showed that, without normalisation for ventricular size, no significant differences were seen between the two groups. However, after normalisation using the volumes of the ventricles, the right lateral ventricles of the two groups are significantly different. They concluded that shape measures reveal new information in addition to size or volumetric differences, which might assist in the understanding of structural differences due to neuroanatomical diseases.

Narr et al. [11] obtained average maps of anatomical differences based on voxel values of the limbic structures and the lateral ventricles of 25 schizophrenic and 28 control subjects. Their analysis showed that significant shape differences occurred in the left lateral ventricles. In particular, there was enlargement of the superior part of the lateral ventricle and the posterior horn. There were also noticeable differences in the part of the lateral ventricles in the vicinity of the caudate head.

Our approach has aspects in common with [2] and [5]. We build PDMs based on corresponding landmark points across a training set. The landmark points are used to generate a small number of shape parameters controlling the modes of variation of the shapes. The use of this parametric description distinguishes our approach from that of [2]. However, the parameters are devised from the training data, unlike those of [5].

3 Materials and Method

3.1 Data

Volumetric T2 MR scans of 30 controls (14-45 years, 13 female, 17 male) and 39 age and sex matched schizophrenics (14-45 years, 9 female, 30 male) were used in this study. The scans were independently acquired in the sagittal, coronal and axial orientations. Each slice had 256 x 256 voxels, with in-plane size of 0.86mm by 0.86mm for sagittal and axial orientations, and 0.78mm by 0.78mm for the coronal orientation. For all orientations the slice thickness was 5mm and the intra-slice gap was 1mm.

All images were corrected for MR inhomogeniety [15], and the three views of each subject were combined by rigid registration and interpolation to give 3D images with effective resolution of 0.78mm x 0.78mm x 0.78mm. The lateral ventricles were segmented using a 3D edge detector [9] to give edge segments which were manually linked to form closed contours in each slice with the guidance of a neuroradiologist. The contours of the left lateral ventricles were reflected to give the same pose as those of the right, resulting in an evaluation set of 138 ventricles for this study.

For each subject, brain size parameters were obtained as follows. Skull stripping was performed on each MR image [12], and ellipsoids were fitted to the resulting brains. The lengths of the three principal axes of the ellipsoids were stored as the brain size parameters. The ventricular surfaces were aligned to a canonical coordinate system using their centroids and the three principal axes obtained from the distribution of the coordinates of their surface points. The brain size parameters were then used to scale each object centred ventricle independently in the three orthogonal directions for normalisation for brain size with respect to the brain size of an arbitrarily chosen template brain. This was necessary to remove the influence of brain shape on ventricular shape.

3.2 Point Distribution Models

A PDM [3] reparameterises a shape described by surface landmark points to a smaller set of shape parameters using equation 1

$$\mathbf{x} = \overline{\mathbf{x}} + \mathbf{P}\mathbf{b} \; . \tag{1}$$

$\mathbf{x}$ is the vector of the coordinates of surface landmarks of a particular shape, $\overline{\mathbf{x}}$ is the average of these vectors over a training set. $\mathbf{P}$ is the matrix whose columns are the eigenvectors corresponding to the largest k eigenvalues of the covariance matrix of the shape vectors. $\mathbf{b}$ is a vector of weights of dimension k. Due to correlations in point positions, k can be much smaller than the number of landmark points. $\mathbf{b}$ then becomes a vector of k shape parameters which are equivalent to $\mathbf{x}$ as a description of the shape.

It is necessary to locate *corresponding* landmark points on all the surfaces in the training set. In the case of 2D PDMs this can be achieved by manual annotation. However, in 3D this becomes difficult and prohibitively labour-intensive.

Davies et al. [4] have shown that the specificity of a SSM depends critically on finding accurately corresponding landmark points. Several approaches have been made towards automatic landmark generation in 3D, including the use of spherical harmonic parameterisation [5] and optimisation of the shape models [4]. Here we identify landmarks from the set of "crest points" on the ventricle surface using a modification of the method due to Subsol et. al. [13]. Correspondence is established using non-rigid registration of the surfaces and minimisation of Euclidean distance expressed as a transportation cost.

3.3 Crest Lines on the Lateral Ventricle Surface

Crest points, which are curvature extrema on the ventricles, are used as anatomical landmarks here. According to the definition of [13] they are points where lines of principal maximal curvature on a surface have maximum values. Crest lines are the locus of crest points and impose an ordering on crest points, which is useful when using crest points to establish correspondence.

To extract the crest points of a ventricular surface, we use the marching lines algorithm [14]. This obtains crest lines directly from the segmented voxel images of the ventricles using the zero crossings of an extremality function of the principal maximal curvature. "Noisy" crest lines and crest points are removed by applying conservative smoothing during their extraction and thresholding using the curvature values at crest points. This results in a consistent set of crest lines across the training set.

3.4 Matching Crest Points as an Assignment Problem

To define correspondence across the training set, crest points and crest lines of the ventricles are matched between ventricles in a pair-wise manner. An ICP-based method for doing this is described in [13]. However, we use a method based on minimising "transportation" costs [1]. In the matching process "injectivity" and "monotonicity" have to be preserved. Injectivity refers to the requirement that in matching the crest points of two ventricles A and B, we create a one-to-one mapping between the crest points. The monotonicity constraint prevents crossovers in the mapping. Furthermore it is desirable to have symmetric matches, in that the matching of surfaces $A \rightarrow B$ and $B \rightarrow A$ give the same pairings.

The ICP-based method in [13] requires steps to impose injectivity and monotonicity, and in general the matches are not symmetric. The transportation method intrinsically enforces injectivity but not monotonicity. Furthermore, If the number of crest points on both examples are the same the matches are symmetric.

In general, the solution to the transportation problem is a global minimum of the transportation cost

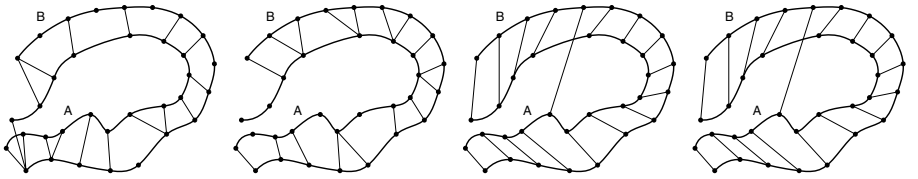
$$z = \sum_{i}^{m} \sum_{j}^{n} D_{ij} x_{ij} \ . \tag{2}$$

where D_{ij} is a cost of transportation of one unit commodity from source i to destination j, and x_{ij} is the quantity transported, subject to the constraint that the sum of commodities generated at all (m) sources is equal to the sum consumed at all (n) destinations

$$\sum_{i=1}^{m} a_i = \sum_{j=1}^{n} b_j \ . \tag{3}$$

where a_i is the output of source i and b_j is the requirement at destination j.

In the present case D_{ij} is the Euclidean distance between point i on one surface and point j on the other (following registration). All a_i and b_j have unit value (each point can match to exactly one other point). In this case the problem reduces to an *assignment problem*. Here we make use of an efficient solution by Achatz et. al. [8]. A global minimum in z is guaranteed. Minimising the assignment cost results in matches that are more numerous and more evenly distributed than those that result from the ICP-based method. Figure 1 illustrates the application of both methods to a pair of synthetic lines.



(a) ICP-based $A \rightarrow B$ (b) ICP-based $B \rightarrow A$ (c) Transportation $A \rightarrow B$ (d) Transportation $B \rightarrow A$

Fig. 1. To illustrate the difference between the ICP-based and the transportation-based methods, both were applied to matching points on a pair of synthetic lines. The above shows initial results (before enforcement of monotonicity and injectivity constraints) for the ICP-based method in both directions (a and b). The initial results of the transportation-based method are always symmetric (c and d) when the number of points in A and B are the same, whereas those of the closest point method are not generally symmetric

3.5 Construction of the 3D PDM of the Lateral Ventricle

One ventricle $\mathbf{v}_t$ was used as a template and its surface represented by vertices and vertex faces defined by triangular triplets of the vertices. The initial triangulation produced about 10,000 vertices, but for computational reasons these were decimated to give about 1,000 vertices. Crest lines were obtained for each

ventricle and normalised with respect to the template as described in section 3.1.

The crest lines of each of the remaining 137 ventricles $\mathbf{v}_i \in \{\mathbf{v}_1, \ldots, \mathbf{v}_{137}\}$ were matched in a pairwise manner to those of the chosen template, $\mathbf{v}_t$. The matches were in both directions i.e. $\mathbf{v}_t \rightarrow \mathbf{v}_i$ and $\mathbf{v}_i \rightarrow \mathbf{v}_t$, using the transportation method and a post-processing step to enforce monotonicity. Matching was performed over 30 iterations: ten iterations each of rigid alignment, affine alignment, and spline warping successively as described in [13].

Although the transportation-based method gives symmetric results for matches in both directions when the number of crest points are equal, the results are not guaranteed to be symmetric when the number of crest points are not equal, which in general is the case with matching ventricles. Therefore, from each matched pair ($\mathbf{v}_t \rightarrow \mathbf{v}_i$ and $\mathbf{v}_i \rightarrow \mathbf{v}_t$), a subset of matches occurring on parts of crest lines that were symmetrically matched in both directions were extracted. Although this decreases the number of matched points used in the subsequent transformation, it gives greater confidence that they are valid matches. For the present case, 1,586 $\pm$ 167 crest points (79% of the total number matched) were on symmetrically matched crest lines for the transportation-based method, and 964 $\pm$ 160 (70% of the total number matched) for the ICP-based method. The symmetric subset of matched points are used to obtain coefficients defining a final spline based warp allowing transformation of the vertex points of $\mathbf{v}_t$ onto the surface of each $\mathbf{v}_i$. The spline based warps are defined in [13].

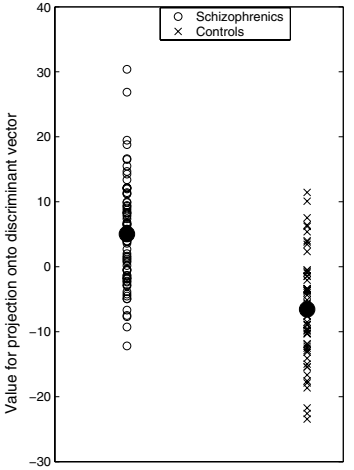
3.6 Shape Analysis

The parameters of the $\mathbf{b}$ vectors are used to define a shape space using the first k eigenvalues in the PDM (k=30 in the present case, explaining over 99% of the observed variance). Each member of the training set is a point within this k-dimensional space, represented by a vector $\mathbf{b}_k$. To characterise shape differences between the groups we conducted linear discriminant analysis (LDA) using Fisher's criterion (see e.g. [6]). This provides a "discriminant vector" in shape space along which the difference between the groups is most marked. We can quantify the shape differences by projecting the individual shape vectors onto the discriminant vector to provide a scalar value representing the individual shapes. The nature of the shape differences between the groups can be visualised by reconstructing the shapes corresponding to the group means. Specific differences correspond to locations where large movements occur between the reconstructed shapes.

4 Results

Figure 2(a) shows the results of projection onto the discriminant vector. The difference in the means was statistically significant ($p < 10^{-13}$ by a Student's t test). Figure 2(b) shows the difference between the means of the schizophrenic group and that of the control group colour-mapped onto a ventricular surface.

The greatest differences were in the region of the tip of the anterior horn (8mm), in the region of the temporal horn (between 2mm and 6mm), around the central part of the main body of the ventricle in the region of the corpus callosum (between 4mm and 6mm).



(a) Projections of the points in 30-dimensional shape space onto the discriminant vector (group means in filled black)

(b) Colour mapped ventricle showing the areas of differences between the schizophrenic group and the control group

Fig. 2. Results of shape analysis

5 Discussion

The results of the morphometric analysis are similar to those of [11] in that they show differences localised to the temporal horn in the region of the parahippocampal fissure, and in the anterior part of the lateral ventricle near the frontal lobe. However, we also found differences in the central part of the lateral ventricle in the region of the corpus callosum. Although [2] also report differences in the temporal horn of male schizophrenics, they did not find differences in the pooled groups of male and females as we have reported here.

Schizophrenia is a complex disease and, as the results of the linear discriminant analysis shows, there is a considerable overlap in the ventricles of schizophrenics and normals. Hence we do not propose we have a method that allows the discrimination of lateral ventricles into schizophrenic and none

schizophrenic groups. However, studies of this sort may help in understanding and monitoring schizophrenia. In this study we have combined left and right ventricles of both males and females. We have also removed all overall volume effects by isotropic scaling of the ventricles prior to shape modelling. The differences we observe are residual differences in shape in addition to any volumetric differences. Future work will include investigating age and gender effects as well as comparing left and right asymmetry.

Acknowledgements. We would like to thank the Epidaure group of INRIA, France for Marching Lines code, Professor P. Klienschmidt of Passau University for code for the assignment problem, and Professors Alan Jackson and Bill Deakin of Manchester University for assistance in interpreting the results.

References

1. Babalola, K. O., Graham, J., Kopala, L., Vandorpe, R.,: Using the Transportation Algorithm to Improve Point Correspondences in the Construction of 3D PDMs, In: Proc. 6th Annual Conference on Medical Image Understanding and Analysis, Portsmouth, UK, pp.141–144, (2002)
2. Buckley, P. F., Dean, D., Bookstein, F.L., Friedman, L., Kwon, D., Lewin, J.S., Kamath, J., Lys, C.,: 3D Magnetic Resonance-Based Morphometrics and Ventricular Dysmorphology in Schizophrenia, Biol. Psychiatry, Vol. 45, pp.62–67, (1999)
3. Cootes, T.F., Hill, A., Taylor, C.J., Haslam, J.,: The Use of Active Shape Models for Locating Structures in Medical Images, Image and Vision Computing Vol. 12, No. 6, pp.355–366, (1994)
4. Davies, R.H., Twining, C.J., Cootes, T.F., Waterton, J.C., Taylor, C.J.,: 3D Statistical Shape Models Using Direct Optimisation of Description Length In: Proc. 7th European Conference on Computer Vision, LCNS Vol.2350(3), pp.3–20 (2002)
5. Gerig, G., Styner, M., Jones, D., Weinberger, D., Lieberman, J.,: Shape Analysis of Brain Ventricles Using SPHARM, In: Proc. IEEE Workshop on Mathematical Methods in Biomedical Image Analysis, pp.171–178 (2001)
6. Hand, D.,J.,: Discrimination and Classification (Ch.4). John Wiley & Sons, (1981)
7. Johnstone, E.C., Crow, T.J., Frith, C.D., Husband, J., Kreel, L., : Cerebral Ventricular Size and Cognitive Impairment in Chronic Schizophrenia, Lancet, Vol. 7992, No. 2, pp.924–926, (1976)
8. Achatz, H., Kleinschmidt, P., Paparizos, K., : A dual forest algorithm for the assignment problem, In: Gritzmann, P., Sturmfels, B., (eds.): DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Vol. 4, pp.1–11, (1991)
9. Monga, O., Deriche, R., Malandain, G., Cocquerez, J.-P., Recursive filtering and edge tracking: two primary tools for 3D edge detection, Image and Vision Computing, Vol. 9, No. 4, pp 203–214, (1991)
10. McCarley, R.W., Wible, C.G., Frumin, M., Hirayasu, Y., Levitt, J.J., Fischer, I.A., Shenton, M.E. : MRI Anatomy of Schizophrenia, Biol. Psychiatry, Vol. 45, pp.1099–1119, (1999)
11. Narr, K.L., Thompson, P.M., Sharma, T., Moussai, J., Blanton, R., Anvar, B., Edris, A., Krupp, R., Rayman, J., Khaledy, M., Toga, A.W.,: Three-Dimensional Mapping of Tempro-Limbic Regions and the Lateral Ventricles in Schizophrenia: Gender Effects, Biol. Psychiatry, Vol. 50, pp.84–97, (2001)

12. Smith, S., Fast robust automated brain extraction, Human Brain Mapping, Vol.17, No.3, pp.143–155,2002
13. Subsol, G., Thirion, J.P., Ayache, N.,: A Scheme for Automatically Building Three-Dimensional Morphometric Anatomical Atlases: Application to a Skull Atlas, Medical Image Analysis, Vol. 2, No. 1, pp.37–60, (1998)
14. Thirion, J.P., Gourdon, A.,: The 3D Marching Lines Algorithm and its Application to Crest Lines Extraction, Graphical Models and Image Processing, Vol. 86, No. 6, pp.503–509, (1996)
15. Vokurka, E.A., Thacker, N.A., Jackson, A., A Fast Model-Independent Method for Automatic Correction of Intensity Nonuniformity in MRI Data, JMRI, Vol.10, No.4, pp.550–562, (1999)

31. **Lateral asymmetry in the shape of brain ventricles in control and schizophrenia groups.** J. Graham, K.O. Babalola, W. Honer, L. Kopala, D. Lang and R. Vandorpe, *Proceedings of the IEEE International Symposium on Biomedical Imaging Arlington VA. April 2006. J Kovacevic and E. Meijering, eds. IEEE. pp 414-417.* doi: 10.1109/ISBI.2006.1624941

LATERAL ASYMMETRY IN THE SHAPE OF BRAIN VENTRICLES IN CONTROL AND SCHIZOPHRENIA GROUPS

J. Graham[1], K.O. Babalola[1], W.G. Honer[2], D. Lang[2], L. Kopala[3], R. Vandorpe[3]

University of Manchester[1], United Kingdom, University of British Columbia[2], Dalhousie University[3], Canada

ABSTRACT

We use point distribution models (PDMs) to investigate lateral asymmetries in the shape of brain ventricles between control subjects and people with schizophrenia. Ventricle surfaces were extracted from T2-weighted MR images and PDMs generated using structural correspondences on the individual surfaces. Using paired linear discriminant analysis we calculate the vector in shape space that maximally separates the shapes of right and left ventricles in the group. The magnitude of the asymmetry is quantified by projection of the individual ventricle shapes onto this vector. We observe significant differences in the magnitude of the asymmetry in both schizophrenia and control groups. There is also a clear difference in the pattern of asymmetry. Male and female subgroups show different magnitudes and patterns of asymmetry, in both groups.

1. INTRODUCTION

Since Johnstone et al.[1] reported volume changes in the brain ventricles associated with schizophrenia, there has been interest in measuring localized shape differences in the ventricles and other neuroanatomical structures between control and schizophrenia groups [2-4]. We have previously published a study in which ventricle shapes were characterised using a Point Distribution Model (PDM) [5]. In that study significant ($p<10^{-13}$) shape changes were observed between schizophrenia and control groups in the temporal and anterior horns in regions close to structures implicated in schizophrenia and in the ventricle body near the corpus calosum. Here we present further analysis using the shape features derived from the PDM to investigate asymmetries between the right and left ventricles in control and schizophrenia groups. Crow [6] hypothesised that loss of asymmetry in schizophrenia accounts for some of the symptoms of the disease, though subsequent evidence has both supported and contradicted this hypothesis. We find significant differences between the shapes of right and left ventricles, and that this asymmetry is in turn significantly different between control and schizophrenia groups.

2. MATERIALS AND METHODS

2.1. Images

This study used volumetric T2 MR scans of 30 controls (14-45 years, 13 female, 17 male) and 39 age and sex matched people with schizophrenia (14-45 years, 9 female, 30 male). The scans were independently acquired in the sagittal, coronal and axial orientations. Each slice had 256 x 256 voxels, with in-plane size of 0.86mm by 0.86mm for sagittal and axial orientations, and 0.78mm by 0.78mm for the coronal orientation. For all orientations the slice thickness was 5mm and the intra-slice gap was 1mm. MR inhomogeneity correction was applied using the method of [7], and the three views of each subject were combined by rigid registration and interpolation to give 3D images with effective resolution of 0.78mm x 0.78mm x 0.78mm. The lateral ventricles were segmented using a 3D edge detector [8] to give edge segments which were manually linked to form closed contours in each slice with the guidance of a neuroradiologist. The contours of the left lateral ventricles were reflected to give the same pose as those of the right, resulting in an evaluation set of 138ventricle surfaces for model building (next section).

Ventricle surfaces were normalized for brain size and shape to minimize the effect of these on measured shape parameters. This was achieved by fitting an ellipsoid to the brain after skull-stripping [9]. The principal axes of the ellipsoid were used to scale each of the dimensions of the ventricles independently to an arbitrarily chosen example from the training set after alignment of their centroids and principal axes.

2.2 Shape Modeling

Point Distribution Models (PDMs), first described by Cootes et al. [10] have been used in a wide range of studies of image segmentation and shape characterization. Shape is parameterised using equation 1.

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}\mathbf{b} \quad (1)$$

$\mathbf{x}$ is the vector of the coordinates of surface landmarks of a particular shape, $\bar{\mathbf{x}}$ is the average of these vectors over a training set. $\mathbf{P}$ is the matrix whose columns are the

eigenvectors corresponding to the largest k eigenvalues of the covariance matrix of the shape vectors. $\mathbf{b}$ is a vector of weights of dimension k. Due to correlations in point positions, k can be much smaller than the number of landmark points. $\mathbf{b}$ then becomes a vector of k shape parameters which are equivalent to $\mathbf{x}$ as a description of the shape. A key point is that the vectors $\mathbf{x}$ for each training shape consist of *corresponding* landmark points. That is, each landmark represents the same anatomical location in each training example. Details of the method of determining corresponding surface points are given in [5]. Briefly, we identify "crest points", loci where the lines of maximal curvature (crest lines) on the ventricle surface have locally maximum values. To obtain correspondences between the crest points on a pair of surfaces an iterative algorithm, combining bi-partite graph-matching [11] and non-rigid surface registration, was used.

The corresponding points derived from this method provide the coordinate vectors $\mathbf{x}$ in equation 1. The eigenvectors $\mathbf{P}$ are used to define a shape space using the first k eigenvalues in the PDM (k=30 in the present case, explaining over 99% of the observed variance). Each $\mathbf{b}$ vector is a k-dimensional feature vector describing the shape of a particular ventricle.

2.3. Paired Discriminant Analysis

In our earlier study [5] we used linear discriminant analysis (LDA) to characterise shape differences between groups. This can be expressed as equation 2.

$$\hat{\mathbf{w}} = \alpha \mathbf{S}^{-1}(\bar{\mathbf{b}}_c - \bar{\mathbf{b}}_s) \quad (2)$$

where $\hat{\mathbf{w}}$ is the direction of the maximally discriminating vector in shape space, $\bar{\mathbf{b}}_c$ and $\bar{\mathbf{b}}_s$ are the mean shape vectors of the two groups (control and schizophrenia) and $\mathbf{S}$ is derived from the covariances of the shape parameters.

In this case we are interested in the difference between *pairs* of ventricles belonging to the same individual. We modify the normal LDA analysis as follows to produce a *paired linear discriminant analysis*. We seek a vector in shape space that maximally separates members of a pair.

$$\hat{\mathbf{w}}_p = \alpha \mathbf{S}_p^{-1} \bar{\mathbf{d}} \quad (3)$$

where $\hat{\mathbf{w}}_p$ is the direction of the vector that maximally separates members of pairs, $\bar{\mathbf{d}}$ is a vector expressing the difference between the shapes of a right-left pair and $\mathbf{S}_p$ is the covariance matrix derived from the difference vectors. Figure 1 illustrates the calculation of $\hat{\mathbf{w}}_p$ for a number of (fictitious) right-left pairs

We can quantify the shape differences by projecting the shape vectors onto the discriminant vector to provide a scalar value for the individual shapes. The group means (right versus left) allow us to quantitate the degree of asymmetry.

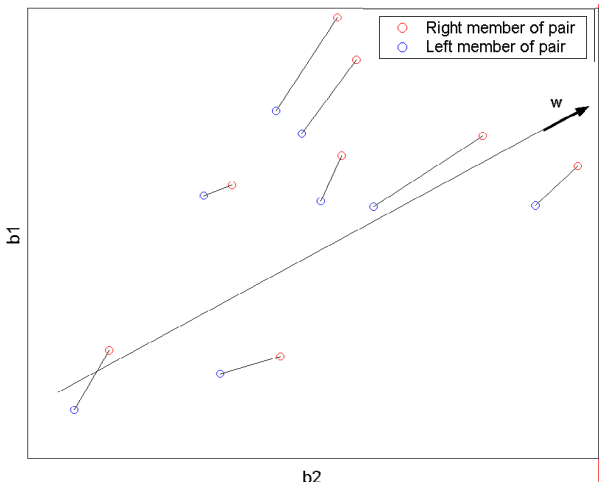


Figure 1. Illustration of the vector maximally separating right-left pairs in a two-dimensional shape space.

3. RESULTS

Shape asymmetries were measured for both groups for the complete data set and for the male and female subgroups separately. Figure 2 shows the paired values for the shape vectors projected onto the maximally discriminating vector for each of the sets. The results are shown quantitatively in table 1, which shows the mean right-left differences for each group, and the p-values derived from the pair-wise t-test comparing the within-group differences in each case. Taken as a whole, the magnitude of the shape difference between ventricle pairs is generally consistent amongst individuals in each group, the magnitude of the difference being greater in the control group than the schizophrenia group. The opposite is true of the male subgroups: asymmetries in shape, as measured by the projected shape differences, are smaller in the control group. For the female subgroup there is no significant difference.

Figures 3-5 show the shape differences in each of the groups colour-mapped onto the surface of the average ventricle, showing where the maximum differences are observed. These figures show that the pattern of the asymmetry between right and left hemispheres is different for control and schizophrenia groups. For example, for the whole set the asymmetry in the schizophrenia group is greatest in the lower part of the main ventricle body and in the anterior horn. For control subjects the asymmetry was greatest at the tips of the temporal and anterior horns.

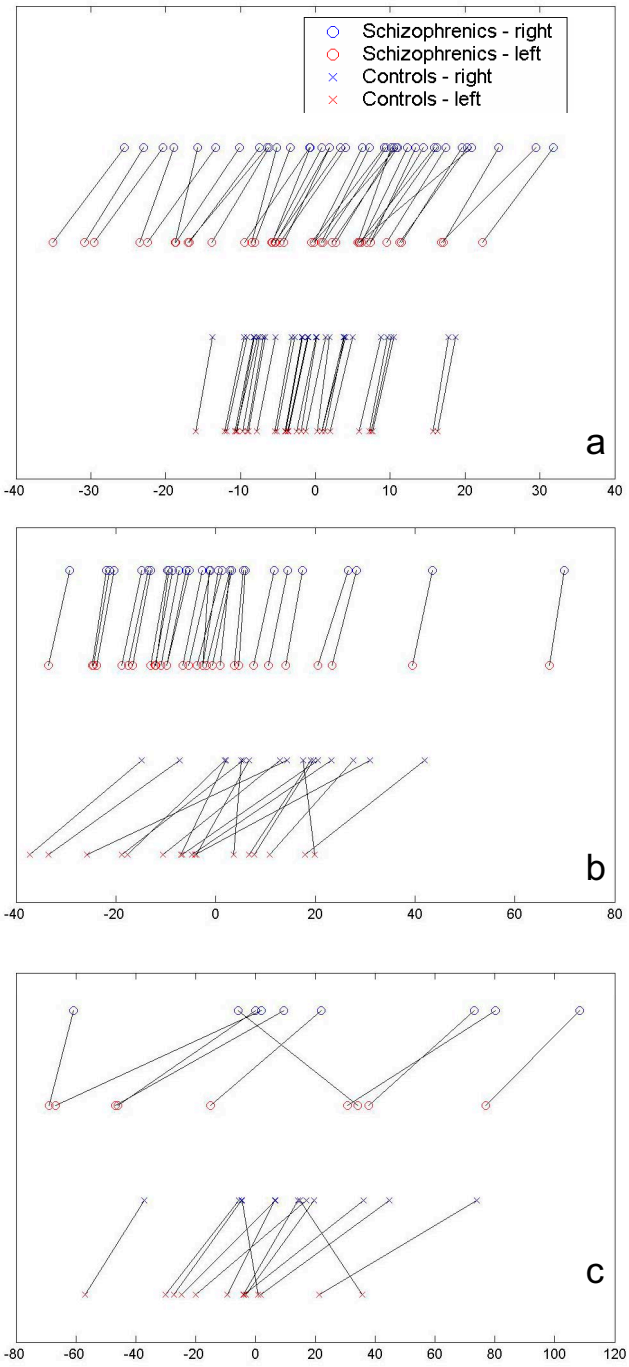


Figure 2. The differences between right and left ventricles in control and schizophrenia groups projected onto the maximally discriminating vector in each case. The scalar values for individual pairs of ventricles are shown connected (the vertical separation is for clarity). (a) All subjects (b) Male subgroup (c) Female subgroup

	$\bar{D}_{schiz}$	$\bar{D}_{cont}$	p
All subjects	7.94 (2.38)	2.49 (0.33)	4.5×10^{-19}
Males	3.59 (1.07)	19.42 (11.23)	8.7×10^{-10}
Females	23.15 (19.65)	32.45 (32.03)	0.43

Table 1. Means and variances (in brackets) of the scalar projections of the three groups shown in figure 2. p is derived from the t-test comparing D_{schiz} and D_{cont}.

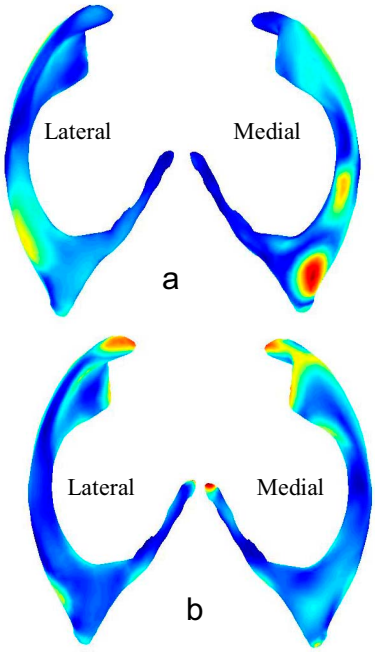


Figure 3. The pattern of mean differences between right and left ventricles in the complete data set colour-mapped onto the mean ventricle shape. Smallest differences are blue, largest are red. (a) schizophrenia group (b) control group, medial and lateral aspects as indicated

4. DISCUSSION

We have used point distribution models to provide shape features that can be used in comparing asymmetries in shape between right and left ventricles of the brain. We observe consistent asymmetries between individuals within the control and schizophrenia groups, but the groups differ in the pattern of the asymmetry observed. That is, the shape vector that maximally distinguishes the ventricle shapes between right and left is different for the control and schizophrenia groups. The same is true when we examine the male and female subgroups. The control group shows greater asymmetry than the schizophrenia group in the complete set and the female subgroup; the reverse is true for the male subgroup (table 1). The observed difference between control and schizophrenia groups is highly significant, except in the female subgroup, where the number of subjects with schizophrenia is very small.

We have demonstrated that this is a potentially useful

technical approach to measurement of lateral shape differences. However, the measured differences both quantitative (table 1) and qualitative (figs 3-5) on this dataset are not consistent between the whole group and the subgroups. This may indicate that there are several vectors in parameter space that would be similarly discriminating between control and schizophrenia groups, the selection between them being stochastic. A more stable selection might result from the use of a larger dataset. Alternatively exploration of vectors that discriminate significantly, if not maximally, may also reveal important lateral changes.

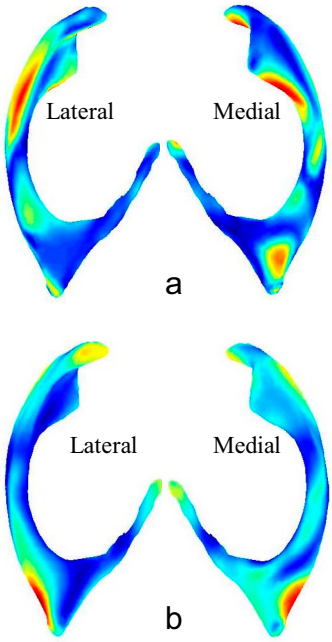


Figure 4. The pattern of mean differences between right and left ventricles in the male subgroup colour-mapped onto the mean ventricle shape. Smallest differences are blue, largest are red. (a) schizophrenia group (b) control group, medial and lateral aspects as indicated

5. REFERENCES

[1] E. C. Johnstone, C. D. Frith, T. J. Crow, et al., "Cerebral ventricular size and cognitive impairment in chronic-schizophrenia", Lancet, 7992, pp. 924-926, 1976.

[2] P. F. Buckley, D. Dean, F. L. Bookstein, et al., "Three-dimensional magnetic resonance-based morphometrics and ventricular dysmorphology in schizophrenia", Biological Psychiatry, 45, pp. 62-67, 1999.

[3] G. Gerig, M. Styner, D. Jones, et al., "Shape analysis of brain ventricles using SPHARM," in IEEE workshop on mathematical methods in biomedical image analysis, pp. 171-178, 2001.

[4] K. L. Narr, P. M. Thompson, T. Sharma, et al., "Three-dimensional mapping of temporo-limbic regions and the lateral ventricles in schizophrenia: Gender effects", Biological Psychiatry, 50, pp. 84-97, 2001.

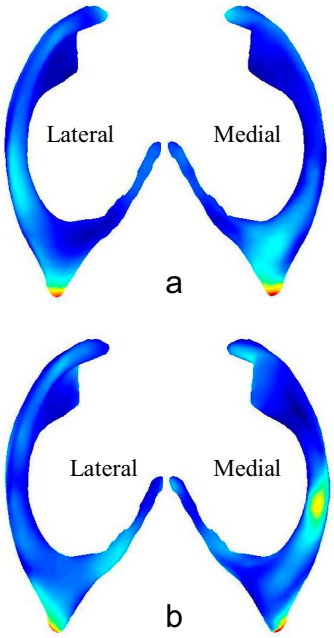


Figure 5. The pattern of mean differences between right and left ventricles in the female subgroup colour-mapped onto the mean ventricle shape. Smallest differences are blue, largest are red. (a) schizophrenia group (b) control group, medial and lateral aspects as indicated

[5] K. O. Babalola, J. Graham, W. Honer, et al., "An investigation of morphometric changes in the lateral ventricles of schizophrenic subjects," in Medical Image Computing and Computer-Assisted Intervention - MICCAI 2003, LNCS vol 2879, (R.E. Ellis and T.M. Peters eds.), pp. 521-529, 2003.

[6] T. J. Crow, "Schizophrenia as failure of hemispheric dominance for language", Trends in Neurosciences, 20, pp. 339-343, 1997.

[7] E. A. Vokurka, N. Thacker and A. Jackson, "A fast model independent method for automatic correction of intensity nonuniformity in MRI data", Journal of Magnetic Resonance Imaging, 10, pp. 550-562, 1999.

[8] O. Monga, R. Deriche, G. Malandain and J. P. Cocquerez, "Recursive filtering and edge tracking - 2. Primary tools for 3D edge-detection", Image and Vision Computing, 9, pp. 203-214, 1991.

[9] S. M. Smith, "Fast robust automated brain extraction", Human Brain Mapping, 17, pp. 143-155, 2002.

[10] T. F. Cootes, D.H. Cooper, C. J. Taylor and J. Graham, "Active shape models - their training and application", Computer Vision and Image Understanding, 61, pp. 38-59, 1995.

[11] H. Achatz, P. Kleinschmidt and K. Paparizos, "A dual forest algorithm for the assignment problem," in DIMACS series in discrete mathematics and theoretical computer science, vol 4, (P. Gritzmann and B. Sturmfels, ed), pp. 1-11, 1991.

Applications of Image Analysis: Proteomics

32. **Statistical models of shape for the analysis of protein spots in 2-D electrophoresis gel images.** M.D. Rogers, J. Graham and R.P. Tonge *Proteomics 3: 879-896, 2003.* doi: 10.1002/pmic.200300421

Mike Rogers[1]
Jim Graham[1]
Robert P. Tonge[2]

[1]Imaging Science and Biomedical Engineering, University of Manchester, Manchester, UK
[2]Protein Science, Enabling Science and Technology (Biology), AstraZeneca, Macclesfield, Cheshire, UK

Statistical models of shape for the analysis of protein spots in two-dimensional electrophoresis gel images

In image analysis of two-dimensional electrophoresis gels, individual spots need to be identified and quantified. Two classes of algorithms are commonly applied to this task. Parametric methods rely on a model, making strong assumptions about spot appearance, but are often insufficiently flexible to adequately represent all spots that may be present in a gel. Nonparametric methods make no assumptions about spot appearance and consequently impose few constraints on spot detection, allowing more flexibility but reducing robustness when image data is complex. We describe a parametric representation of spot shape that is both general enough to represent unusual spots, and specific enough to introduce constraints on the interpretation of complex images. Our method uses a model of shape based on the statistics of an annotated training set. The model allows new spot shapes, belonging to the same statistical distribution as the training set, to be generated. To represent spot appearance we use the statistically derived shape convolved with a Gaussian kernel, simulating the diffusion process in spot formation. We show that the statistical model of spot appearance and shape is able to fit to image data more closely than the commonly used spot parameterizations based solely on Gaussian and diffusion models. We show that improvements in model fitting are gained without degrading the specificity of the representation.

Keywords: Spot modelling / Statistical models / Two-dimensional gel image analysis PRO 0421

1 Introduction

Image analysis of 2-DE gels requires the identification of a large number of individual spots (possibly in excess of 3000 *per* gel). These must be characterized for further analysis of the sample, such as comparison across a set of gels. Currently, many commercial and academic 2-DE image analysis packages are available [1–7], each with an associated spot identification and characterization algorithm. One of the first steps in any spot detection algorithm is the segmentation of individual spots from the background. After the segmentation step, spots are characterized and represented as a list of parameters over which further analysis can be carried out. Spot characterization algorithms can be divided into two categories: parametric and nonparametric. Nonparametric methods [6, 8–11] carry out various heuristic post-processing routines on the raw segmentation boundaries to delineate the spots. Spots are then represented by a set of measurements calculated over the detected spot regions. No explicit constraints on the shape of the boundaries or the appearance of the spot are imposed. The flexibility of this approach is outweighed by the relative lack of robustness when complex images are analyzed. Parametric methods use models to parameterize protein spots. The use of models to aid the interpretation of complex data is a well established technique in image interpretation [12–15]. Models represent prior knowledge which is used to impose constraints on the analysis procedure, which in turn improves the robustness of the solution.

Commonly, nonparametric methods are used to provide an initial spot segmentation, which can then be refined using model based, parametric quantification. In 2-DE gel analysis, the most commonly used spot model is a Gaussian function [7, 16]. Figure 1(a) shows an example of a typical protein spot with a Gaussian profile. This model is assumed to provide a good representation of most spots present in most gel images. However, it has been shown that Gaussian models produce an inadequate fit to some protein spots, most notably large volume, saturated spots [17]. The Gaussian formulation is insufficiently flexible to represent the true variation in spot appearance. Figure 1(b) shows an example of a high volume protein spot exhibiting a saturated, 'flat-top' shape. Bettens [17, 18] addressed this shortcoming by proposing a model

Correspondence: Dr. Mike Rogers, Imaging Science and Biomedical Engineering, Stopford Building, University of Manchester, Oxford Road, Manchester, M13 9PT, UK
E-mail: mike.rogers@man.ac.uk
Fax: +44-161-275-5145

Abbreviations: PDM, point distribution model; **PCA**, principal components analysis

© 2003 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim 0173-0835/03/0606–887 $17.50+.50/0

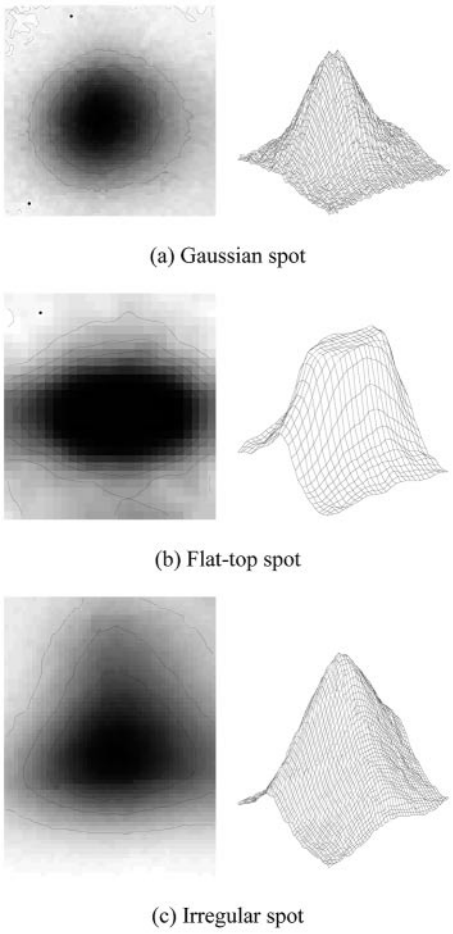


Figure 1. Examples of electrophoresis gel spots. The first column shows the appearance of the spot in the image with contours of constant grey-level overlayed. The second column shows a 3-D mesh representation of the same data. (a) Gaussian, (b) 'Flat-top', (c) Irregular.

based on the physics of the spot formation. Protein spots are formed by a diffusion process, which is only adequately represented by a Gaussian when the initial concentration distribution occupied by the sample has a small area. Bettens' diffusion model more adequately represents spots in the gel when this assumption is not met.

Both the Gaussian and diffusion models assume perfect diffusion across the gel medium. Spots created by a perfect diffusion process will be regular and symmetric. In practice, the diffusion process is not perfect and spots can be formed with unpredictable, unusual shapes. An example of such a spot is shown in Fig. 1(c). Further examples are shown in Fig. 7(a). To represent more adequately the full range of observed spot shape, we have developed a new protein spot model that is both flexible enough to represent irregular shape variation and specific enough to retain usable constraints on the interpretation of gel images. The physical process by which irregular spots are formed is extremely complex. It would be a daunting task to directly estimate all the physical variables affecting spot formation. Therefore, our model is built using a statistical analysis of the resulting spot appearance, trained from examples. The model has two main parts, the first representing valid variation in spot shape, and the other pertaining to the diffusion process that forms spot appearance. In the following subsections we briefly describe the Gaussian and diffusion models.

1.1 Gaussian model

The most common protein spot model is based on the 2-D Gaussian function:

$$S(x,y) = B + I\exp\left(-\frac{(x-x_0)}{2\sigma_x^2}\right)\exp\left(-\frac{(y-y_0)}{2\sigma_y^2}\right) \qquad (1)$$

where B is background intensity and I is spot intensity, x_0 and y_0 control spot location and σ_x and σ_y control the spread of the Gaussian independently in x and y directions. Protein spot formation is a diffusion process. Under ideal conditions, the bivariate Gaussian model represents diffusion of an initial concentration focussed at a single point in two independent directions.

1.2 Diffusion model

In practical situations, the assumptions that must hold for a Gaussian model to accurately represent protein spots are often broken. Bettens [17, 18] addressed this with a more detailed theoretical model of anisotropic diffusion. Again, the model assumes perfect diffusion characteristics. In this case the initial concentration is assumed to be uniformly distributed across a circular disk, leading to the formula:

$$S(x,y) = B + \frac{C_0}{2}\left[\mathrm{erf}\left(\frac{a'+r'}{2}\right) + \mathrm{erf}\left(\frac{a'-r'}{2}\right)\right] + \\ + \frac{C_0}{r'\sqrt{\pi}}\left[\exp\left(-\left(\frac{a'+r'}{2}\right)^2\right) - \exp\left(-\left(\frac{a'-r'}{2}\right)^2\right)\right] \qquad (2)$$

with

$$r' = \sqrt{\frac{(x-x_0)^2}{D'_x} + \frac{(y-y_0)^2}{D'_y}}$$

where B is background intensity, C_0 is initial concentration, D'_x and D'_y are related to the diffusion constants in the two main directions of diffusion, x_0 and y_0 control location and a' is the area of the disc containing the protein material.

Note that as $a' \to 0$, Eq. (2) reduces to the bivariate Gaussian. The model becomes more accurate than a Gaussian model when the initial single point assumption is invalid. This is most likely for proteins with relatively high concentrations, which tend to appear in gel images as spots with large volumes exhibiting a 'flattened' profile.

2 Materials and methods

2.1 Statistical model

The diffusion model of spot formation is combined with a model of spot shape defined by a training set. The shape representation is compact, yet sufficiently comprehensive to represent the full range of observed spot shapes.

2.1.1 Modelling shape

Methods of representing shape variation have received much attention in machine vision in the past [15, 19–21]. Cootes *et al.* [12] introduced one of the most widely used techniques called point distribution models (PDMs). A PDM represents the statistics of the observed variation in a training set of shapes, and is constructed in three steps: first parameterizing the shapes by placing landmark points on object boundaries in a training set of images, then aligning the landmarks, and finally analyzing the remaining variation amongst the aligned training data.

2.1.1.1 Landmarking the training set

The landmark points provide a vector which represents the shape of a spot: $\mathbf{x}_i = (x_1, x_2, x_3, \ldots, x_n, y_1, y_2, \ldots, y_n)^T$. Details of how we determine the landmark positions are described in Section 2.2. For the moment, note that we use 25 points to represent the spot shape.

2.1.1.2 Aligning the landmarks

As we wish to model shape variation, it is first necessary to remove other sources of variation. Spot shapes are aligned with respect to their centre of gravity and scale. In the general case of shape modelling the Procrustes alignment method [22] is commonly applied to exclude orientation variation (for details see [12]). In the case of gel spots, we already incorporate orientation into the landmarking process so this step is unnecessary.

2.1.1.3 Modelling the shape variation

The shapes in the training set are represented by a set of aligned shape vectors, x_i with dimensionality $2n$ (in our case 50). The number of degrees of freedom with which the shapes can vary is typically much less than $2n$. This is because the variation in landmark position between examples is usually highly correlated. PDMs use principal component analysis (PCA) to capture these correlations and therefore reduce the number parameters required to represent the shape. The approach is as follows. The $2n \times 2n$ covariance matrix, S, of the data is:

$$\mathbf{S} = \frac{1}{s-1}\sum_{i=1}^{s}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \quad (3)$$

where $\bar{\mathbf{x}}$ is the mean of the vectors. The aligned training data, $\mathbf{x}$, forms a cloud of points around $\bar{\mathbf{x}}$ in a $2n$-D space. The eigenvectors, $\mathbf{p}_j$, and corresponding eigenvalues, λ_j ($j = 1, \ldots, 2n$), of $\mathbf{S}$ represent a set of orthogonal axes which are aligned with the principal modes of variation of the cloud. The eigenvectors corresponding to the largest eigenvalues represent the most significant modes along which the shape can vary. λ_j gives the variance along the jth component. Most of the shape variation can be represented by selecting a smaller number, $n_s < 2n$, of these axes which explain a large proportion of variation. Often n_s is chosen so the selected axes, or modes, explain at least, say, 95% of the variance exhibited in the training set.

Neglecting any alignment steps, any shape, $\mathbf{x}_j$, in the training set can then be approximated by a weighted sum of the first n_s eigenvectors and the mean shape:

$$\mathbf{x}_i \approx \bar{\mathbf{x}} + \mathbf{P}\mathbf{b}_s \quad (4)$$

where $\mathbf{P} = (p_1, p_2, \ldots, p_{n_s})$ is the matrix of the first n_s eigenvectors, and $\mathbf{b}_s$ is a n_s dimensional vector of weights, normally referred to as shape parameters. PDMs are generative models. New examples can be constructed by varying the values of the shape parameters, $\mathbf{b}_s$, in Eq. (4). The shape parameters, $\mathbf{b}_s$, which best match the model to a particular shape vector, $\mathbf{x}_i$, can be calculated as follows:

$$\mathbf{b}_s = \mathbf{P}^T(\mathbf{x}_i - \bar{\mathbf{x}}) \quad (5)$$

2.1.2 Modelling appearance

The PDM only represents shape, but we require a full model of spot appearance. Protein spot formation in 2-DE gels is a diffusion process which is equivalent to convolution of an initial concentration distribution with a 2-D Gaussian kernel. We have assumed the initial concentration distribution can be represented as a flat 2-D shape

within the boundary represented by the shape model. This flat shape is convolved with a bivariate Gaussian kernel giving a final model with the form:

$$S(x,y) = B + F(\bar{\mathbf{x}} + \mathbf{P}\mathbf{b}_s) * G(x,y,\sigma_x,\sigma_y,I) \quad (6)$$

where $*$ is the image convolution operator,

$$G(x,y,\sigma_x,\sigma_y,I) = I \exp\left(-\frac{(x-x_0)}{2\sigma_x^2}\right) \exp\left(-\frac{(y-y_0)}{2\sigma_y^2}\right) \quad (7)$$

and $F(\bar{\mathbf{x}} + \mathbf{P}\mathbf{b}_s)$ is an image with value 1 within the shape and 0 outside. We define our model using the parameter vector $\mathrm{p} = (B,I,x_0,y_0,\sigma_x,\sigma_y,s,\mathbf{b}_s)$, where B is an additive background term, I is spot intensity, x_0 and y_0 control location, σ_x and σ_y control the spread of the Gaussian along the two directions of diffusion, s is a scaling for the spot shape (from the alignment procedure) and $\mathbf{b}_s$ is a vector of shape parameters.

Figure 2 shows an example of the full spot model. A flat spot shape generated using the statistical shape model is convolved with a bivariate Gaussian function resulting in an irregularly shaped 'flat-top' protein spot. This model is equivalent to the bivariate Gaussian when $s \rightarrow 0$, and is equivalent to the diffusion model when the shape parameters, $\mathbf{b}_s$, represent an elliptical shape.

2.2 Building a spot model

Section 2.1 described the basis of the models we use. Here we address the practical issue of building the model: determining the training shapes from spot images and calculating the distributions of parameter values. In many applications of PDMs, manual marking of landmark points has been used. Due to the complexity of the images, and the number of spots required to build a model, this is an impractical strategy in this case. We proceed by segmenting the spots in the training images, smoothing the boundaries obtained using a general shape representation and making the landmark points evenly spaced around the resulting boundary. As the boundaries are extracted from real image data, a number of overlapping spots will be represented. These need to be detected and excluded from the training data, as their inclusion would bias the model and result in reduced specificity.

2.2.1 Automatic generation of training shape vectors

Raw spot boundaries are obtained by thresholding the Laplacian of Gaussian transform of the training gel images (Gaussian σ = 5). The resulting boundaries are smoothed using a Fourier shape descriptor [15] resulting in a parameterization of the spot shape by the Fourier coefficients. The Fourier coefficients represent spot shapes using five harmonics. Spot appearance is modelled by convolving this smoothed shape with a Gaussian kernel (Eq. (6)). The parameters of the joint model are then optimized to improve the fit to the original image data using a Levenberg-Marquardt gradient descent algorithm. This provides an adjusted parameterization of the shape matched to the image data. In this way the shapes used to build our statistical model are derived from our model of spot appearance, rather than the somewhat arbitrary data-driven segmentation.

The PDM landmark representation is obtained from the resulting spot shapes by placing 25 evenly spaced points around the boundary. The position of the first point is the topmost intersection of the boundary with the line $x = \bar{x}_b$, where $\bar{x}_b$ is the x coordinate of the shape's centre of gravity. The number of points and their placement are defined rather arbitrarily. It is, in principle, possible to optimize the landmark positions [23], but this was judged unnecessary for this study, as the spot shapes are consistently smooth and relatively featureless.

2.2.2 Robust model building

Automatic generation of training shapes carries the danger that incorrect shapes may be included in the model. These shapes may be the result of incorrect data-driven segmentation or the inclusion of overlapping spots as

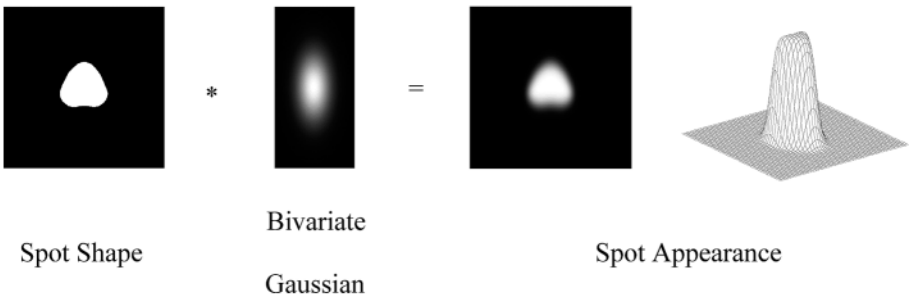


Figure 2. Spot model formation. A flat shape is convolved with a bivariate Gaussian kernel, which is equivalent to a diffusion process.

single objects. The Fourier shape representation imposes no explicit shape constraints, other than smoothness, so it is not possible to filter these incorrect segmentations at that stage. We could filter the resulting shapes by hand, but this would be a highly time consuming and subjective process.

Rather, we have chosen to reduce the influence of such shapes by using robust principal component analysis [24] in the model building. We expect the number of incorrect shapes to be small, and therefore they can only influence the model as outliers in the shape distribution. Robust PCA iteratively reduces the influence of outliers on the resulting model. The effect of the robust PCA can be seen in Fig. 3. The figure shows two PDMs, one built using standard PCA (Fig. 3(a)) and one built using robust PCA (Fig. 3(b)). The models were generated from the same training data. Both models represent the spots by principal components that retain 99% of the observed variance, in the robust case this is 99% of the variance remaining after the iterative weighting procedure.

The standard model represents the retained variance in the training data using ten modes, whereas the robust model requires only six modes. The contribution of each mode to the total variance of the training set is shown for each model. The centre of each row shows the mean model shape; on either side of the mean are shown the change in shape obtained by varying the corresponding shape parameter b_k by $\pm 2\sigma_k$ about the mean. The first mode of the standard model represents a large variation in aspect ratio with an apparent 'waist' becoming visible at the extremes of the mode. This mode would allow the model to represent multiple overlapping spots, which is undesirable. There is no mode in the robust model that allows shapes with 'waists'. The first robust mode corresponds to mode 2 of the standard model, and represents shape 'skew', although with a somewhat less extreme trend than the standard model. Aspect ratio variation is represented by robust mode 2. The rest of the robust modes represent less significant shape variation, including some squaring-off and trends towards triangular spot shape. Figure 4 shows examples of shapes that have been treated as outliers by the robust analysis. They all represent highly uncharacteristic shapes and several are clearly multiple spots.

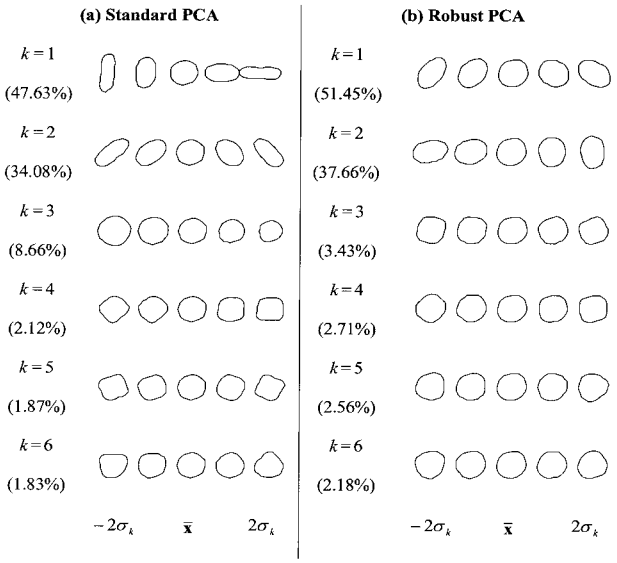


Figure 3. Modes of variation of PDM models. The proportion of total model variance each mode represents is given. Each row is formed by varying the appropriate shape parameter $\pm 2\sigma$ by from the model mean, whilst keeping all other parameters fixed. (a) The first six of 10 modes PDM built using standard PCA, retaining 99% of total training set variance. (b) All six modes of a PDM built using robust PCA. Both models were trained with the same data.

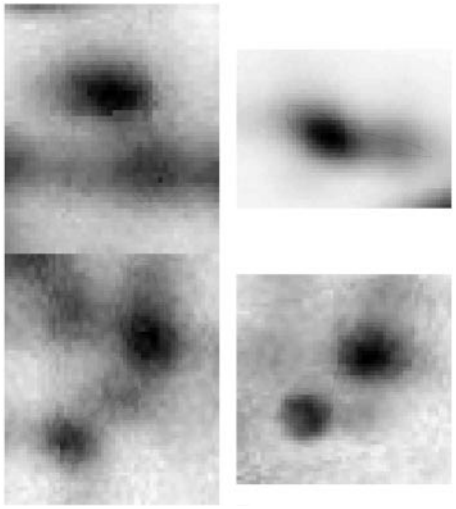
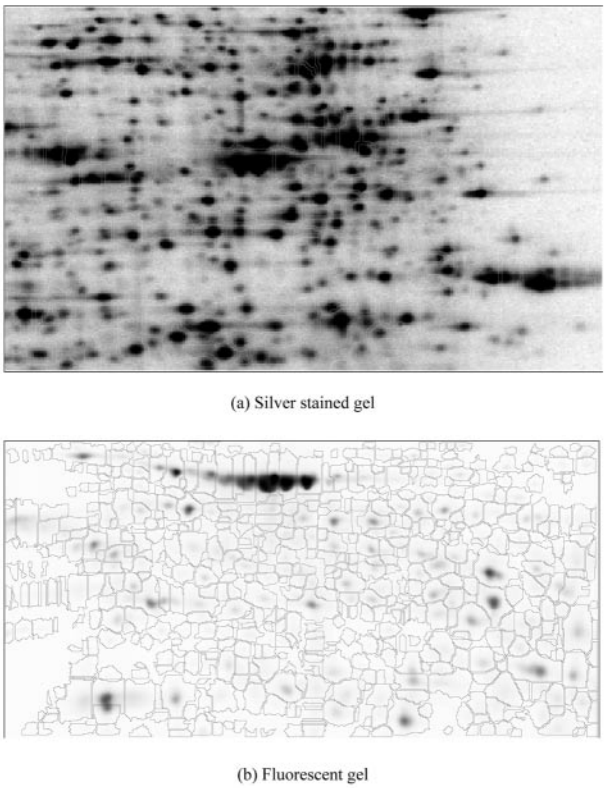


Figure 4. Four examples of shapes that have been downweighted by robust PCA. Each shape is superimposed over the image patch used in its generation.

2.3 Evaluation of models

We have compared the results for fitting the statistical spot model to image data with those achieved using the Gaussian and diffusion models. The experimental procedure was as follows: spot regions were detected in a test image using a watershed algorithm [18] (Fig. 5). Each of the spot models was fitted to each spot region using a Levenberg-Marquardt nonlinear optimization algorithm to determine the best model parameters, minimizing the following residual:



(a) Silver stained gel

(b) Fluorescent gel

Figure 5. Example training images with watershed boundaries. (a) A silver stained image with 403 delineated fitting regions, downloaded from http://www.2dgels.com/benchmark/. (b) A fluorescent dye image with 573 fitting regions.

$$r = \arg\min_{\mathrm{p}} \left[\frac{\sum_{x,y \in R} (\mathrm{S}(x,y|\mathrm{p}) - I(x,y))^2}{n_R \left(I_R^{\max} - I_R^{\min}\right)^2} \right] \quad (8)$$

where R is the region of the image over which fitting takes place, $x,y \in R$ are the coordinates of the pixels within the fitting region, $I(x,y)$are image values, $S(x,y|\mathrm{p})$ are the model values given the parameter vectors p (Section 2.1.2), $I_R^{\max}$, $I_R^{\min}$ are the maximum and minimum image values within the region, and n_R is the number of pixels within the region. This residual provides a measure of model fit error that is normalized with respect to the intensity of the spot (which we have approximated as $I_R^{\max}$, $I_R^{\min}$) and the size of the fitting region (the number of pixels n_R). This residual form allows direct comparisons of fit quality to be made between high and low volume spots. The three models were fitted to 403 watershed delineated spots shown in Fig. 5(a) and 573 spots shown in Fig. 5(b). Fig. 5(a) is a section from a silver stained *E. coli* gel under standard conditions. (The image is available for download at http://www.2dgels.com/benchmark/.) The section is 375×228 pixels in size with a bit depth of 8. Fig. 5(b) is a section from a gel stained with a fluorescent dye. The section is 2896×2485 in size with a bit depth of 24. The silver image contains many saturated and overlapping spots, whereas the fluorescent image contains a much higher proportion of regular Gaussian-like spots with fewer saturated or overlapping spots.

Table 1. Mean residual after model fitting to 403 spots in the silver image and 573 spots in the fluorescent image

Model	Silver $\bar{r}$	Fluorescent $\bar{r}$
Gaussian	8.3×10^{-3}	5.11×10^{-3}
Diffusion	7.83×10^{-3}	4.94×10^{-3}
Statistical	7.49×10^{-3}	3.63×10^{-3}

3 Results and discussion

3.1 Accuracy

The mean residuals $\bar{r}$ for each model after fitting to all regions in both images are shown in Table 1. In general the fitting results for the fluorescent image are better due to the higher resolution of the image data. The statistical model results in the smallest average residual after fitting for both images. Figure 6 shows the mean residual for each spot model and image, grouped by volume. Group 1 contains the smallest 10% of spots by volume, rising to group 10 which contains the largest 10% of spots by volume. For the silver image, each group contains 40 spots (with 43 in group 10); for the fluorescent image, each group contains 57 spots (with 60 in group 10). In our evaluation images, spots correspond to dark regions, so volume was defined as the sum of the inverse pixel intensities over the watershed region:

$$v = \sum_{x,y \in R} (1 - I(x,y)) \quad (9)$$

where spot intensity $I(x,y)$ is in the range (0, 1). Figure 6 shows that, in both cases, the largest improvements in fit made by the statistical model are associated with the largest spot volumes. We have assumed that high volume spots are more likely to produce unusual spot shapes, which, we have argued, are the best represented by the statistical model. For groups 9 and 10, with highest spot volume, statistical models resulted in average decrease in fitting residuals of 33% and 23.9% over the diffusion and Gaussian models respectively for the silver image, with 57.9% and 57.3% improvements for the fluorescent image. For the silver image, small and medium volume spots (groups 1–6) give fits for the Gaussian, diffusion

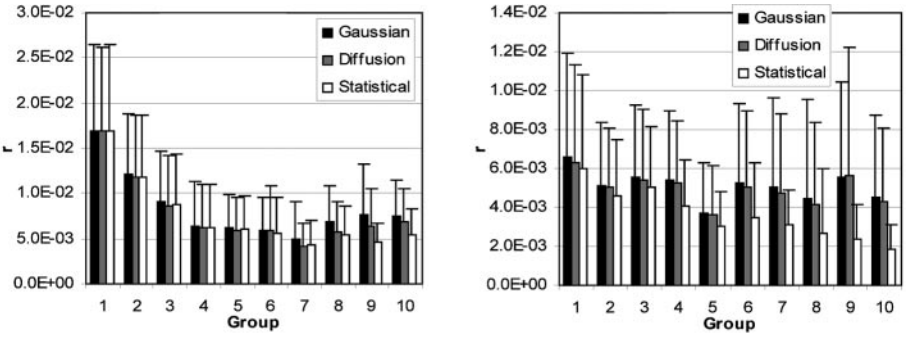


Figure 6. Residual *r* of model fit plotted by increasing spot volume for each model. Spot volume group 1 contains the smallest 10% of spots by volume, rising to group 10 which contains the largest 10% of spots by volume.

and statistical diffusion models that are almost equivalent. However, the statistical model results in reductions in residual for all volume groups of the fluorescent image. This suggests that in the fluorescent image all spot groups contain shape variation away from Gaussian assumptions, even the smallest spots by volume. The statistical model is able to fit to these subtle spot shape changes in the higher resolution fluorescent image. This trend is not visible in the silver image data and this may be because the low resolution of the silver image data introduces local minima into the model fitting search space, which prevents full convergence. Under-converged results will not show a quantitative difference between the models' final residuals for small to medium spots.

For all spot volume groups the statistical model results in fits that are better than or equivalent to the fits of the other two models. This is achieved in both images despite large visual and resolution differences. These results demonstrate that the statistical model is able to fit well to a wide variety of gel image types.

3.2 Specificity

The results show that our statistical model fits more closely to the image data than the other models. This is to be expected, as the model has the most degrees of freedom. An important question is whether the reduction in residual corresponds to a decrease in model specificity. Both images contain watershed fitting regions with multiple spots. A specific model should not represent these regions well. We have argued that the images also contain regions with singular, but irregularly shaped spots. Figure 7 shows five examples of regions containing irregular, single spots and five examples of regions containing multiple spots, together with the fits and residuals of each model. For each of the single spot regions, the lowest residual is achieved with the statistical model. On average, the residual of the statistical model is 65% and 63.9% lower than the fits of the Gaussian and diffusion models respectively. The fits of all models to multispot regions are visually poor (Fig. 7(b)). Here, the decrease in average residual achieved by the statistical model is 16.9% and 1.7% compared to Gaussian and diffusion models respectively. Clearly, the statistical model improves the fit for single spots significantly more than for multiple spot regions. The careful training of the model gives a representation that is specific to single spots, and therefore cannot represent multiple spot regions significantly better than the other models. These selective fitting improvements lead to an increase in the separability of the two types of fitting regions. For the statistical model, four of the five single spot fits have lower residuals than all of the multispot regions. For the Gaussian model, only three single spots have a residual that is lower than all the Gaussian multispot fits. For the diffusion only two single spot fits are lower than all multispot regions. This example shows that in general it is not possible to set a single threshold on model fit residual that will identify all fitting regions containing multiple spots. However, a model that is specific to the observed appearance of gel spots, such as our statistical shape model, will fit more closely to genuine single spots than a more general heuristic model, such as a simple Gaussian function. A specific model also fits poorly to invalid data, increasing the likelihood of detecting invalid model fits.

These results suggest that the statistical model can improve the likelihood of detecting erroneously fitted models. However the figure shows only a small set of example spots. It is necessary to view the full set of residuals to determine whether this trend is genuine. Figure 8 shows reconstructions of the silver image (Fig. 5(a)) using the Gaussian, diffusion and statistical models. The images have been constructed using model values within the fitting regions (as defined by the watersheds in Fig. 5) and have been filled by interpolation in the other areas of the image. The images are displayed together with a map of the model fit residual for each fitting region. These error maps have been constructed by setting each pixel with each fitting region with the value of its associated residual after model fitting. The error maps give a visual impres-

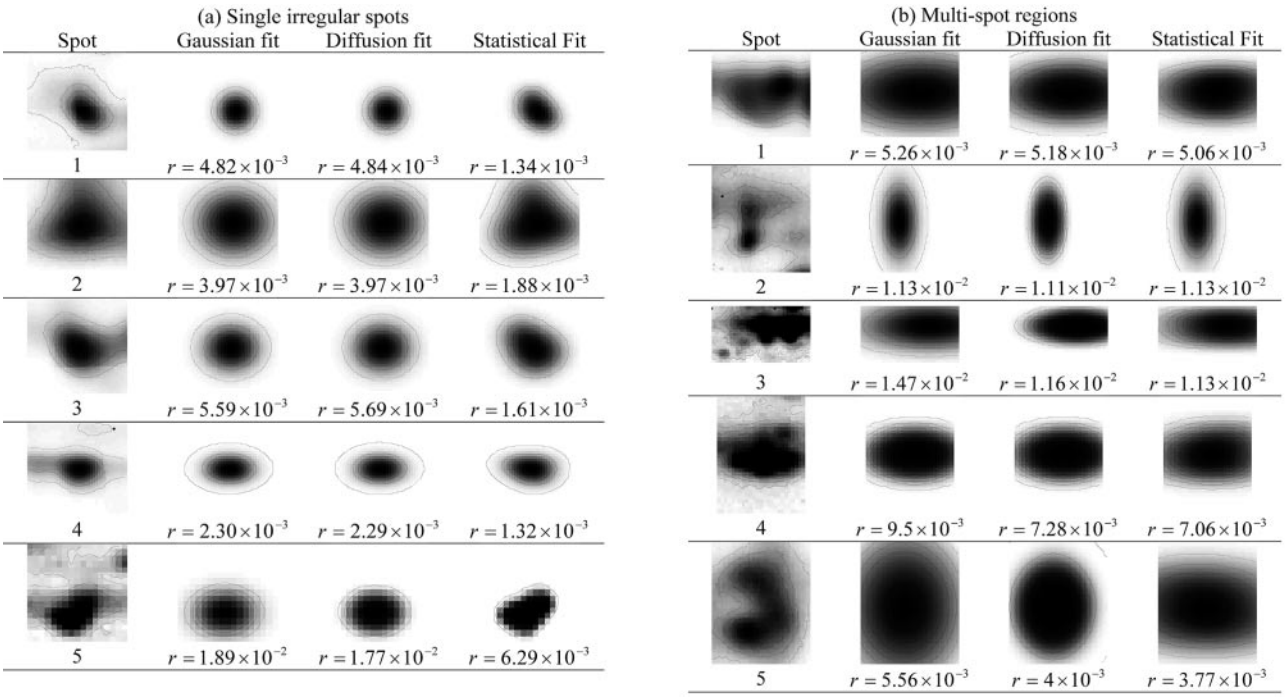


Figure 7. Example fits of each model to spot regions from the images shown in Fig. 5, with resulting fit residuals for each model. (a) shows examples of regions containing single spots with irregular shape, and (b) shows regions containing multiple spots.

sion of the spatial distribution of residual values. High values (light pixels) in these images indicate poor model fits. The appearance of the statistical reconstruction (Fig. 8(c)) away from very badly fitted regions (multiple spot regions) is rather more convincing than the reconstructions of the Gaussian and diffusion models ((Figs. 8(a) and (b)). The pattern of high residuals in the error maps for each of the models is similar. If a threshold value were chosen for each model to discriminate regions where the model had been incorrectly fitted to multiple spots, approximately the same regions would be identified, regardless of the model used. However, the statistical model can better represent irregular single spots leading to significantly lower fitting residuals in these regions. A Gaussian model would result in high fit residuals for these single spots and thresholding may erroneously identify the regions as containing multiple spots. An area of each image in Fig. 8 has been highlighted with an ellipse. The highlighted area contains many spots which have been adequately separated into fitting regions by the original watershed process (see Fig. 5(a)). The difference between the Gaussian and statistical reconstruction images is visually apparent as a general 'sharpening' of the highlighted area, together with some differences in the shapes of some spots. The Gaussian reconstruction of the highlighted area contains several regions with relatively high residual and many regions with moderate residual values. Setting a sensible threshold on fit residual for the image would probably identify several multiple spot regions in this area. The same highlighted area in the diffusion reconstruction has generally improved fits for all spots, with some substantial reductions in residual. A threshold similar to that chosen for the Gaussian model would identify fewer multiple spot regions in the area. The same highlighted area in the statistical reconstruction contains only one region of high residual, which is present in both other reconstructions, indicating that only this spot should be classified as a multiple spot region. In this case, the high residual region corresponds to a relatively low volume spot group. Fit residuals in other regions have been substantially reduced. The statistical model fits more accurately to genuine single spots, reducing their residuals significantly and therefore reducing the likelihood of them being identified as multiple spot regions. This in turn increases model specificity and improves the ability to discriminate between single and multiple spot regions.

For statistical models, specificity is entirely determined by the training data and model generation process. The results of this evaluation show that our automatic model building scheme (Section 2.2) retains specificity that is at least as high as and in general better than that of Gaussian and diffusion models whilst increasing quantification accuracy.

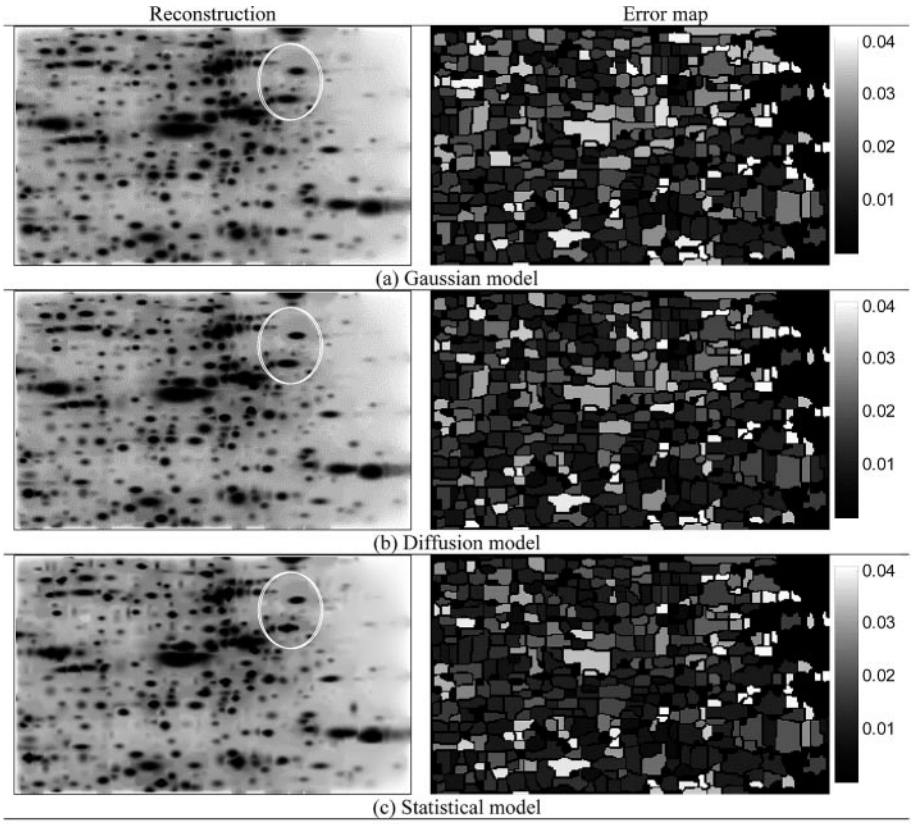


Figure 8. Reconstructed synthetic images and error maps corresponding to Fig. 5(a). (a) Gaussian protein spot model, (b) diffusion protein spot model and (c) statistical protein spot model. Values outside fitting regions have been filled by interpolation. Error maps are generated by filling fitting regions with corresponding residual value. An area of difference between the three images has been highlighted with a circle in each image.

4 Concluding remarks

In this paper, we have introduced a new statistical model of spot appearance. This model is both flexible and specific enough to represent the true range of protein spot appearance found in complex 2-DE gel images. The model provides more accurate descriptions of irregularly-shaped single spots without losing the specificity to distinguish multiple spot groups. Furthermore, the need to develop a sophisticated theoretical model of the physical processes driving irregular spot formation has been circumvented by learning the resulting shape variation in a statistical manner.

The authors are grateful for the financial support from the Biotechnology and Biological Sciences Research Council.

Received September 3, 2002

5 References

[1] Smilansky, Z., *Electrophoresis* 2001, *22*, 1616–1626.

[2] Pleißner, K.-P., Oswald, H., Wegner, S., in: Pennington, S. R., Dunn, M. J., (Eds.), *Proteomics: From Protein Sequence to Function*, BIOS, Oxford 2001, pp. 131–149.

[3] Mahon, P., Dupree, P., *Electrophoresis* 2001, *22*, 2075–2085.

[4] Appel, R. D., Hochstrasser, D. F., Funk, M., Vargas, R. J. *et al.*, *Electrophoresis* 1991, *12*, 722–735.

[5] Lemkin, P. F., Lipkin, L. E., *Comput. Biomed. Res.* 1981, *14*, 272–297.

[6] Wu, Y., Lemkin, P. F., Upton, K., *Electrophoresis* 1993, *14*, 1351–1356.

[7] Garrels, J. I., *J. Biol. Chem.* 1989, *264*, 5269–5282.

[8] Conradsen, K., Pedersen, J., *Biometrics* 1992, *48*, 1273–1287.

[9] Lemkin, P. F., Myrick, J. E., Upton, K. M., *Appl. Theor. Electrophor.* 1993, *3*, 163–172.

[10] Tyson, J. J., Haralick, R. H., *Electrophoresis* 1986, *7*, 107–113.

[11] Rowlands, D. G., Flook, A., Payne, P. I., Hoff, A. V., Niblett, T. *et al., Electrophoresis* 1988, *9*, 820–830.

[12] Cootes, T. F., Taylor, C. J., Cooper, D. H., Graham, J., *Comput. Vis. Image Underst.* 1995, *61*, 38–59.

[13] Edwards, G. J., Cootes, T. F., Taylor, C. J., in: *Proc. Eur. Conf. Computer Vision*, Springer, Freiburg, 1998, pp. 581–595.

[14] Szekely, G., Kelemen, A., Brechbuhler, C., Gerig, G., *Med. Image Anal.* 1996, *1*, 19–34.

[15] Staib, L. H., Duncan, J. S., *IEEE Trans. Pattern Analysis and Machine Intelligence* 1992, *14*, 1061–1075.

[16] Anderson, N. L., Taylor, J., Scandora, A. E., Coulter, B. P., Anderson, N. G., *Clin. Chem.* 1981, *27*, 1807–1820.

[17] Bettens, E., *Peak Characterisation Using Parameter Estimation Methods* Ph.D. thesis, University of Antwerp, 1999.

[18] Bettens, E., Scheunders, P., Dyck, D. V., Moens, L., Osta, P. V., *Electrophoresis* 1997, *18*, 792–798.

[19] Pentland, A. P., Sclaroff, S., *IEEE Trans. Pattern Analysis and Machine Intelligence* 1991, *13*, 715–729.

[20] Widrow, B., *Pattern Recognit.* 1973, *5*, 175–211.

[21] Yuille, A. L., Cohen, D. S., Hallinan, P. W., in: *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, IEEE Computer Soc., San Diego, CA June 1989, pp. 104–109.

[22] Gower, J. C., *Psychometrika* 1975, *40*, 33–51.

[23] Davies, R. H., Twining, C. J., Cootes, T. F., Waterton, J. C., Taylor, C. J., *IEEE Trans. Med. Imaging* 2002, *21*, 525–537.

[24] Campbell, N. A., *Appl. Stat.* 1980, *29*, 231–237.

33. **Automatic construction of statistical shape models for protein spot analysis in electrophoresis gels.** M. Rogers, J. Graham and R.P. Tonge *Proceedings of the British Machine Vision Conference, University of East Anglia, 2003. R. Harvey and J.A. Bangham (eds.) BMVA Press, pp 369-378.* doi:10.5244/C.17.36

Automatic Construction of Statistical Shape Models for Protein Spot Analysis in Electrophoresis Gels

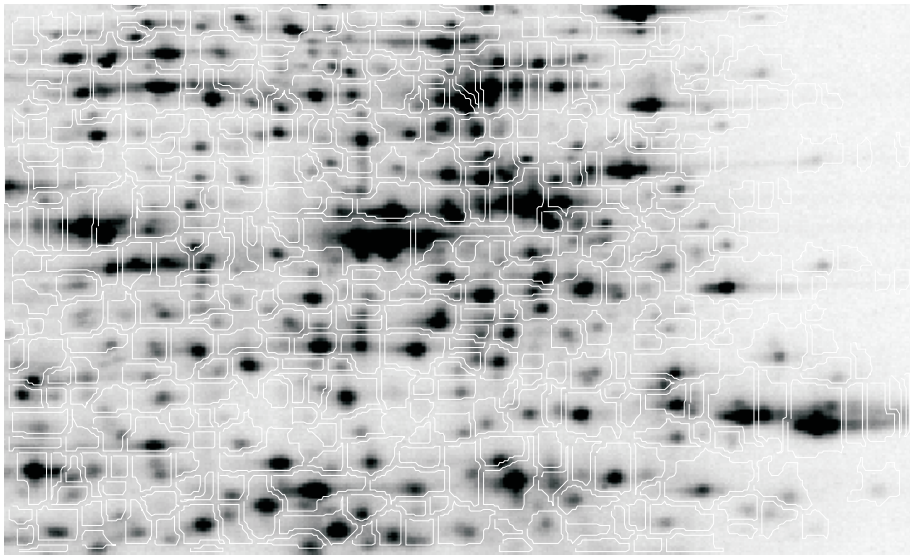
Mike Rogers[a], Jim Graham[a] and Robert P. Tonge[b]
[a]Imaging Science and Biomedical Engineering,
University of Manchester, Manchester, UK.
[b]Protein Science, Enabling Science and Technology (Biology),
AstraZeneca, Alderley Park, Macclesfield, Cheshire, UK.
mike.rogers@man.ac.uk
http://www.isbe.man.ac.uk/∼mdr/personal.html

Abstract

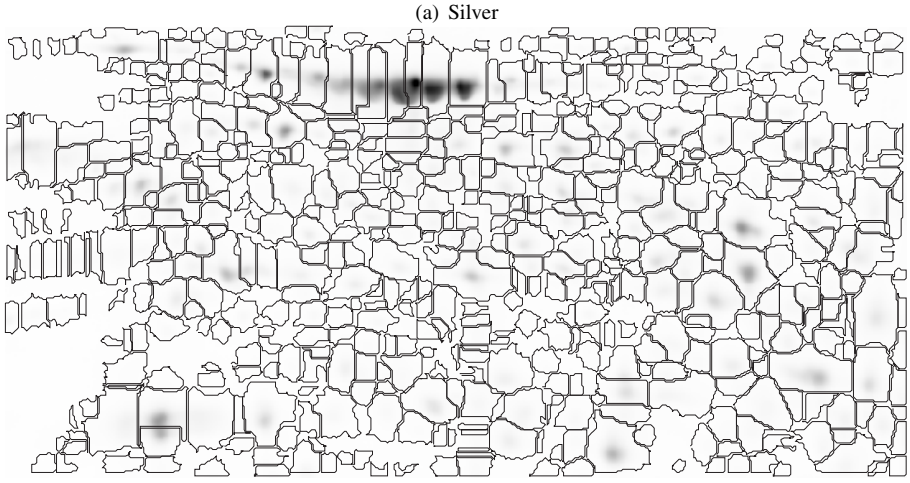
Proteomics research relies heavily on electrophoresis gels, which are complex images containing many protein 'spots'. The identification and quantification of these spots is a bottleneck in the proteomics workflow. We describe a statistical model of protein spot appearance that is both general enough to represent unusual spots, and specific enough to introduce constraints on the interpretation of complex images. We propose a robust method of automatic model construction that is used to circumvent manual model construction which is subjective and time-consuming. We show that the statistical model of spot appearance is able to fit to image data more closely than the commonly used spot parameterisations which are based solely on Gaussian and diffusion formulations.

1 Introduction

Proteomics is the study of the complete set of proteins in a cell or organism throughout the entire life-cycle. It is hoped that this research will enhance understanding of cell function in general and, more specifically, it will also identify proteins that can be used as drug targets and disease markers. The main barrier to proteomics research is complexity. It is estimated that total number of proteins in a human cell could be as large as 500,000. Key to any analysis are separation and detection technologies. A well-established and widely used technology is 2-Dimensional Electrophoresis (2-DE). This process separates protein mixtures by iso-electric point (pI) and molecular weight (MW). Separation results from two separate diffusion processes which are driven along orthogonal axes in a polyacrimide gel, resulting in a grid of protein strains. The separated proteins are visualised by pre or post staining, yielding an image, containing protein 'spots'. Figure 1 shows two segments of 2-DE gel images stained using different techniques. In practice, 3,000-4,000 spots can be visualised on a single gel image, each representing an individual protein strain. The analysis of these complex gel images is a significant bottleneck in the proteomics research workflow [6].

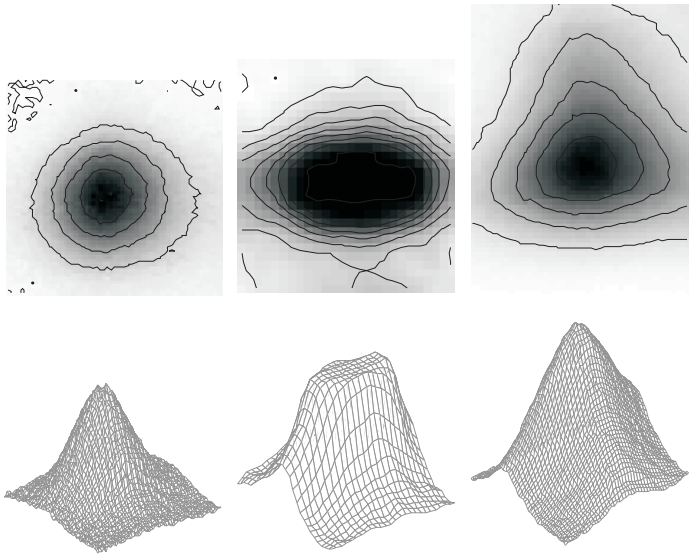


(a) Silver



(b) Fluorescent

Figure 1: Example electrophoresis images with watershed boundaries. (a) A sliver stained image with 403 delineated fitting regions. (b) A fluorescent dye image with 573 fitting regions.



(a) Gaussian (b) Flat-Top (c) Irregular

Figure 2: Examples of electrophoresis gel spots. The top row shows the appearance of the spot in the image with contours of constant gray-level overlayed. The bottom row shows a 3D mesh representation of the same data. (a) Gaussian, (b) 'Flat-top', (c) Irregular.

Image analysis of 2-DE gels requires the identification of a large number of individual spots. These must be characterised for further analysis of the sample. One of the first steps in any spot detection algorithm is the segmentation of individual spots from the background. After the segmentation step, spots are quantified and represented as a list of parameters over which further analysis can be carried out. Commonly, protein spot models are used to aid quantification by imposing constraints, which in turn improves the robustness of the solution. The most commonly used spot model is a Gaussian function [4]:

$$S(x,y) = B + I \exp\left(-\frac{(x-x_0)}{2\sigma_x^2}\right) \exp\left(\frac{(y-y_0)}{2\sigma_y^2}\right) \tag{1}$$

where B is background intensity, I is spot intensity, x_0 and y_0 control spot location and σ_x and σ_y control the spread of the Gaussian independently in x and y directions. Figure 2(a) shows an example of a typical protein spot with a Gaussian profile. This model is assumed to provide a good representation of most spots present in most gel images. However, it has been shown that Gaussian models produce an inadequate fit to some protein spots, most notably large volume, saturated spots [1]. Figure 2(b) shows an example of a high volume protein spot exhibiting a saturated, 'flat-top' shape. Bettens [1] addressed this shortcoming by proposing a model based on the physics of the spot formation process. Protein spots are formed by a diffusion process, which is only adequately represented by a Gaussian when the initial concentration distribution occupied by the sample has a small area. Bettens' diffusion model more adequately represents spots in the gel when

this assumption is not met:

$$\begin{aligned} S(x,y) = \quad & B + \frac{C_0}{2}\left[\operatorname{erf}\left(\frac{a'+r'}{2}\right) + \operatorname{erf}\left(\frac{a'-r'}{2}\right)\right] \\ & + \frac{C_0}{r'\sqrt{\pi}}\left[\exp\left(-\left(\frac{a'+r'}{2}\right)^2\right) - \exp\left(-\left(\frac{a'-r'}{2}\right)^2\right)\right] \end{aligned} \tag{2}$$

where $r' = \sqrt{\frac{(x-x_0)^2}{D'_x} + \frac{(y-y_0)^2}{D'_y}}$, B is background intensity, C_0 is initial concentration, D'_x and D'_y are related to the diffusion constants in the two main directions of diffusion, x_0 and y_0 control location and a' is the area of the disc containing the protein material. As $a \to 0$ equation 2 reduces to the bivariate Gaussin (eqn. 1).

Both the Gaussian and diffusion models assume perfect diffusion across the gel medium. Spots created by a perfect diffusion process will be regular and symmetric. In practice, the diffusion process is not perfect and spots can be formed with unpredictable, unusual shapes. An example of such a spot is shown in Figure 2(c). To represent more adequately the full range of observed spot shape, we have developed a new protein spot model that is both flexible enough to represent irregular shape variation and specific enough to retain usable constraints on the interpretation of gel images. The physical process by which irregular spots are formed is extremely complex. It would be daunting task to directly estimate all the physical variables affecting spot formation. Instead, we have used a Point Distribution Model (PDM) [3] to represent observed variation in spot shape. Gaussian convolution simulates the diffusion process and forms a full model of spot appearance. In section 2 we describe the model, together with an automatic method for model construction. Results of an evaluation of the model and a discussion are presented in sections 3 and 4.

2 Modelling Protein Spot Shape and Appearance

To represent observed variation in protein spot shape we have used a PDM trained with a set of protein spot boundaries. The PDM only represents shape, but we require a full model of spot appearance. Protein spot formation in 2-DE gels is a diffusion process which is equivalent to convolution of an initial concentration distribution with a 2-D Gaussian kernel. We have assumed the initial concentration distribution can be represented as a flat 2-D shape within the boundary represented by the shape model. This flat shape is convolved with a bi-variate Gaussian kernel giving a full model of spot appearance. Figure 3 shows an example of the full spot appearance model. We define our model using the parameter vector $\vec{p} = (B, I, x_0, y_0, \sigma_x, \sigma_y, s, \vec{b_s})$, where B is an additive background term, I is spot intensity, x_0 and y_0 control location, σ_x and σ_y control the spread of the Gaussian along the two directions of diffusion, s is a scaling for the spot shape (from the alignment procedure) and $\vec{b_s}$ is a vector of PDM shape parameters. This model is equivalent to the bi-variate Gaussian when $s = 0$, and is equivalent to the diffusion model when the shape parameters, $\vec{b_s}$, represent an elliptical shape.

2.1 Automatic Spot Model Construction

Section 2 described the basis of the models we use. Here we address the practical issue of building the model: determining the training shapes from spot images and calculating

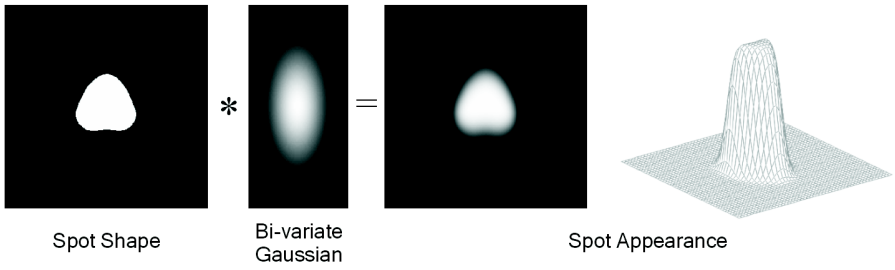


Figure 3: Spot model formation. A flat shape is convolved with a bi-variate Gaussian kernel, which is equivalent to a diffusion process.

the distributions of parameter values. In many applications of PDMs, manual marking of landmark points has been used. Due to the complexity of the images, and the number of spots required to build a model, this is an impractical strategy in this case. We proceed by segmenting the spots in the training images, smoothing the boundaries obtained using a general shape representation and making the landmark points evenly spaced round the resulting boundary. As the boundaries are extracted from real image data, a number of overlapping spots will be represented. These need to be detected and excluded from the training data, as their inclusion would bias the model and result in reduced specificity.

2.1.1 Generating the Training Set

Raw spot boundaries are obtained by thresholding the Laplacian of Gaussian transform of the training gel images. The resulting boundaries are smoothed using a Fourier shape descriptor [5] resulting in a parametrisation of the spot shape by the Fourier coefficients (5 harmonics). Spot appearance is modelled by convolving this smoothed shape with a Gaussian kernel, in the same way described in section 2. The parameters of this spot appearance model are then optimised to improve the fit to the original image data using a Levenberg Marquardt gradient descent algorithm. This provides an adjusted parametrisation of the shape matched to the image data. In this way the shapes used to build our statistical model are derived from our model of spot appearance, rather than the somewhat arbitrary data-driven segmentation. Using a Fourier representation in this strategy does not impose any explicit shape constraints on the boundaries extracted. The PDM landmark representation is obtained from the resulting spot shapes by placing 25 evenly spaced points around the boundary.

2.1.2 Robust Model Building

Automatic generation of training shapes will include incorrect shapes in the model. These shapes are the result of unseparated overlapping multi-spot groups. The Fourier shape representation imposes no explicit shape constraints, other than smoothness, so it is not possible to filter these incorrect segmentations at that stage. We could filter the resulting shapes by hand, but this would be a highly time consuming and subjective process. Rather, we have chosen to reduce the influence of such shapes by using Robust Principal

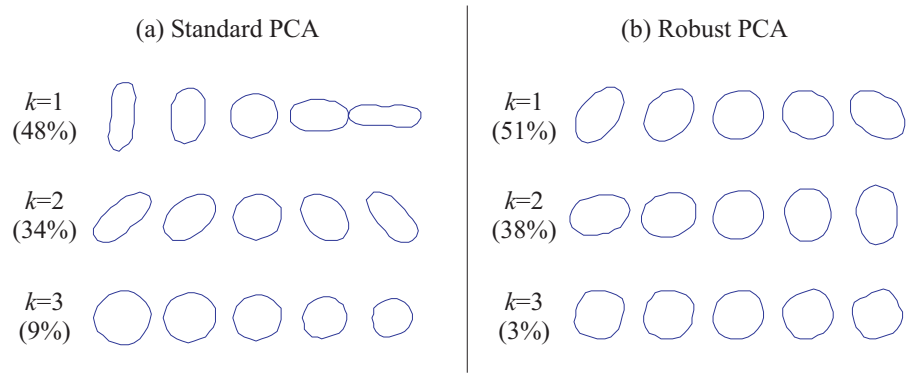


Figure 4: Robust PCA. (a) The first 3 of 10 modes (± 2 std.dev.) PDM built using standard PCA. (b) The first 3 modes of a PDM built using Robust PCA. Both models were trained with the same data.

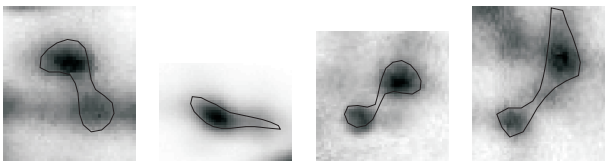


Figure 5: Four examples of shapes that have been downweighted by robust PCA. Each shape is superimposed over the image patch used in its generation.

Component Analysis [2] in the model building. We expect the number of incorrect shapes to be small and their shape to be unusual, and therefore they can only influence the model as outliers in the shape distribution. Robust PCA iteratively reduces the influence of outliers on the resulting model. The effect of the robust PCA can be seen in Figure 4. The figure shows two PDMs, one built using standard PCA (Figure 4(a)) and one built using robust PCA (Figure 4(b)). The models were generated from the same training data. Both models represent the spots by principal components that retain 99% of the observed variance, in the robust case this is 99% of the variance remaining after the iterative weighting procedure. The standard model represents the retained variance in the training data using 10 modes, whereas the robust model requires only 6 modes. The contribution of each mode to the total variance of the training set is shown for each model. The first mode of the standard model represents a large variation in aspect ratio with an apparent 'waist' becoming visible at the extremes of the mode. This mode would allow the model to represent multiple overlapping spots, which is undesirable. There is no mode in the robust model that allows shapes with 'waists'. Figure 5 shows examples of shapes that have been treated as outliers by the robust analysis. They all represent highly uncharacteristic shapes and several are clearly multiple spots.

3 Evaluation of Models

We have compared the results for fitting the statistical spot model to image data with those achieved using the Gaussian and diffusion models. The experimental procedure was as follows. Spot regions were detected in a test image using a watershed algorithm. Each of the spot models was fitted to each spot region using a Levenberg-Marquardt non-linear optimisation algorithm to determine the best model parameters, minimising the following residual: $r = \sum_{x,y \in R} \left[(S(x,y|\vec{p}) - I(x,y))^2 / \left(n_R (I_R^{max} - I_R^{min}) \right) \right]$ where R is the region of the image over which fitting takes place, $x,y \in R$ are the coordinates of the pixels within the fitting region, $I(x,y)$ are image values, $S(x,y|\vec{p})$ are the model values given the parameter vectors, I_R^{max}, I_R^{min} are the maximum and minimum image values within the region, and n_R is the number of pixels within the region. This residual provides a measure of model fit error that is normalised with respect to the intensity of the spot (which we have approximated as $I_R^{max} - I_R^{min}$) and the size of the fitting region (the number of pixels n_R). This residual form allows direct comparisons of fit quality to be made between high and low volume spots. The three models were fitted to 403 watershed delineated spots from a silver stained E.coli gel (375x228 pixels, 8 bit) and 573 spots from a gel stained with a fluorescent dye (2896x2485 pixels, 24 bit). The silver image is low-resolution and contains many saturated and overlapping spots, whereas the fluorescent image is much higher quality and contains fewer saturated or overlapping spots.

The mean residuals $\bar{r}$ for each model after fitting to all regions in both images are shown in Table 1. In general the fitting results for the fluorescent image are better due to the higher resolution of the image data. The statistical model results in the smallest average residual after fitting for both images. Figure 6 shows the mean residual for each spot model and image, grouped by volume. Group one contains the smallest 10% of spots by volume, rising to group 10 which contains the largest 10% of spots by volume. In both cases, the largest improvements in fit made by the statistical model are associated with the largest spot volumes. We have assumed that high volume spots are more likely to produce unusual spot shapes, which, we have argued, are the best represented by the statistical model. For the silver image, small and medium volume spots (groups 1-6) give fits for the Gaussian, diffusion and statistical diffusion models that are almost equivalent. However, the statistical model results in reductions in residual for all volume groups of the fluorescent image. This suggests that in the fluorescent image all spot groups contain shape variation away from Gaussian assumptions, even the smallest spots by volume. This trend is not visible in the silver image data and this may be due to the low-resolution of the image preventing full convergence. For all spot volume groups the statistical model results in fits that are better than or equivalent to the fits of the other two models. This is achieved in both images despite large visual and resolution differences.

These results demonstrate that the statistical model is able to fit well to a wide variety of gel image types. This is to be expected, as the model has the most degrees of freedom. An important question is whether the reduction in residual corresponds to a increase in model specificity. Both images contain watershed fitting regions with multiple spots. A specific model should not represent these regions well. We have carried out the following experiment to quantify the specificity of each type of model. Our aim is to determine the relative ability of the models to distinguish between single and multiple spots, using their model fit residual value. We have manually classified each fitting region in the fluorescent image (Figure 1(b), 573 regions) into one of two classes: single spot regions (472 regions)

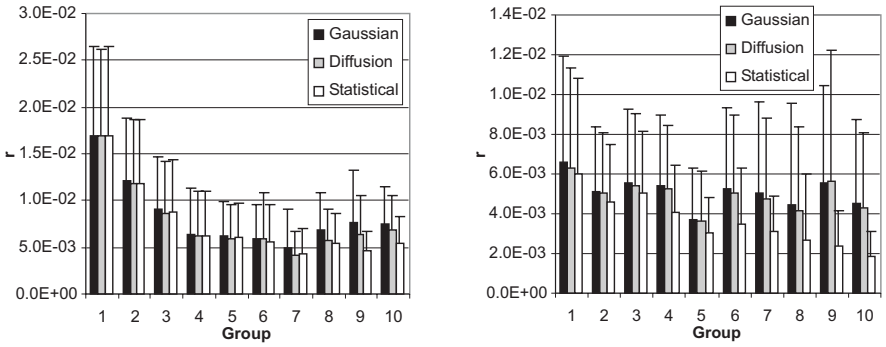
Model	Silver $\bar{r}$	Fluorescent $\bar{r}$
Gaussian	8.3×10^{-3}	5.11×10^{-3}
Diffusion	7.83×10^{-3}	4.94×10^{-3}
Statistical	7.49×10^{-3}	3.63×10^{-3}

Table 1: Mean residual after model fitting to 403 spots in the silver image and 573 spots in the fluorescent image.

or multiple spot regions (101 regions). Figure 7 shows five examples of the single spot region class, containing irregular, single spots and five examples of regions containing multiple spots, together with the fits and residuals of each model. For each of the single spot regions, the lowest residual is achieved with the statistical model. The fits of all models to multi-spot regions are visually poor (Figure 7(b)). Examination of the residuals of these 10 regions illustrates that, in general, it is not possible to define a threshold on residual value that perfectly discriminates between the two groups. This is the case for all the models. Figure 8 shows the estimated discrete probability distributions for each model for each region class. The separation of the class distributions is not good for any of the models, However, a more specific model will increase the separation between the two distributions. The distributions are non-normal, so to quantify the difference between each class we have chosen to use the non-parametric Kolmogorov-Smirnov (K-S) test. The K-S test measures the similarity between two datasets by finding the maximum discrepancy between their cumulative frequency distributions, which is called the d-statistic. The d statistic ranges between 0 and 1, the smaller the value of d, the more similar the two distributions. The discrete probability distributions (using 75 bins) and K-S distance measures for the class distributions of each model are given in Figure 8. The statistical model results in a K-S distance of $d = 0.672$, indicating that the distributions of single and multiple spot residuals are more distinct than those of the Gaussian and diffusion models ($d = 0.536$ and $d = 0.515$ respectively). This results shows that, as well as giving a more accurate quantification of 2-DE protein spots, the statistical model is more specific than the other models. The careful training and robust construction of the model results in a representation that is specific to single spots, and therefore that can not represent multiple spot regions significantly better than the other models. These selective fitting improvements lead to an increase in the separability of the two types of fitting regions.

4 Concluding Remarks

In this paper, we have described a statistical model of protein spot appearance, together with a automatic construction algorithm which takes into account the complexity of the image data. This model is both flexible and specific enough to represent the true range of protein spot appearance found in complex 2-DE gel images without the need to develop a sophisticated theoretical model of the physical processes driving irregular spot formation.



(a) Silver (b) Fluorescent

Figure 6: Mean residual $\bar{r}$ of model fit with error bars showing $+1$ std. err., plotted by increasing spot volume for each model. Spot volume group 1 contains the smallest 10% of spots by volume, rising to group 10 which contains the largest 10% of spots by volume.

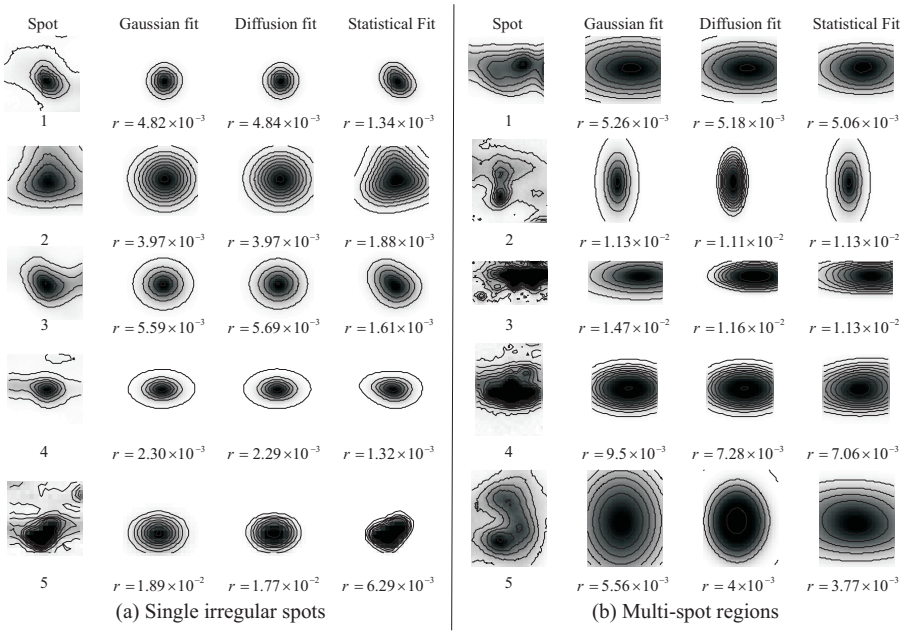
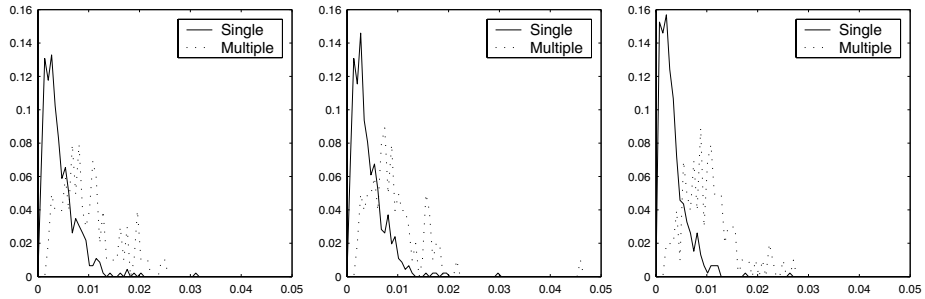


Figure 7: Example fits of each model to spot regions from the images shown in Figure 1, with resulting fit residuals for each model. (a) Shows examples of regions containing single spots with irregular shape. The improved fit of the statistical model is clear in each case. (b) Shows regions containing multiple spots. None of the models generate an adequate fit to these spots.



(a) Gaussian ($d = 0.536$) (b) Diffusion ($d = 0.515$) (c) Statistical ($d = 0.672$)

Figure 8: Discrete probability distribution (75 bins) of fit residual for single and multiple spot fitting regions with K-S distance measure. (a) Gaussian model (b) Diffusion model and (c) statistical model.

Acknowledgements

The authors are grateful for the financial support for the Biotechnology and Biological Sciences Research Council.

References

[1] E Bettens, P Scheunders, D Van Dyck, L Moens, and P Van Osta. Computer analysis of two-dimensional electrophoresis gels: A new segmentation and modelling algorithm. *Electrophoresis*, 18:792–798, 1997.

[2] N A Campbell. Robust procedures in multivariate analysis I: Robust covariance estimation. *Applied Statistics*, 29(3):231–237, 1980.

[3] T F Cootes, C J Taylor, D H Cooper, and J Graham. Active Shape Models - their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, January 1995.

[4] J I Garrels. The QUEST system for quantitative anaylsis of two-dimensional gels. *Journal of Biological Chemistry*, 264(9):5269–5282, March 1989.

[5] L H Staib and J S Duncan. Boundary finding with parametrically deformable models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(11):1061–1075, November 1992.

[6] T Voss and P Haberl. Observations on the reproducability and matching efficiency of two-dimensional electrophoresis gels: Consequences for comprehensive data analysis. *Electrophoresis*, 21:3345–3350, 2000.

34. **Using statistical image models for objective evaluation of 2D gel image analysis.** M.D. Rogers, J. Graham and R.P. Tonge *Proteomics 3: 879-886, 2003.* doi: 10.1002/pmic.200300420

Mike Rogers[1]
Jim Graham[1]
Robert P. Tonge[2]

[1]Imaging Science and Biomedical Engineering, University of Manchester, Manchester, UK
[2]Protein Science, Enabling Science and Technology (Biology), AstraZeneca, Macclesfield, Cheshire, UK

Using statistical image models for objective evaluation of spot detection in two-dimensional gels

Protein spot detection is central to the analysis of two-dimensional electrophoresis gel images. There are many commercially available packages, each implementing a protein spot detection algorithm. Despite this, there have been relatively few studies comparing the performance characteristics of the different packages. This is in part due to the fact that different packages employ different sets of user-adjustable parameters. It is also partly due to the fact that the images are complex. To carry out an evaluation, "ground truth" data specifying spot position, shape and intensities needs to be defined subjectively on selected test images. We address this problem by proposing a method of evaluation using synthetic images with unambiguous interpretation. The characteristics of the spots in the synthetic images are determined from statistical models of the shape, intensity, size, spread and location of real spot data. The distribution of parameters is described using a Gaussian mixture model obtained from training images. The synthetic images allow us to investigate the effects of individual image properties, such as signal-to-noise ratios and degree of spot overlap, by measuring quantifiable outcomes, *e.g.* accuracy of spot position, false positive and false negative detection. We illustrate the approach by carrying out quantitative evaluations of spot detection on a number of widely used analysis packages.

Keywords: Commercial gel analysis packages / Objective evaluation / Spot detection / Spot modelling/ Statistical models / Two-dimensional gel image analysis PRO 0420

1 Introduction

The complexity of data in two-dimensional electrophoresis (2-DE) gel images requires the use of computer aided analysis packages. Central to any gel analysis package is its protein spot detection algorithm. There are many commercially available applications implementing spot detection algorithms designed for 2-DE gel analysis. Despite the importance of spot detection, few studies have compared the relative performances of the various available systems [1]. This may be due to the fact that the images contain extremely complex spot patterns, making it difficult to determine the definitive correct segmentation. Without a 'ground truth' annotation, objective performance evaluation is impossible.

A potential solution to this problem is to evaluate performance using a synthetic data set, generated with precisely known parameters. This eliminates the need for time-consuming and subjective annotation of real data, allowing quantitative and objective evaluations to be performed. If this approach is to be used, and valid conclusions reached, it is important that the synthetic data truly reflects all the characteristics of the real data.

In this paper we describe methods for creating a synthetic data set with properties that match real gel images. We propose several evaluation measures and illustrate them using a number of widely used gel image analysis packages: ImageMaster (Nonlinear Dynamics, Newcastle upon Tyne, UK), Melanie (Geneva BioInformatics, Geneva, Switzerland), PDQuest [2] (Bio-Rad Laboratories, Hercules, CA, USA), Progenesis (Nonlinear Dynamics) and Z3 [3] (Compugen, Tel Aviv, Israel). We stress that this paper does not represent a definitive evaluation of any of these packages. Each makes use of a number of internal parameters and processes that cannot be readily controlled, making a detailed comparative evaluation meaningless. Rather we present the results as examples of using objective evaluation criteria. The synthetic data we have used in the evaluation have been generated using a statistical model of protein spot appearance*. This model has been shown to represent the variation in spot appearance and shape more completely than previously proposed models.

Correspondence: Dr. Mike Rogers, Imaging Science and Biomedical Engineering, Stopford Building, University of Manchester, Oxford Road, Manchester, M13 9PT, UK
E-mail: mike.rogers@man.ac.uk
Fax: +44-161-275-5145

Abbreviations: FP, false positive; **FN**, false negative; **TP**, true positive; **TPF**, true positive fraction

* Rogers, M., Graham, J., Tonge, R. P., *Proteomics* 2003, this issue.

© 2003 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim 0173-0835/03/0606–879 $17.50+.50/0

2 Materials and methods

2.1 Synthetic data

The parameters describing individual spots and the characteristics of collections of spots in gel images are derived from training images. Parameters describing individual spots are derived from a statistical spot model described in our companion paper in this issue [11]. Spatial relationships – spot locations and relative spot intensities – are described using a statistical gel model and scaled to a common reference frame. Each spot in a synthetic image is represented by a vector containing parameters describing shape, intensity, size, location and point spread. The vectors contain 12 variables, six representing spot shape.

These vectors can be thought of as a discrete set of samples drawn from an underlying continuous probability density function (PDF). This PDF represents the likelihood of a spot with particular characteristics occurring in a given image. Our model is designed to represent this PDF. The simplest model of the PDF is to assume a uniform distribution across the entire space, in which case every possible spot configuration is equally likely. This is not a realistic model of the gel image spot characteristics, and ignores all training data. In general, the training samples are not well represented by any standard parametric distribution, such as Gaussian. We have chosen to estimate the form of the PDF by modelling it as a mixture of Gaussians. This approach is a well established statistical technique in a number of applications [4, 5]. A weighted sum of multivariate Gaussian kernels is used to approximate the PDF given the training samples. It is not possible to calculate the parameters for each kernel analytically from the data, instead we utilize an iterative Expectation Maximisation (EM) algorithm for this purpose [6]. This algorithm requires the number of kernels to be fixed in advance. We have chosen to use 40 kernels in our model.

Samples can be drawn at random, or in a more systematic way, from the model to construct artificial spots with precisely known characteristics. This method can be used to generate copies of real images, or entirely new images. Figure 1 shows a segment of a 2-DE gel image, taken from the set used by the manufacturers of Z3 in their evaluations [3]. Figure 2 shows a quantified copy of the image formed by fitting protein spot models to each spot in the original. In this figure, the intense spot train in the centre of the figure is not well represented in the reconstruction. During the fitting phase, the spots in this feature were not separated and the model was fitted to an image patch containing multiple spots. The statistical spot model only represents valid shape variation and

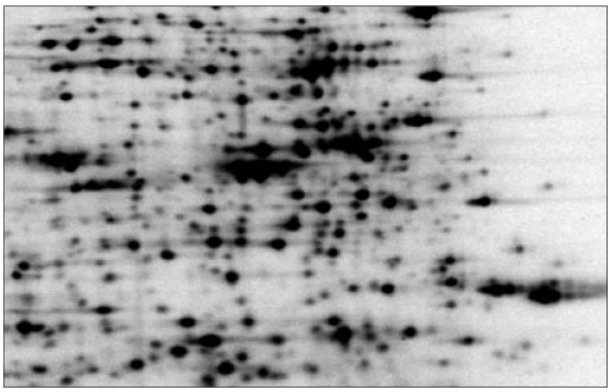


Figure 1. A section from a real 2-DE gel image, downloaded from http://www.2dgels.com/benchmark/. The image is a section from a silver stained gel of *Escherichia coli* under standard conditions.

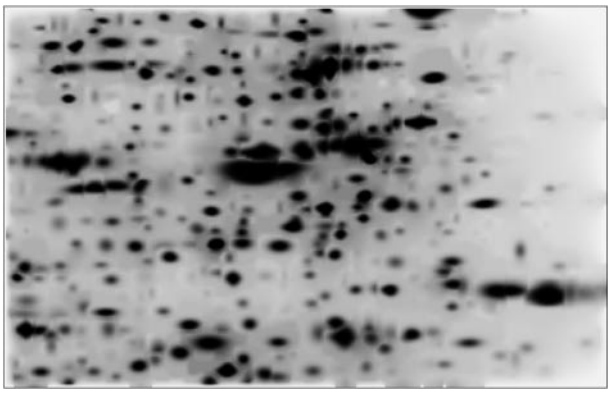


Figure 2. A model-generated synthetic gel image generated using parameters calculated from the spots in Fig. 1. Multiple spots are not represented in the model, thus regions containing overlapping spots are not reconstructed well.

therefore does not reconstruct the multiple spot regions well. The model is specific to valid single spot regions. A Gaussian spot reconstruction of this image would have a similarly poor representation of multiple spot regions. Figure 3 shows a completely synthetic image formed by taking random samples from the PDF of the gel image model. In this case, the spots have been placed on a uniform background. It is also possible to generate a background model (Fig. 4), although that has not been used in the evaluation described here.

The evaluation has been designed to examine spot detection performance. We wish to remove as many confusing factors from this evaluation as possible, so as to focus on precisely known and quantifiable effects and trends. To this end we have not included any characteris-

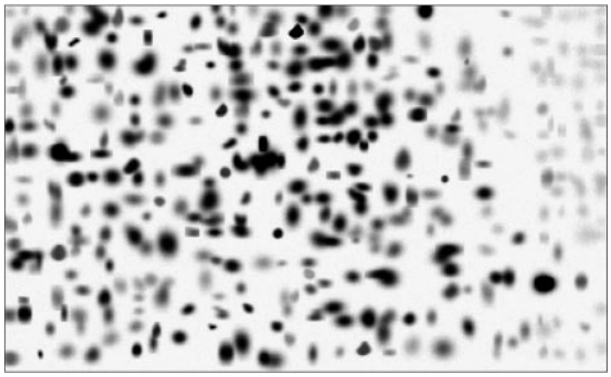


Figure 3. A synthetic gel image generated using random parameters drawn from a PDF learned from Fig. 1, on a flat background.

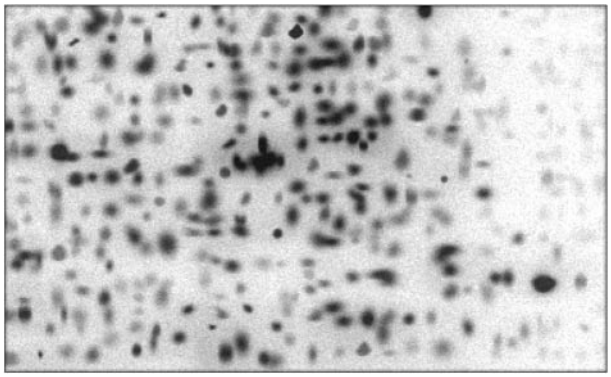


Figure 4. A synthetic gel image generated using random parameters drawn from a PDF learned from Fig. 1, with smoothly varying synthetic background with random noise.

tics that are not being directly studied from the images used in each experiment. For example, we have not performed an evaluation of the effects of varying background on spot detection. Therefore all generated images contain spots placed on a uniform background. Images produced in this way may not appear entirely realistic, as they do not contain background patterns observed in real data. However, they allow a nonsubjective, precisely quantified evaluation of the effects of individual image characteristics on spot detection performance. Trends identified with these techniques may be useful in explaining the performance differences between gel image analysis systems when applied to real data.

2.2 Evaluation

There are many different ways of quantifying spots that have been detected on a gel, and different approaches are taken by each of the applications we have evaluated. Many of the measures are not directly comparable between packages. For this reason, we have based our analysis on the coordinates of detected spot centers, which is the simplest and most consistent measure available in each package. We make two types of measure on spot centers: accuracy and detection/failure rate. Accuracy is defined as being the Euclidian distance between the detected spot centre and the 'ground truth' coordinates of the synthetic spot center. This distance is normalized as a percentage of the known extent of the spot. We have defined the extent of a generated spot to be:

$$SE = \bar{r} + 2\bar{\sigma} \quad (1)$$

where $\bar{r}$ is the mean size of the spot shape through the *x* and *y* axes, and

$$\bar{\sigma} = (\sigma_x + \sigma_y)/2 \quad (2)$$

the mean of the spot models' point spread parameters σ_x, σ_y which correspond to the standard deviations of a bivariate Gaussian function. In a successful detection there is an exact correspondence between the detected coordinate and the 'ground truth'.

Detection rate corresponds to the number of spots successfully detected. We have defined a spot to be successfully detected when the distance between its measured center and the nearest 'ground truth' coordinate is less than a defined threshold. For the experiments below, we define the threshold to be 10% of the extent of the spot. A successfully detected spot is a true positive (TP) detection, a missed spot (no detection within the threshold distance from its center) is a false negative (FN). The proportion of TP to the true number of spots is the true positive fraction (TPF). A false positive (FP) is a detection which does not occur close enough to any 'ground truth' centre, or where more than one detected spot corresponds to a single center.

2.3 Experimental design

Using synthetic images of isolated spots, and synthetic gel images we conducted several quantitative evaluations.

2.3.1 Sensitivity and accuracy

Some packages provide user-adjustable parameters controlling spot detection. Table 1 gives a summary of the parameters available for each package. We have evaluated the sensitivity of detection performance on various parameter choices using a gel image with approximately 400 spots (Fig. 3). For each application, parameters were set at a range of values and the TPF and number of FPs

Table 1. Summary of the software packages used in the evaluation[a]. The spot detection parameters used in the FROC analysis are given together with the values at which detection performance was evaluated.

Package	Parameters Evaluated [Values]	Notes
ImageMaster 2D Elite™ V3.01[b]	*Sensitivity* [0, 2000, 4000, 6000, 8000, 10 000], *Operator Size* [5, 7, 9, 11, 15, 21, 25]	Other parameters available are: *noise factor* and *background* which were held constant at values of 3 and 1 respectively. Has a parameter estimation wizard to help determine parameter values.
Melanie™ V3.08g	*Laplacian Threshold* [0, 20, 40, 60, 80, 100], *Partials Threshold* [0 20 40 60 80 100]	Other parameters available are: *smoothing, minimum perimeter, saturation* and *peak height* which were held constant at values of 2, 90, 100 and 10 respectively. Optional *Gaussian fitting* after detection was disabled.
PDQuest™ V7.0.0	*Sensitivity* [0.1, 0.2, 0.4, 0.6, 1, 5, 10, 100, 1000]	Also has a *peak height* parameter which was held constant at 10. Optional parameters control preprocessing such as *smoothing* and *streak removal*. All preprocessing was disabled.
Progenesis™ Workstation V2002.01	None	Parameter free spot detection.
Z3™ V2.1	*Dynamic Range* [2, 5, 8, 12, 15]	Spot detection results can be filtered by *size* and *confidence*. For this evaluation all spots detected were used with minimum filtering values of 1 and 0 respectively.

a) The software packages are ordered alphabetically by brand name
b) ImageMaster 2D Elite is essentially the same package as Phoretix 2D Advanced™, marketed under different brand names. Please contact vendors for details of any differences

measured. The parameters resulting in the best ratio of TPF to FP were used to produce an accuracy measurement for each package. The measure consisted of the distance between detected and known spot position as a proportion of spot extent for each TP detection.

2.3.2 Sensitivity to signal-to-noise ratio and spot overlap

The two previous experiments used a full image with many spots. To evaluate some of the specific characteristics of the detection algorithms we have performed evaluations on images containing only one or two spots. These have been designed to test the sensitivity of the packages to the S/N and the amount of spot overlap. In each case, the best package parameters, determined from the full image analysis, were used for detection.

2.3.2.1 Signal-to-noise

An image set was created consisting of a single spot with additive Gaussian random noise of variable magnitude. The S/N of the images was raised from 7 dB to 30 dB. We have used the definition:

$$S/N = 10 \log [V_s/\sigma_n] \quad (3)$$

where V_s is the magnitude of the signal and σ_n is the standard deviation of the noise. Figure 5 shows examples of images from this image set. The number of TP, FP and FN were measured for each package at each S/N.

2.3.2.2 Spot overlap

In this experiment, a synthetic image of two overlapping spots was created. The proportion of overlap was varied between 0–100%. We have defined overlap in terms of the distance between the two spot centers:

$$100\left(1 - \left(\frac{d}{SE_1 + SE_2}\right)\right) \quad (4)$$

where d is the distance between spot centres and SE_1, SE_2 are the extents of the two spots.

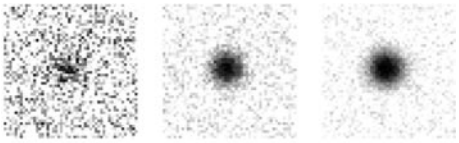


Figure 5. Three examples from the signal-to-noise evaluation set with their S/N.

Multiple images were created for each overlap proportion with different spot size ratios. The size ratio was varied between 1:1 to 1:4. In this experiment, the expected failure mode is in detecting the overlapping spots as a single spot. The number of failed detections was measured at each overlap condition. Figure 6 shows six examples from the set of 40 images.

3 Results and discussion

3.1 Sensitivity and accuracy

Results are presented as free-response receiver operator characteristic (FROC) curves, which show the relationship between TPF and number of FPs as the parameter is varied. Figs. 7–10 show FROC curves for each package and parameter.

The FROC analysis of the effect of varying parameter values demonstrates the different amount of control offered by each application. The form of the FROC curves vary significantly from the ideal [7] indicating the complex effect that the parameters have on spot detection. ImageMaster parameters (Fig. 7) control detection performance through a wide range, as do PDQuest and Melanie. Changes in ImageMaster parameter values result in fairly smooth changes in performance. Conversely, Melanie's laplacian threshold and PDQuest's sensitivity (Figs. 8(a) and 9) both contain points of discontinuity, around which small increases in value can lead to a disproportionately large increase in the number of FP detections. It may be difficult for an operator to identify such critical parameter values in the absence of known 'ground truth' data. It may therefore be difficult to find the optimal performance of these packages manually for real images. Figure 10 shows the effect of altering Z3's dynamic range parameter. This parameter appears to have little effect on the performance of the detection.

The best value of each parameter for each package was determined from these curves as the point with the largest TPF/FP ratio. These values were used to calculate the accuracy of each package. Accuracy results, together with the TPF and number of failures of each type are presented in Table 2. For this data, the two packages with the highest TPF were ImageMaster (85.1%) and Progenesis (84.6%). Z3 achieved the lowest TPF value of 68.2%. This is somewhat surprising, as a previous study [1] found that Z3 outperformed Melanie in terms of TPF. However, in [1] the evaluation was carried out on real images after manually identifying spots. Detection rates were also determined manually, making the study highly subjective.

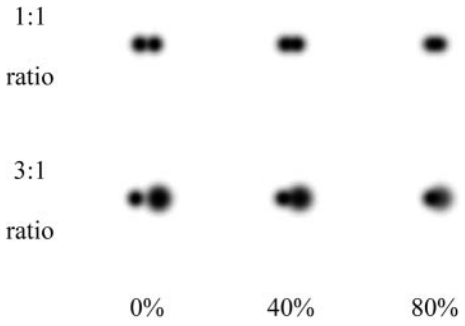


Figure 6. Example images from the spot overlap image set.

Table 2. Accuracy of spot detection using best parameters. The table shows the average accuracy (as a proportion of spot extent) for detected spots, with the number of False Positive (FP) detections and the number of False Negative (FN) detections

Package	Accuracy	TPF	FP	FN
ImageMaster 2D Elite	3.15%	85.1%	40	58
Melanie	3.40%	81.5%	26	72
PDQuest	5.02%	78.2%	20	85
Progenesis	3.46%	84.6%	26	60
Z3	4.46%	68.2%	30	124

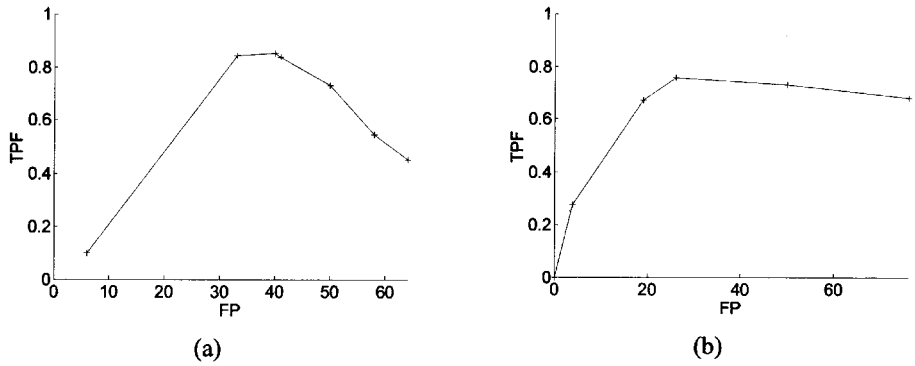


Figure 7. FROC curves for ImageMaster spot detection parameters. (a) Operator size at values [5, 15, 25, 35, 45, 51], sensitivity parameter was held constant at a value of 8475. (b) Sensitivity at values [1, 2000, 4000, 6000, 8000, 10 000], operator size parameter was held constant at a value of 31.

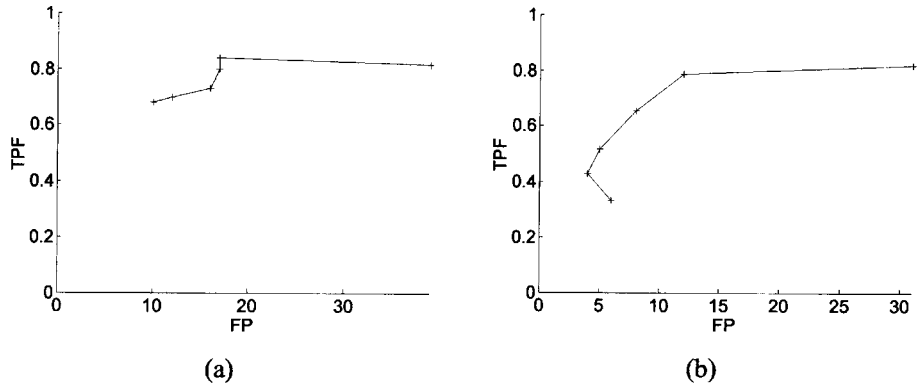


Figure 8. FROC curves for Melanie spot detection parameters. (a) Laplacian threshold at values [0, 20, 40, 60, 80, 100], partials threshold parameter was held constant at a value of 20. (b) Partials threshold at values [0, 20, 40, 60, 80, 100], Laplacian threshold parameter was held constant at a value of 20.

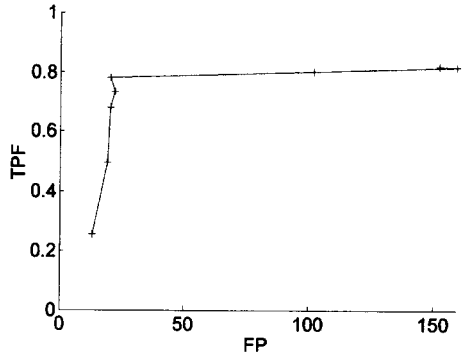


Figure 9. FROC curve for PDQuest spot detection parameter sensitivity at values [0.1, 0.2, 0.4, 0.6, 1, 10, 100, 1000]

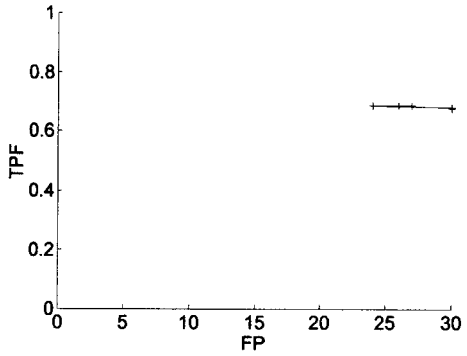


Figure 10. FROC curve for Z3 spot detection parameter dynamic range at values [2, 5, 8, 12, 15].

The evaluations were also influenced by many unquantified image characteristics occurring in the set of images used. Our experiment investigated only a subset of known image factors. In particular no background variation was included in our evaluation. It may be supposed, therefore, that Z3 is particularly sensitive to background variation. However, the unquantified nature of the real data makes a direct comparison between the two studies difficult. Further investigation into the effects of quantified background and other image characteristics is required.

The most accurate package was ImageMaster, which measured the centers of all detected spots to be located within 3.15% of their true extent. PDQuest detected the fewest FP spots, but was the least accurate on the TP spots it detected. Overall, the accuracy of ImageMaster, Melanie and Progenesis are roughly similar. However Progenesis has the distinct advantage of being a parameter free detection, removing the need to manually determine the best set of parameters. ImageMaster and PDQuest each provide a semi-automatic method of determining parameter values with a small amount of user interaction. However, for both packages, the semi-automatic parameter values did not correspond with the values determined from the FROC analysis, which may reflect the subjectivity still contained in the process.

3.2 S/N and spot overlap

Table 3 presents the TPF and number of failures for the entire S/N data set. Figure 11 shows the number of FP detections for each package at each S/N value. Three of the five packages have successfully detected the correct spot in all the images. Progenesis and PDQuest failed to

Table 3. Detection at varying signal-to-noise ratios. Detection rates across all ten images. Parameter values used were the same as for Table 2.

Package	TPF	FP	FN
ImageMaster 2D Elite	100%	15	0
Melanie	100%	15	0
PDQuest	70%	12	3
Progenesis	70%	14	3
Z3	100%	11	0

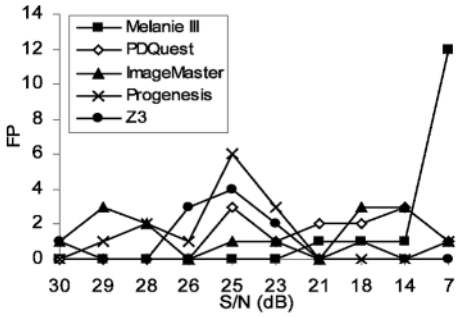


Figure 11. Number of FPs for each package in analyzing an image containing a single spot over a range of S/N.

detect the spot in three images with low S/N. Progenesis does not allow user control of spot detection parameters. It is therefore impossible to ensure that the internal parameters used in the detection remained the same for each image. Each other application had all parameter values fixed, so direct comparison of these results may be misleading. Z3 resulted in the fewest FP detections overall, but these all occurred at relatively high S/N values (see Fig. 11). This supports the theory that Z3 is sensitive to particular background patterns, as random noise causes an increase in FPs even at low magnitudes. Similar effects are observable for the distribution of FPs for the other packages, indicating sensitivity to difference background patterns. Melanie resulted in the most FP detections, however the majority (12 of 15) were detected only in the image with the lowest S/N. If this image had not been included, Melanie would result in the best performance.

Figure 12 shows the TPF for each package at each overlap proportion. All packages fail for some images when the overlap proportion is 60% or greater. In this evaluation, Melanie appears the most sensitive to spot overlap, failing in some images with only a 20% overlap. PDQuest, one of the worst packages in the S/N evaluation, is the most robust to spot overlap.

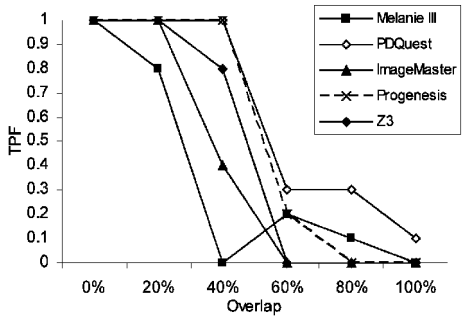


Figure 12. Detection sensitivity in analyzing an image of two spots with different amounts of overlap.

4 Concluding remarks

Evaluation of system performance has in the past been highly subjective [1] and focussed on specific aspects of individual systems [8–10]. This is partly due to the fact that the operating parameters of image analysis systems are largely hidden from users. This is a perfectly sensible state of affairs if a system is to be used for normal laboratory purposes, but creates difficulties for objective evaluation.

Part of the reason for difficulty of evaluation is also the complex nature of the images. The wide variability of spot shape and size, background and density of spots makes it very difficult to approach the problem in the standard scientific manner of isolating individual parameters for evaluation. We have sought to deal with this problem by creating synthetic gel images in which all parameters are under experimental control. For this approach to be useful it is important that the synthetic images are realistic representations of true gel images. The ability to model the parameters of spot appearance is an essential contribution to this end. It may be considered that the appearance of Fig. 3 is not realistic, in the sense that there is no background variation. It would be possible to model and superimpose a synthetic background on the image. While being more realistic in appearance, this would introduce additional complexity from the point of view of quantitative comparison. There would be further parameters to vary in the ‘complete gel’ experiments. It may be appropriate and informative to extend these experiments in this way. In the present study, the issue of random background variation has been investigated by means of the S/N parameter in the ‘single spot’ studies.

We have presented an experimental approach to evaluating system performance, and applied this to a number of commonly used commercial systems. These were used as evaluation copies obtained from the manufacturers. Our intent was not to conduct a definitive comparative evaluation. There are sufficient hidden parameters in each system to make a direct comparison difficult. Rather we sought to show that meaningful performance measures can be obtained using this objective approach. The (perhaps unsurprising) conclusion from the evaluation is that the different packages show different strengths and weaknesses. ImageMaster is the most accurate package, Z3 the most robust to poor S/N and PDQuest the most robust to spot overlap. Melanie performs well in all evaluations and Progenesis has the advantage of a parameter free spot detection, while also performing well in most evaluations.

The scope of the objective evaluation presented here is not complete. Other interesting parameters to model and synthesize would include the nature of the background of

gel images. However, we feel that this generic approach provides a valuable mechanism for understanding the capabilities of image analysis methods.

The authors are grateful for the financial support for the Biotechnology and Biological Sciences Research Council.

Received September 3, 2002

5 References

[1] Raman, B., Cheung, A., Marten, M. R., *Electrophoresis* 2002, *23*, 2194–2202.

[2] Garrels, J. I., *J. Biol. Chem.* 1989, *264*, 5269–5282.

[3] Smilansky, Z., *Electrophoresis* 2001, *22*, 1616–1626.

[4] McLachlan, G. J., Basford, K. E., *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker, New York 1988.

[5] Everitt, B. S., Hand, D. J., *Finite Mixture Distributions*, Chapman and Hall, New York 1981.

[6] Bilmes, J. A., *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models*, Technical Report ICSI-TR-97-021 International Computer Science Institute, Berkeley, CA, USA 1998.

[7] MacMillan, N. A., Creelman, C. D., *Detection Theory: A Users Guide*, Cambridge University Press, New York 1991.

[8] Lemkin, P. F., Lipkin, L. E., *Comput. Biomed. Res.* 1981, *14*, 272–297.

[9] Garrels, J. I., *J. Biol. Chem.* 1979, *254*, 7961–7977.

[10] Tonge, R., Shaw, J., Middleton, B., Rowlinson, R. *et al., Proteomics* 2001, *1*, 377–396.

[11] Rogers, M., Graham, J., Tonge, R. P., *Proteomics* 2003, *3*, 887–896.

35. **Robust and accurate registration of 2-D electrophoresis gels using point matching.** M. Rogers and J. Graham, *IEEE Transactions on Image Processing 16: 624-635, 2007.* doi: 10.1109/TIP.2007.891342

Robust and Accurate Registration of 2-D Electrophoresis Gels Using Point-Matching

Mike Rogers and Jim Graham, *Member, IEEE*

***Abstract*—Point-matching is a widely applied image registration method and many algorithms have been developed. Registration of 2-D electrophoresis gels is an important problem in biological research that presents many of the technical difficulties that beset point-matching: large numbers of points with variable densities, large nonrigid transformations between point sets, paucity of structural information and large numbers of unmatchable points (outliers) in either set. In seeking the most suitable algorithm for gel registration we have evaluated a number of approaches for accuracy and robustness in the face of these difficulties. Using synthetic images we test combinations of three algorithm components: correspondence assignment, distance metrics and image transformation. We show that a version of the iterated closest point (ICP) algorithm using a non-Euclidean distance metric and a robust estimation of transform parameters provides best performance, equalling SoftAssign in the presence of moderate image distortion, and providing superior robustness against large distortions and high outlier proportions. From this evaluation we develop a gel registration algorithm based on robust ICP and a novel distance metric combining Euclidean, shape context and image-related features. We demonstrate the accuracy of gel matching using synthetic distortions of real gels and show that robust estimation of transform parameters using M-estimators can enforce inverse consistency, ensuring that matching results are independent of the order of the images.**

***Index Terms*—Biomedical image registration, iterated closest point (ICP), robust point matching (RPM), 2-D electrophoresis (2-DE) gels.**

I. Introduction

TWO-dimensional electrophoresis (2-DE) is a method of protein separation used in the field of Proteomics. The technique results in a matrix of diffuse spots which can be visualized by pre or post staining. Each of these spots is a separated protein strain. The volume of each spot is proportional to the amount of each protein in the original sample. In practice, 3 000–4 000 spots can be visualized on a single gel image. Many recent studies involve differential analysis of sets of up to 100 2-DE gels. To carry out a differential investigation, it is necessary to determine correspondence between spots on sets of gel images. The production of 2-DE gels is inherently variable. As a result complex nonlinear deformations are often required to align comparable gels. These deformations are often quite large, difficult to identify manually and extremely time-consuming to correct. Fig. 1 shows two examples from a replicate gel set [Fig. 1(a) and (b)] together with two synthetically deformed gels [Fig. 1(c) and (d)] used in our evaluation (Section III-C). The amount of deformation introduced is typical of the type of deformations encountered in gel analysis, however, the magnitude of the synthetic deformation shown in Fig. 1(d) is larger than that typically encountered in replicate gel sets.

A typical gel pair will take 4–8 h of manual spot matching correction using current generations of software, if automatic analysis fails. This level of user-input quickly becomes untenable as the size of the gel sets increase. The analysis of these complex gel images is a significant bottleneck in the proteomics research workflow [1], which can be alleviated by improved automatic analysis techniques.

In this work, we have focused on the problem of aligning a pair of gels. Matching a pair of gel images is an image registration problem, and as such, algorithms that are directly driven by image-based metrics, such as mutual information or sum of squared differences (SSD), could be applied. However, the numerous small spot features in the images produce a search space that contains many local minima. We contend that it is unlikely that any image-based local optimisation algorithm can be sufficiently robust against these minima. The inherent local convergence properties of such algorithms, combined with the relatively large nonrigid and nonsystematic deformations found in gel images, make even complex multiresolution search methods, which utilize multilevel regularisation, unlikely to determine a globally optimal registration. Instead, and as rough segmentation of protein spot features is fairly easy to achieve, we have taken the approach of matching these features directly using a robust, constrained point-matching algorithm. Point-matching of protein spot features presents all of the technical difficulties that beset such applications: large numbers of points with varying densities, large nonrigid transformations between point sets, significant numbers of unmatchable points (outliers) in either set, and paucity of structural information to help identify correspondences. In this study, we carry out an extensive evaluation of the components of a number of point-matching algorithms using point sets that simulate the appearance of gels. In this way, we control parameters such as deformation and outlier proportions, and identify how each

Manuscript received September 22, 2005; revised September 22, 2006. This work was supported in part by the Biotechnology and Biologicial Science Research Council (U.K.), award number 34/E14651, and in part by the Interdisciplinary Research Collaboration "From Medical Images and Signals to Clinical Information" (Engineering and Physical Sciences Research Council award number Gr/N14248/01 and Medical Research Council award number D2025/31, U.K.). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Nicolas Rougon.

M. Rogers is currently with Image Metrics, Ltd., Castlefield, Manchester M3 4SW U.K. (e-mail: mike.rogers@image-metrics.com).

J. Graham is with the The Division of Imaging Science and Biomedical Engineering, The University of Manchester, Manchester M13 9PT U.K. (e-mail: jim.graham@manchester.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2007.891342

1057-7149/$25.00 © 2007 IEEE

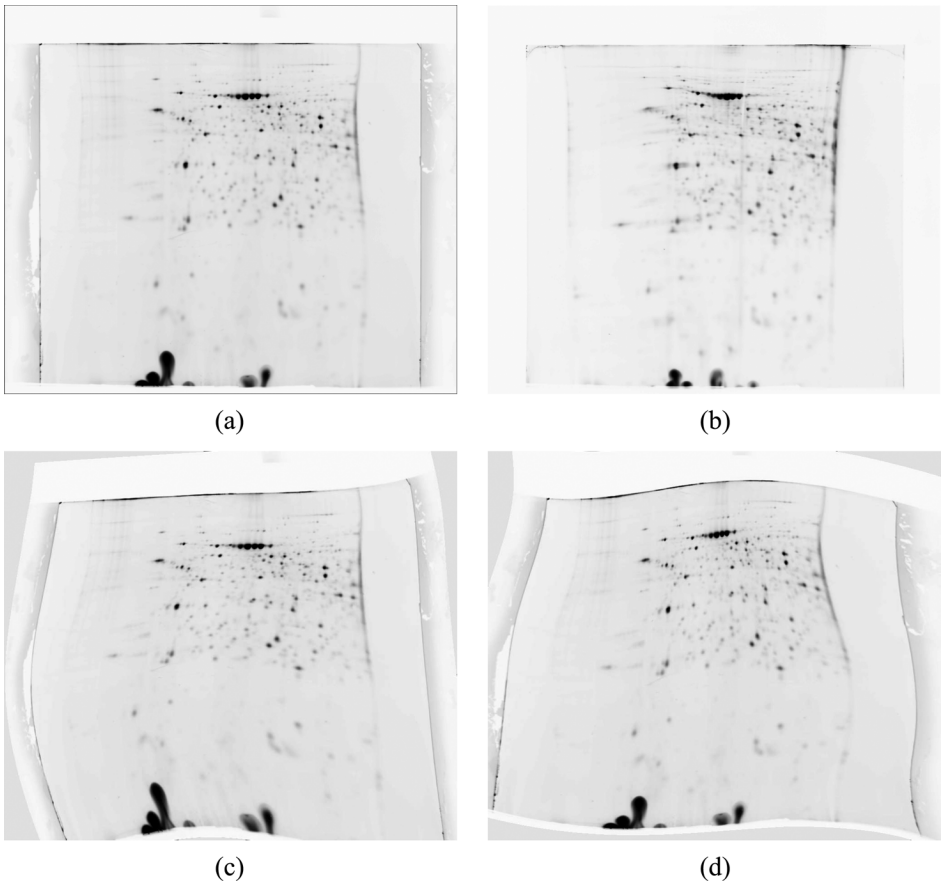


Fig. 1. Examples of 2-DE gel images and synthetic deformation. (a), (b) Real gel images; (c), (d) synthetically deformed versions of (a). (a) Gel 1; (b) Gel 2; (c) medium deformation ($E = 0.275$); (d) large synthetic deformation ($E = 0.5$).

component of the matching algorithm is affected by them. This allows us to specify a point matching scheme which is well tuned to gel matching: an iterative, multiresolution algorithm that uses a novel point distance measure. An important contribution is that we have demonstrated inverse consistency constraints can be introduced into our point matching scheme, ensuring that matching results are independent of the order of the images. We also develop a set of image-based features to enhance our point matching algorithm. The resulting system is compared with, and shown to be more robust and accurate than, a commonly used fully image-based registration algorithm [2], supporting our initial contention.

The rest of this paper is laid out as follows. In the rest of this section, we briefly review existing work on point matching. Descriptions of our methods, evaluations, and results using synthetic point sets and gel images are given in Sections II and III, respectively. Section IV contains a summary and conclusions.

A. Related Work

The registration of point sets is a common problem in computer vision [3]–[7] and a complete review is beyond the scope of this paper. In this work, we focus on the class of algorithms derived from "iterative closest point" (ICP) [3]. ICP is attractive because of its simplicity and good convergence properties; however it is a local refinement technique and requires a reasonable starting point. Granger and Pennec [4] developed an algorithm called EM-ICP which uses Gaussian weighted multiple matches. Their algorithm was developed within a rigid transformation framework. Closely related to this is the *SoftAssign* point matching method [6], [8] in which Gaussian weighted matches are normalized to produce probability estimates for each possible correspondence. An alternative to closest point correspondence estimation is optimal bi-partite graph matching (BGM). Belongie *et al.* [9] used BGM to produce a one-to-one correspondence at each iteration of their algorithm. They also used rich descriptors of the distribution of points called *shape context* as an alternative distance measure to Euclidian. Robust statistical methods have also been applied to address the presence of outliers. Zhang [7] used M-estimators on the residual distance distribution to estimate parameters of a rigid transform. The methods mentioned above do not impose symmetry of the recovered solution. That is, a different result would be obtained if the two point sets were swapped. Johnson and Christensen [5] used inverse consistency constraints within the context of intensity-based image registration to improve registration performance by producing more correct correspondences.

II. Point Matching for Gel Registration

In this section, we evaluate the performance of several candidate point matching algorithms using synthetic data which simulates spot patterns extracted from 2-DE gels. There are often important, genuine, differences in the patterns of spots in gels being aligned. As well as these differences, automatic spot detection algorithms often introduce additional spurious spots

whilst missing or, more commonly, merging overlapping spots. These factors will create *outlier* points in each gel of a comparable pair that have no counterpart in the other. These outliers typically lie within the spatial range of the true spot pattern.

With this in mind, our evaluation focusses on the sensitivity of the algorithms to both *shape deformation* and the presence of unmatchable *outliers*. While basing our investigation on the class of algorithms derived from the ICP method [3], we extend this comparative framework to include algorithms that apply deterministic annealing.

The rest of this section is structured as follows. Section II-A defines the framework of a general point matching problem. Section II-B describes the point-matching algorithms that we have evaluated, separated into components to identify the sources of strengths and weaknesses in each. A description of global deterministic annealing methods that we have also evaluated is given in Section II-C. Section II-D describes the experiments we have carried out, and presents an analysis of the results obtained.

A. Point Matching Definition

The problem of matching two point sets with unknown correspondence can be described as follows. Given two sets of 2-D points, $x = (x_1, \ldots, x_N)^T$, and $y = (y_1, \ldots, y_M)^T$, where $x_i = (x_{i1}, x_{i2})^T$ and y, equivalently, we wish to find a $N \times M$ correspondence matrix, m, and transformation, T, which minimize the following function:

$$E(m,T) = \sum_{i=1}^{N} w_i \left(x_i - T\left(y_i'\right)\right)^T S_i^{-1} \left(x_i - T\left(y_i'\right)\right) + \lambda J_T \tag{1}$$

where

$$y_i' = \frac{m_i y}{\sum_j m_{ij}}, \quad w_i = \sqrt{\sum_j m_{ij}^2} \tag{2}$$

and $S_i = (\sum_j m_{ij}(S_{i,x} + S_{j,y}))/\sum_j m_{ij}$.

Here, m is matrix with values in the range [0,1], representing the confidence of each possible correspondence. The terms y' and w are the mean location and combined weight for a weighted combination of original points y where m_i is the ith row of m. In this way, m is used to produce a weighted one-to-one correspondence between y_i' and x_i. Covariance matrices, $S_{i,x}$ and $S_{j,y}$, represent uncertainty on the position of each point, and are used to produce a Mahalanobis distance metric. The point covariances can be set manually using prior assumptions, or estimated from data as appropriate depending on the specific application. Note that in (1), S_i is a covariance associated with two corresponding points, x and y', and is obtained from a weighted combination of estimated point covariances. The final term J_T represents transformation smoothness and is formalized within the context of spline transforms [10]; its form depends on the transform being used. Throughout this process we have used clamped-plate splines (CPSs) [11] to parameterize these transformations; in this case, the smoothness term is proportional to the integral of the second derivatives of the transform.

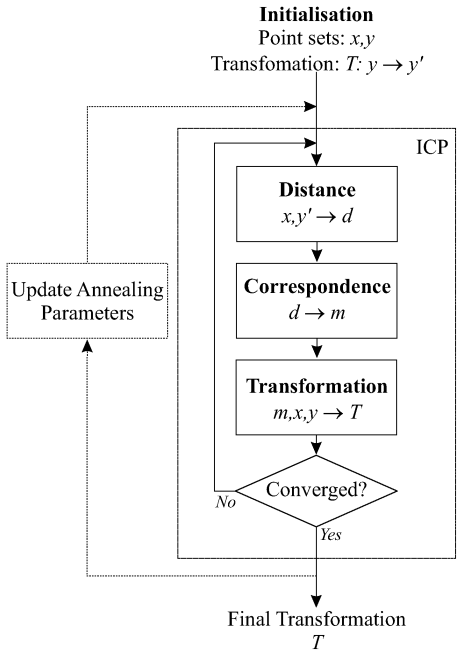


Fig. 2. Point matching algorithm structure. ICP type algorithms can be split into Distance, correspondence and transformation components. Annealing can be used to provide a global-to-local optimisation.

As stated, the correspondence matrix m is used to estimate N points with one-to-one correspondence from the two sets of feature points (2). These can be used to estimate the parameters of a transform between the two point sets and are henceforth referred to as *control points*. With N source and target control points, $x = (x_1, \ldots, x_N)^T$ and $y' = (y_1', \ldots, y_N')^T$, CPS c_s, and affine parameters c_a can be calculated to maximize smoothness whilst interpolating through the control points by solving the following linear system:

$$\begin{bmatrix} K + \lambda W^{-1} & P \\ P^T & O \end{bmatrix} \begin{bmatrix} c_s \\ c_a \end{bmatrix} = \begin{bmatrix} y' \\ o \end{bmatrix} \tag{3}$$

where y' is an $N \times 1$ vector of one component of the spline target points, K is an $N \times N$ matrix of Green's function values ($K_{ij} = G(\|x_i - x_j\|)$), P is an $Nu \times 3$ matrix whose rows are $(1, x_i^T)$, o is a 3×1 vector of zeros and O is a 3×3 matrix of zeros. The matrix W^{-1} is a weighting matrix with a diagonal structure: $W^{-1} = \mathtt{diag}(S_1, S_2, \ldots, S_N)$. For anisotropic control point errors W^{-1} has a block diagonal structure and the form of the solution remains similar to (3) [10], [12].

The tradeoff between smoothness and the accuracy of control point matching is controlled by the parameter λ. In our work, we chose to manipulate λ as a function of $p : \lambda = p/(1-p)$, in which case $0 \leq p < 1$, with $p = 0$ representing an interpolating spline (no smoothing) and $p \to 1$ increases smoothing until only affine parameters are obtained. All transformations have been calculated within the coordinate range: [−0.5,0.5], scaling and centring image coordinates to this region.

No closed form solution to minimizing (1) exists. Therefore, a point-matching algorithm such as ICP must combine (2) and (3) in an iterative framework. Fig. 2 illustrates the iterative algorithm used here. Briefly, the feature point correspondence matrix m is estimated using a distance measure between point

TABLE I
DISTANCE MEASURES WITH EVALUATION PARAMETERS

Abbrev.	Name	Parameters
Euc	Euclidian distance	*none*
SC	Shape Context	Bins: 16 radial $\times 4$ annular, Min Radius: 0.005, Max Radius: 0.25

TABLE II
CORRESPONDENCE ESTIMATION WITH EVALUATION PARAMETERS

Abbrev.	Name	Parameters
CP	Closest Point	*none*
kCP	k-nearest closest points	Static: $k = 4$, Annealing: $4 \geq k \geq 1$
BGM	Bi-partite Graph Matching	*none*
Gauss	Normalised Gaussian weighted multiple matches	Static: $\sigma = 0.025$, Annealing: $0.1 \geq \sigma \geq 0.025$
SA	SoftAssign: Sinkhorn normalised Gaussian weighted multiple matches	Static: $t_s a = 0.0013$, Annealing: $0.02 \geq \sigma \geq 0.0013$, Slack Row\Column initialisation: 1×10^{-12}

TABLE III
TRANSFORMATION PARAMETERS WITH EVALUATION PARAMETERS

Abbrev.	Name	Parameters
CF	Closed form solution to (3)	*none*
Mest	Iterative M-estimation using residual distance distribution	Max Iterations: 30
Mestinv	Iterative M-estimation using residual distance and inverse consistency distribution	Max Iterations: 30, $\alpha = 0.5$, $\beta = 0.5$

sets. Equation (2) is then applied to give control point locations. Equation (3) is then used to give an estimate of transformation parameters. The transformation is applied to the point sets and the distance measure recalculated. This process is repeated until a convergence measure is satisfied. The resulting transformation and correspondence estimates are assumed to minimize equation (1). The following sections describe the structure and components of such an algorithm in more detail.

B. Local Point Matching Methods

Fig. 2 shows the basic structure of our algorithms applied to solving the registration defined by (1). The boxed portion of the figure represents the class of algorithms derived from the ICP scheme. The figure also shows how this basic methodology can be extended to allow global-to-local annealing algorithms, such as robust point matching (RPM) [6], to be implemented by allowing ICP to repeatedly converge at different parameter values. We have explicitly separated the procedure into components: *distance measure*, d_{ij}, and methods of *estimating correspondence*, m_{ij}, and *calculating transformation parameters*, T, each of which can be evaluated in terms of its contribution to the effectiveness of the overall algorithm. To enable us to label the different combinations systematically, we provide in Tables I–III a labelled list of the techniques we have implemented for each component of the algorithm, together with any appropriate adjustable parameters and the values of these that we have used in this study. Each full algorithm will be described in terms of these labels, for example, standard ICP is called Euc-CP-CF: Euclidian distance measure (Euc), closest point correspondence estimation (CP) and closed form transformation estimation (CF); whereas robust M-estimator ICP would be Euc-CP-Mest. We now briefly describe each component option.

Distance: Euclidian (Euc). $d_{ij} = |x_i - y_j|_2$, where $|\ldots|_2$ is the 2-*norm* of a vector.

Distance: Shape Context (SC). SC [9] provides a semi-global description of the spatial distribution of neighboring points by counting the number of points in radial regions, yielding histograms that can be made invariant to affine deformations [Fig. 7(a)]. The method also includes an explicit treatment of outliers. The χ^2 statistic between histograms is used as a distance between features.

Correspondence: Closest Point (CP). CP correspondence produces a binary matrix. For each i, $m_{ij} = 1$ if d_{ij} is the minimum for all j, and 0, otherwise. Therefore, correspondence is assigned between x_i and a single member of y. This procedure does not guarantee one-to-one correspondence.

Correspondence: k-Closest Points (kCP). kCP correspondence also produces a binary matrix. For each i, $m_{ij} = 1$ if d_{ij} is one of the k smallest values for all j, and 0, otherwise. In this case, a single x_i is estimated to correspond to k members of y with equal likelihood. The parameter, k, can be thought of as a circular binary influence, or scale, parameter with a radius determined at each data point as the Euclidian distance of the kth furthest point.

Correspondence: Bi-partite Graph Matching. We have used an efficient implementation of the unit-supply transportation algorithm [13] to calculate an optimal BGM solution. The result is a binary matrix with guaranteed one-to-one correspondence between sets if the same number of points are present in each. In our implementation supply must match demand so when $N \neq M$ extra "slack" rows or columns are added to m, to represent null, or outlier, correspondences. These "slack" variables have a high transportation cost to ensure a maximal graph matching is obtained.

Correspondence: Gaussian Weighted (Gauss). Derived from Bayesian assumptions of uniform Gaussian noise on data points [4], each possible correspondence is weighted by a normalized Gaussian term: $m_{ij} = (\exp(-d_{ij}^2/2\sigma^2)/\sum_k \exp(-d_{ik}^2/2\sigma^2))$. This method is also known as *softmax* [14]. The scale parameter, σ, can be decreased in an annealing schedule, and for each i, the m_{ij} corresponding to the minimum distance, d_{ij}, approaches 1 while all others approach 0. In the limit, $\sigma \to 0$, all m_{ij} will be 0 except the m_{ij} corresponding to the minimum distance. This normalisation has the effect of imposing a single winner-takes-all constraint.

Correspondence: SoftAssign (SA). In the same way as *softmax*, SA uses a Gaussian-weighted correspondence ma-

trix. SA forms a double stochastic matrix using Sinkhorn's method of repeated row and column normalisation [15], which imposes two-way constraints. If $N \neq M$, extra "slack" rows or columns can be added to m, representing null, or outlier, correspondences. These "slack" variables can be included or excluded from the Sinkhorn normalisation process and must be initialized with a distance value. The Gaussian scale parameter controls the "binaryness" of the matrix in a similar way to *softmax*, again allowing an annealing approach. As $t_s a \rightarrow 0$ the matrix m tends towards a binary permutation matrix.

Transformation: Closed-Form (CF). Given the correspondence matrix, m, transformation parameters, c and d, can be calculated by minimizing (1), the analytical solution of which is given by solving (3) using the block diagonal matrix $W^{-1} = \mathtt{diag}(S_i/w_i)$.

Transformation: M-estimator (Mest). The closed-form solution to (3) is only appropriate when data, x and y', contain no outliers, i.e., when m contains only correct correspondences. The closed-form solution is a least-squares model parameter estimation method and, therefore, assumes a Gaussian distribution of residuals. When outliers are present in the data the residual distribution will not be Gaussian, and least squares methods are not appropriate. We have used an M-estimator approach to to reduce the influence of outlying values making the residual distribution conform more closely to a Gaussian model. This is achieved by iteratively re-weighting each correspondence to reduce or remove the influence of correspondences not consistent with our transformation model. The weighting function is based on the residual between the transformed point, $T(y'_i)$ and its target position, x_i : $r'_{ij} = r_{ij} - \bar{r}$, where $r_{ij} = (x_i - T(y_j))^T S_{ij}^{-1}(x_i - T(y_j))$, S_{ij} is the combined covariance of point features x_i and y_j and $\bar{r}$ is a robust estimate of the mean of all r. We have used one of the most consistent and widely used forms of weighting function, which was devised by Huber [16].

Transformation: M-estimator with inverse consistency (Mestinv). A desirable property of any point matching scheme is inverse consistency. That is, transform $T_1 : x \rightarrow y'$, is the inverse of the transform $T_2 : y' \rightarrow x$, i.e., $T_1 = T_2^{-1}$. In general, the correspondence methods described above will not give a consistent result if the point sets are swapped. Also, CPS transforms are not inverse consistent, even with fixed and consistent correspondence. Johnson and Christensen [5], [17] have shown that improved results are obtained when inverse consistency constraints are introduced during image registration. We have incorporated an inverse consistency constraint into our M-estimator residuals as follows: $r_{ij} = \alpha((x_i - T_1(y_j))^T S_{ij}^{-1}(x_i - T_1(y_j))) + \beta((T_2^{-1}(y_j) - T_1(y_j))^T \hat{S}_j^{-1}(T_2^{-1}(y_j) - T_1(y_j)))$, where T_1 is the forward transform, T_2^{-1} is the inverse of the backwards transform. The first term can be thought of as an accuracy constraint, and the second measures inverse consistency at the point location. Parameters α and β control the relative importance of each term. Using this residual in an M-estimation scheme has the effect of reducing the weights of inaccurate and inconsistent correspondences.

C. Global Deterministic Annealing Algorithms

The correspondence estimation methods kCP, Gauss and SA each have some kind of scale parameter. Their performance is dependent on setting a good value for this parameter. For this reason, and to produce a global optimisation, these algorithms can be applied in a deterministic annealing context where their scale is lowered in an annealing schedule. In the context of this work, the scale parameter can be thought of as controlling the extent of a weighted mean for the calculation of y' (1). For example, as the Gauss σ parameter tends towards a large value, all m_{ij} become equal and all y'_i tend towards $\bar{y}$. The resulting transformation will transform every point to the center of gravity of y. As σ decreases, the region of influence of the weighted mean decreases and the method becomes more locally influenced. The settings for maximum and minimum scale parameter value control the annealing schedule. The parameter is initialized at the maximum value, and is decreased until the minimum value is reached. In this way, a coarse-to-fine global search strategy is achieved. Parameter settings for each annealing method are given in Table II.

D. Evaluation and Results

We have performed several experiments to evaluate the performance of different combinations of point-matching components using synthetic point sets designed to simulate protein spot patterns.

Each experiment uses two point sets, referred to as the source and target sets. Each source set consists of 100 points drawn randomly from a uniform distribution within the unit circle. A uniform distribution has been used to simulate protein spot distribution as we do not wish to bias our analysis towards a particular biological sample or gel production conditions. Each target set is copy of the corresponding source set, giving us a ground truth correspondence. Varying degrees of shape deformation can be introduced by transforming the target set using a random Gaussian RBF spline with a fixed transformation energy, E. Fig. 1 illustrates the amount of deformation typically observed between a pair of gels [Fig. 1(a) and (b)] alongside the deformation resulting from a spline transformation with $E = 0.275$ and $E = 0.5$ [Fig. 1(c) and (d)]. Fig. 3 shows that deformations up to $E = 0.5$ were applied in the evaluation experiments, a range considerably larger than that commonly observed between comparable gels.

Outliers have been introduced to either set by removing or adding points in the target at random. During gel registration outliers are likely to occur in both sets simultaneously. For this reason, outliers are generated by both adding a percentage of random points to the target set *and* removing a percentage of the target set.

In all experiments, 100 alignments of random source and target pairs were performed for each method at each deformation energy or outlier level. Unless specified, parameters for each algorithm are given in Tables I–III. These parameters were chosen manually to produce a good alignment performance for each algorithm. Parameter settings for SA and Gauss are equivalent. During this evaluation all covariance matrices are set to $S_i = I_2$ and a CPS smoothing parameter of $p = 0.05$ was used.

Each figure in this section shows residual mean-squared Euclidian distance, r (± 1 s.d. error bars), between ground-truth and corresponding points after alignment. These values are either plotted against deformation energy, E, or against

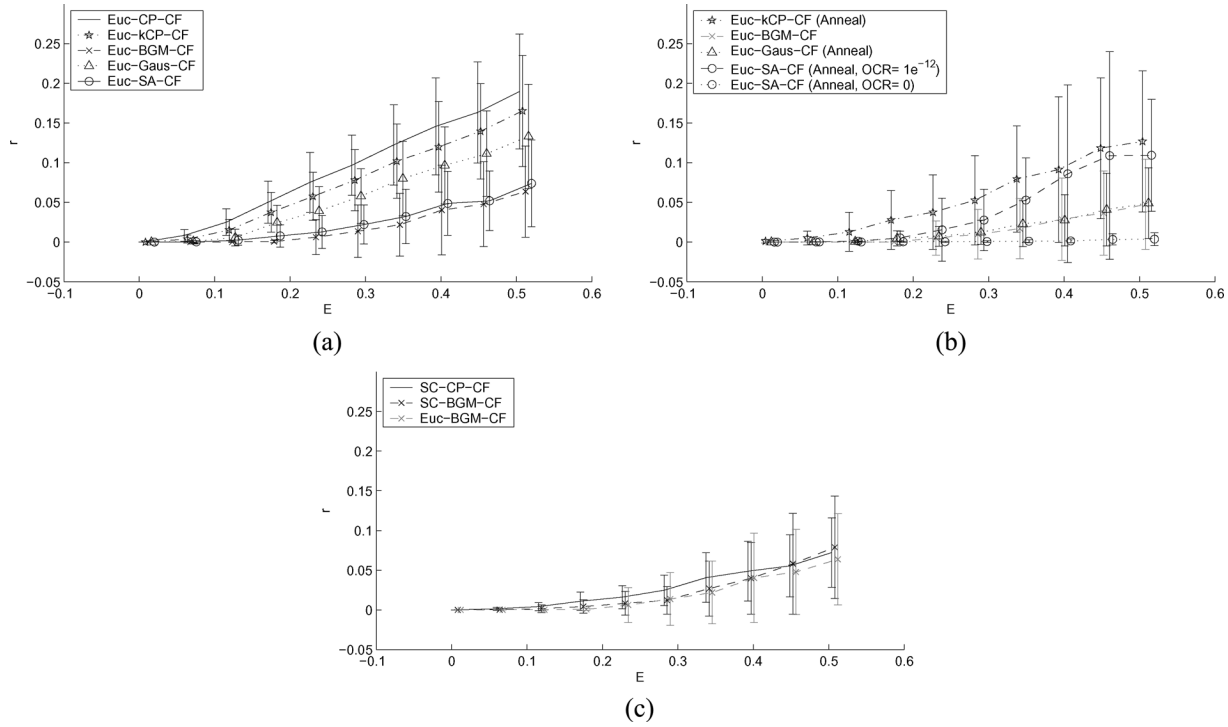


Fig. 3. Effects of deformation on correspondence estimation and distance measure. (a) Without annealing; (b) with annealing; (c) shape context.

percentage outliers added and removed from the target set depending on the experiment. A value of $r = 0$ indicates a perfect recovery of the deforming transformation at the data points.

1) Deformation: Fig. 3(a) shows the effect of increasing deformation on different methods of correspondence estimation with no outliers. CP is the most sensitive to deformation. BGM and SA are the least sensitive to deformation and have similar performance when the SA scale parameter is set to discount all but local correspondences (see Table II). This equivalence demonstrates the fact that both algorithms produce a local optimal solution w.r.t. initialisation. Gauss is less robust to deformation than SA. Parameters for these two algorithms are equivalent, so the improvement observed in SA is the sole result of the dual row and column constraints of Sinkhorn normalisation, as opposed to the single constraint of Gauss.

Fig. 3(b) shows results using annealing methods. BGM results are repeated on the figure for visual comparison [in Fig. 3, Euc-BGM-CF is almost co-incident with Euc-Gauss-CF (Anneal)]. Gauss with annealing has a performance equivalent to BGM. For SA, the initialisation value of the slack row and column has a bearing on performance. Setting this value to 0 $(\mathrm{OCR} = 0)$ has the effect of disallowing outliers, as multiplicative Sinkhorn normalisation has no effect (see Section II-B). This setting produces the best performance of any algorithm evaluated as it matches the data, which has no outliers. When data is expected to contain few, or no, outliers SA without slack row and columns produces optimal results. Our data is expected to contain a large proportion of outliers, so this parameter choice is not appropriate. Choosing a value allowing outlier estimation $(\mathrm{OCR} = 1 \times 10^{-12})$ and propagating the slack row and column values throughout iterations results in a performance less accurate than BGM. This is a more reasonable parameter setting for our data and has been evaluated for data containing outliers, the results of which are presented later in this section. Numerous other possibilities exist for controlling slack row and column values between iterations which we have not investigated here. The two performances of SA presented here represent the extremes of performance that can be achieved with this algorithm by parameter tuning.

Fig. 3(c) shows the effect of deformation using the SC distance measure with CP and BGM correspondence methods. Euc-BGM-CF has been repeated for visual comparison. Euc-CP-CF is the worst performing algorithm in Fig. 3(a), and Euc-BGM-CF the best. However, using SC as a distance measure produces similar performance for both CP and BGM to that obtained by Euc-BGM-CF. Using a more descriptive distance measure, allows a simpler correspondence method to be used without degradation of results.

kCP, Gauss, and SA are not included on this figure. Using SC with any of these non-one-to-one correspondence methods produces unstable results, and the algorithm is not guaranteed to converge. Using Euclidian distance, the scale parameters of these algorithms control the width of a smoothing kernel during calculation of y' (1). However, correspondences with similar SC distances are not guaranteed to lie close to one another in Euclidian space. In this case the scale parameters have little meaning with respect to smoothing estimates of y'. In the form described in Section II-A, none of these algorithms can be used with non-Euclidian distance measures. It may be possible to modify these correspondence techniques to use a non-Euclidian distance measure, but this is beyond the scope of this comparison.

2) Outliers: Fig. 4 shows the effect of outliers on point matching algorithms with varying amounts of deformation.

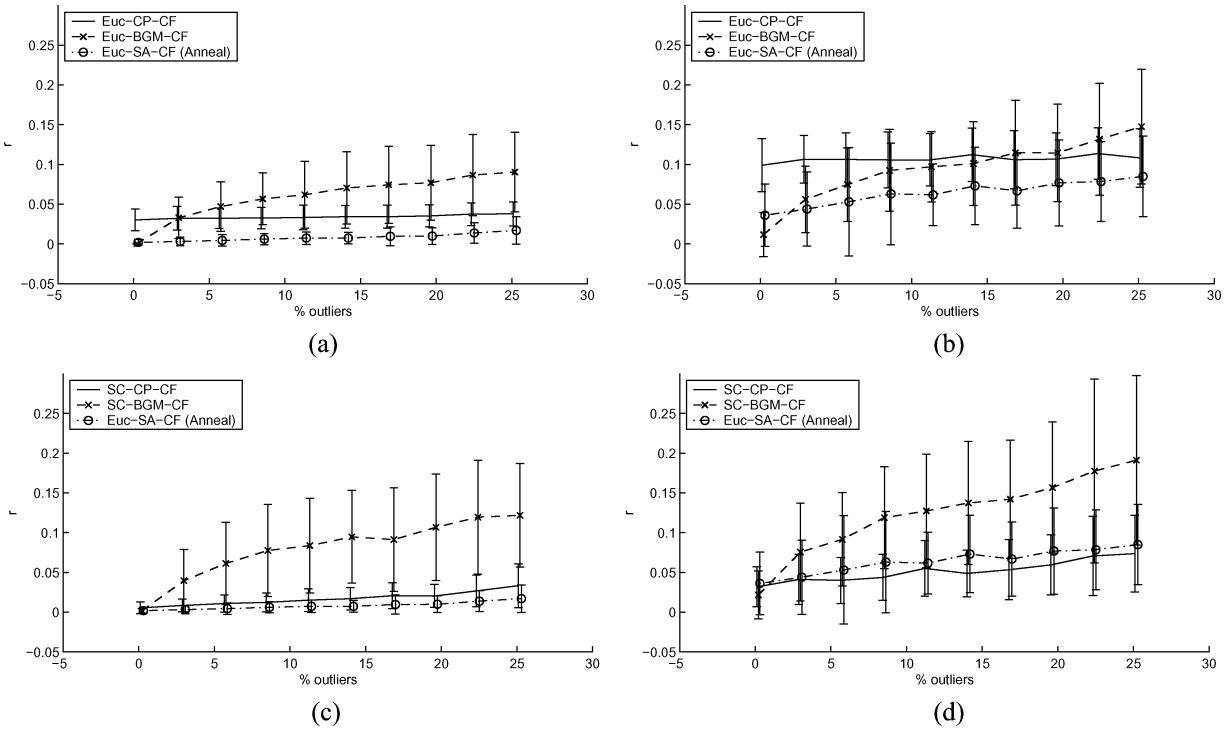


Fig. 4. Effects of outliers on point matching algorithms with varying amounts of deformation using (a), (b) Euclidian and (c), (d) shape context distance measures. (a) Medium deformation, $E = 0.1$, Euc; (b) large deformation, $E = 0.25$, Euc; (c) medium deformation, $E = 0.1$, SC; (d) large deformation, $E = 0.25$, SC.

Point sets were created by *both* removing original points *and* adding extra random points to the target set. A representative set of algorithms from the deformation study have been evaluated. Correspondence methods used were the simple CP method, the BGM method and SA with annealing. Both Euc [Fig. 4(a) and (b)] and SC [Fig. 4(c) and (d)] distance measures were used with each correspondence method.

Previously, we have shown that BGM is sensitive to the proportion of outliers even when no deformation is present when the target set contains both missing *and* extra points [18] (data not shown). Inspection of these results shows "chains" of incorrectly assigned, but relatively consistent correspondences which introduce error into the global transformation. In contrast, where outliers from only additional *or* missing points are present, BGM is able to find correct matches even at quite high outlier proportions (data not shown). However, the presence of "unmatchable" points on both reference and target images corresponds more closely to the real situation in gel matching. We do not believe that this test of BGM has been reported previously. Euc-CP-CF is insensitive to the number of outliers, its performance depending only on deformation [Fig. 4(a) and (b)]. Using the Euclidian distance measure (Euc), SA is the best performing correspondence method in terms of robustness to outliers.

As shown in the case of deformation (Fig. 3), improving the distance measure can allow the use of a less complex method of correspondence estimation. This is again shown in these results. SC cannot be used with SA, but Euc-SA is included as a benchmark in Fig. 4(c) and (d). SC-CP-CF has a comparable sensitivity to outliers as that of Euc-SA-CF (Anneal). The SC distance measure degrades slightly as outliers corrupt the local point patterns. At larger deformations [$E = 0.25$, Fig. 4(d)], SC-CP-CF produces the best alignment of all the algorithms, more accurate than that of the more complex SA correspondence method.

In summary, when deformation is small Euc-CP-CF and Euc-SA-CF are the best performing algorithms and are both insensitive to the percentage of outliers. When deformation is large SC-CP-CF and Euc-SA-CF with annealing are the best performing algorithms in terms of alignment accuracy.

3) M-Estimation: The outlier sensitivity experiments were repeated using M-estimation (Mest) to reduce the influence of outliers on the CP and BGM schemes. Fig. 5 shows results from these experiments in the presence of large amounts of deformation. Results from Euc-SA-CF are repeated on the figure to aid visual comparison. In all results, Mest results in improved alignment over CF for each method. This improvement is greatest for BGM methods, where Mest removes most of the chains of erroneous correspondences.

Unlike SA, Mest separates the measure used to determine correspondence from the measure used to determine consistency with the transformation model. Therefore, we are able to use a non-Euclidian distance measure to determine correspondence more effectively, whilst still imposing consistency constraints in terms of Euclidian residuals. When deformation is large (Fig. 5, $E = 0.25$) both SC-CP-Mest and SC-BGM-Mest produce more accurate alignments than Euc-SA-CF and show no significant increase in residual with respect to increasing numbers of outliers.

4) M-Estimation With Inverse Consistency: Fig. 6 shows the effects of adding an inverse consistency term to the M-estimator residual distribution. Results for Mest are repeated for visual

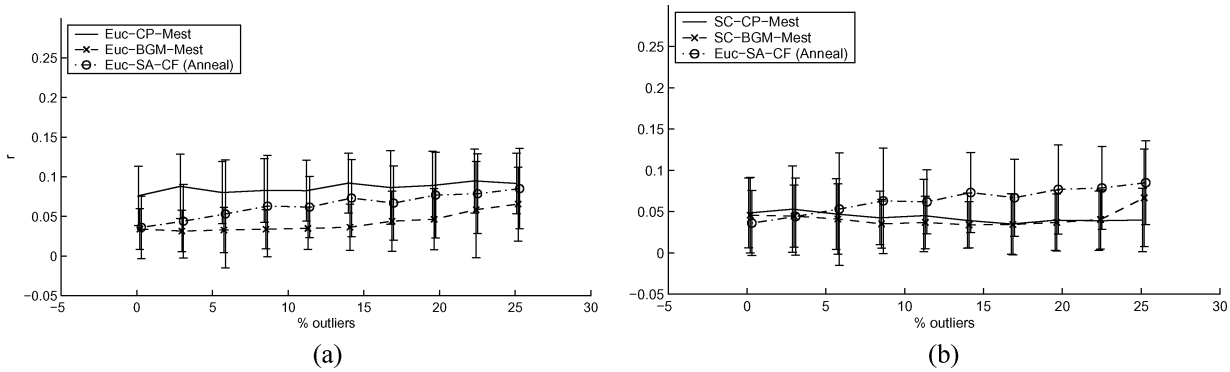


Fig. 5. Effects of outliers on point matching algorithms with varying amounts of deformation using M-estimation with (a) Euclidian and (b) shape context distance metrics to calculate transformation parameters. (a) Large deformation, $E = 0.25$, Euc; (b) large deformation, $E = 0.25$, SC.

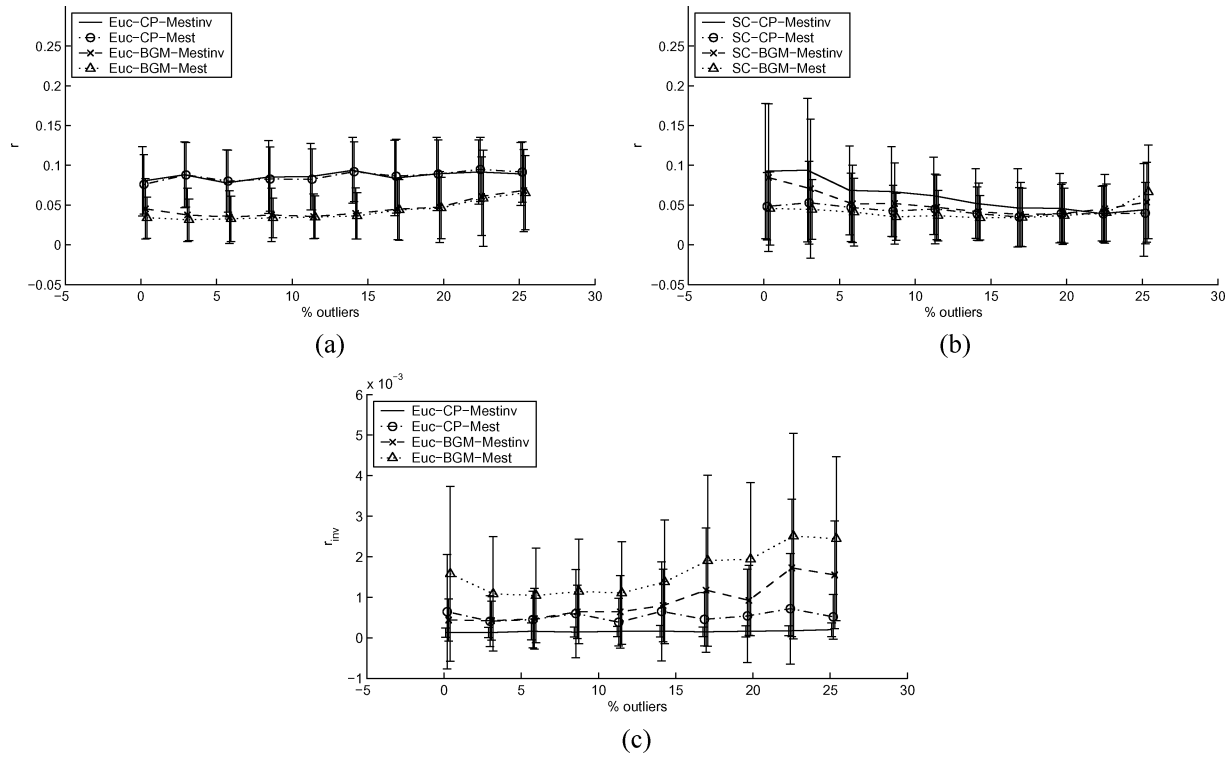


Fig. 6. Effects of additional *and* removed outliers on point matching algorithms using M-estimation with inverse consistency and (a), (c) Euclidian and (b)shape context distance metrics. Large deformation, $E = 0.25$. (a) Accuracy, Euc; (b) Accuracy, SC; (c) Inverse Consistency, Euc.

comparison. Using Mestinv does not improve alignment accuracy using either Euc [Fig. 6(a)] or SC [Fig. 6(b)] distance measures. The accuracy of Mest and Mestinv are equivalent when Euc is used a distance measure regardless of correspondence method. The relationship is less clear when SC is used as a distance measure. When the number of outliers are small, Mestinv degrades accuracy; however, the performance of all algorithms converge at higher outlier levels.

Fig. 6(c) shows a measure of the consistency of forward and reverse transforms, $r_{\text{inv}}(x) = T_1(x) - T_2^{-1}(x)$, plotted against outlier percentage. We have evaluated r_{inv} at a regular grid of 10×10 locations covering the range of the transforms. A value of $r_{\text{inv}} = 0$ indicates perfectly consistent forward and reverse transformations and that swapping the target and source point sets with one another would have no effect on the results of alignment. In all cases, this measure shows that Mestinv results in more inverse consistent transformations than standard Mest (SC data not shown).

These results show that forcing inverse consistency when most correspondences are good, as will be the case using SC when the number of outliers is low [Fig. 6(b)], can erroneously down-weight the number of "correct" matches in favour of incorrect but inverse consistent correspondences. However, when a larger number of outliers are present, the Mestinv technique can produce a much more inversely consistent result, without reducing the accuracy of the final solution, thus ensuring that matching results are independent of the order of the images.

5) Summary: A summary of algorithms best suited for use on data with given characteristics, as determined by our evaluation, is given in Table IV.

TABLE IV
DATA CHARACTERISTICS ALGORITHM SUMMARY

Expected Outliers (%)	Expected Deformation (E)	Algorithm	Notes	Figures
Any ($\sim$ 0 to 25%)	Small ($\sim< 0.1$)	Any CP	When deformation is small, CP correspondence estimation gives the best performance and is invariant to the number of outliers. A more complex algorithm is not required.	4(a)-(d)
Few ($\sim$ 0 to 5%)	Medium ($\sim$ 0.1)	Any BGM	For larger deformation but few outliers Euc-BGM-CF is recommended. Using a more complex distance measure or an iterated annealing method does not improve performance.	3(a)-(c)
Many ($\sim$ 5 to 25%)	Medium ($\sim$ 0.1)	Euc-SA-CF (Anneal)	When a medium amount of deformation is present, SoftAssign with annealing is most invariant w.r.t. larger numbers of outliers.	4(a),4(c)
Many ($\sim$ 5 to 25%)	Large ($\sim> 0.25$)	SC-CP-Mest	When large deformation and the presence of outliers is expected most benefit is gained from a more descriptive distance measure (SC), combined with a robust parameter estimation technique (Mest).	5(b)

III. GEL REGISTRATION

Using the results of our evaluation of point matching algorithms, we have developed a method of gel registration. The entire algorithm is set in a multiresolution framework, with the final transform from the current resolution being used to initialize the next highest. At each level of image resolution, spots' center regions are detected from both gels using a simple threshold on the Laplacian image. We estimate anisotropic covariance for each detected spot center using image partial derivatives [19]. A single point is extracted from each connected region, its position defined to be the location of the darkest pixel. All but the 400 most intense spots are discarded. Increasing the number of spots used to align our gel images does not significantly improve the performance of our algorithm. We have shown that using a descriptive measure of distance between points can produce good alignment results. Using this finding, we estimate correspondence using a distance measure combining Euclidean distance, shape context and two novel measures describing the context of the distribution of spot intensities and sizes. These new, gel-specific, measures are described later in this section. The transformation parameters arising from closest point correspondence with M-estimation are calculated and refined using local image correlation. The resulting transformation is used to initialize the next point matching step. The process is iterated until convergence.

A. Point Matching for Gel Registration

Our point matching scheme uses CP correspondence estimation. We use a combination of Euc and SC distance measures. The evaluation of point matching methods in the presence of outlier features presented in Section II-D shows that when deformation is expected to be large the most appropriate distance measure is SC, and when deformation is small Euclidian distance yields the highest accuracy and robustness (see Table IV for summary). We postulate that using a distance measure tailored to a specific task will produce a better point-matching result than the simple application of a general approach. For this reason, in addition to Euc and SC, we have added two more distance measures designed to represent the local image structure of each gel-spot feature. The first of these is illustrated in Fig. 7(b). Following the SC histogram binning method, we have have developed semi-global image intensity and feature information descriptors. Rather than counting the number of feature points in a specific bin, we have used the average image intensity within the region as an element of an attribute vector. We use the robust least median of squares (LMedS) [20] measure to calculate the distance between vectors. Similarly, we define a third distance measure associated with the binary thresholded Laplacian image used to determine point locations [Fig. 7(c)]. We use the term *image context* (IC) for the first of these two measures, as it contains information about the intensity distribution and *feature context* (FC) for the second, which encapsulates information about the extent of surrounding spot center features. Each of these measures describes a different aspect of a spot's local environment, and they are combined into a single distance between features using the following formula (neglecting normalisation): $d' = \alpha d_{\mathrm{Euc}} + (1-\alpha)(d_{\mathrm{SC}} + d_{\mathrm{IC}} + d_{\mathrm{FC}})/3$, where α is a weighting factor between the two measures, d_{Euc} is Euclidian distance, d_{SC} is shape context χ^2 distance, d_{IC} and d_{FC} are image context and feature context distances calculated using LMedS. All measures are normalized over the set of all distances to have mean 0 and standard deviation 1, which ensures equal influence for each measure when $\alpha = 0.5$. It is expected that the required deformation will be large at low image resolutions, and will become progressively smaller as the gels come into alignment at higher resolutions. For this reason, we vary α linearly between 0 (lowest resolution, entirely context driven) and 1 (highest resolution, entirely euclidian driven) with resolution level during registration. Section III-C gives results of experiments showing the effect of including each of these different distance measures on gel registration accuracy. We also adjust the smoothness of the CPS transform. Starting with a strongly constrained smooth transform at coarse resolutions, the value of the smoothing parameter, p, is decreased at each resolution level, ending with a less constrained transform. In this work, we vary p linearly between 0.25 and 0.01.

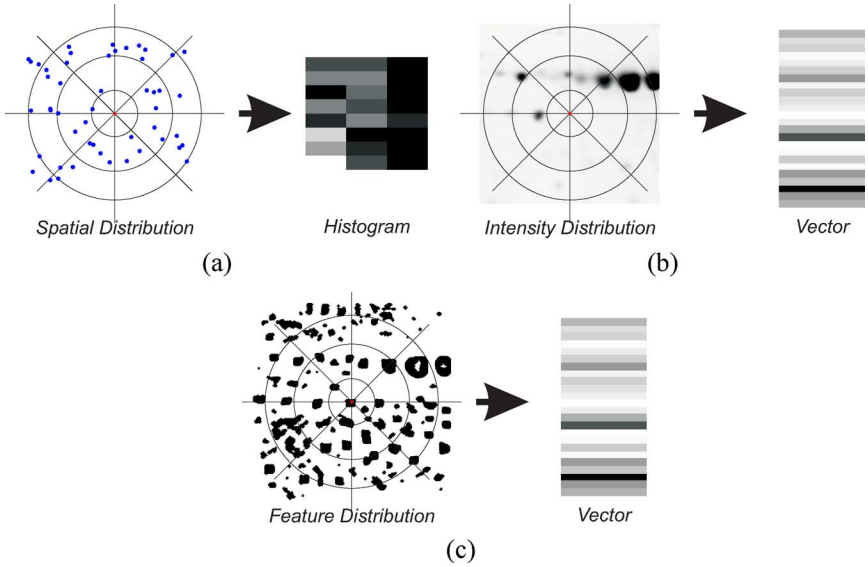


Fig. 7. Attributes for feature distance calculation. (a) Shape Context. Similarity of shape context is measured by constructing a histogram of numbers of points within radially arranged bins. This gives a semi-global indication of the similarity of the shape environments of points being compared. (b) Image context. Similarly to shape context, a measure of similarity of the image environment of comparable points is obtained by averaging the image intensities within radial bins to provide a feature vector. (c) Feature context provides a measure of similarity of the extent of local features, determined by a feature detection step. The elements of the vector in this case are counts of within-feature pixels.

Due to genuine differences in spot pattern and the simplicity of our automatic feature extraction scheme, we know that correspondences will contain errors. We have shown that M-estimation can be used to improve alignment accuracy in the presence of outlier correspondences. Mestinv results in equivalent accuracy for Euc distance with improved inverse consistency. However, alignment accuracy is slightly impaired when using SC as a distance measure. For this reason, we have used Mest as our transform parameter estimation method. Results from Section II-D also suggest that SA could be used as a point matching algorithm, and an evaluation of its performance is given in Section III-C.

B. Local Image-Based Refinement

A further refinement to our scheme addresses the inconsistency of point localisation that arises from using the darkest pixel of regions within the binary feature image. The darkest pixel of corresponding feature areas in two different gels may not be in the same position on each gel. For this reason, we optimize the position of each point in one image with respect to the location of the corresponding point in the other. We maximize the cross-correlation between local image patches by adjusting the location of a point in one of the images. This process is applied, following the determination of correspondence and transformation parameters, on corresponding feature pairs with high weight at each resolution level. The new feature locations are used in subsequent feature matching iterations. For this work, we have used an image region of 15×15 pixels, corresponding to an image region slightly larger than the largest expected protein spot. Results of applying this technique are given in Section III-C.

C. Evaluation and Results

As the analysis of 2-DE gels requires the comparison of corresponding protein spots, the effectiveness of gel matching algorithms should be measured in terms of the accuracy of alignment of protein spots. To perform this evaluation we require a large set of gel image pairs with annotated spot positions and known correspondence. Ideally, the matching difficulty for each pair should be known and should represent the true range found in real data. Data meeting these requirements is not available and, due to the complexity of the images, would be extremely time consuming to produce. Instead, we have used DIGE gel pairs with known spot locations and introduced varying amounts of synthetic deformation to form our test data set. DIGE gels [21] are produced using protein mixtures that are prestained with different fluorescent dyes, chosen to have different U.V. excitation frequencies. This allows pairs of images to be produced with perfect correspondence but showing genuine sample differences.[1]

In this evaluation, we have used five pairs of DIGE gel images, each with ~650 annotated spot positions. Using these images, we created a large evaluation data set by introducing varying amounts of shape deformation to one image using a random Gaussian RBF spline with a fixed transformation energy, E. Fig. 1 shows an example of a synthetically deformed image. In our evaluation, E has been varied linearly in five steps between $0.05 \rightarrow 0.5$. By observation, the top end of this range exceeds the maximum amount of deformation required to align corresponding gels in practice. At each value of E, we have created 5 randomly deformed images from each DIGE pair. This gives a total of $5 \times 5 \times 5 = 125$ gel alignments, each with ~650 spots.

After gel alignment, the recovered transformation is used to transform the spot locations to their estimated position in the un-deformed gel, giving a residual Euclidian distance between the transformed spots and their ground-truth positions

[1]Gel matching is still required to compare between different DIGE gel pairs or when a large number of comparisons is required.

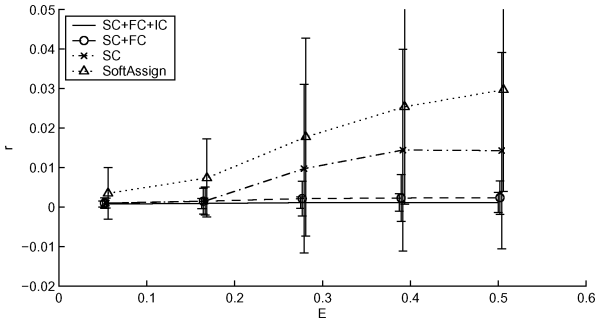


Fig. 8. Effect of distance measure on gel registration error. SC is shape context only, SC + FC is an equal combination of shape context and feature context, and SC + FC + IC is an equal combination of all three distance measures. SoftAssign shows results obtained using SA with Euclidian distance.

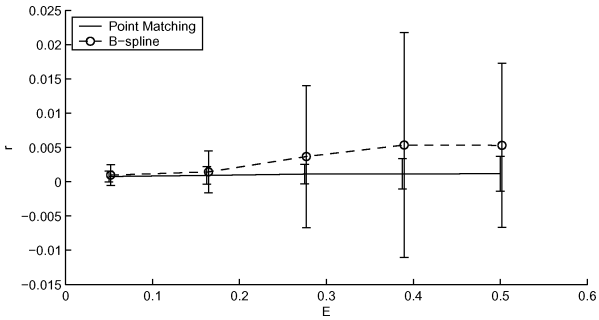


Fig. 9. Comparison of point-based gel registration with image-based gel registration.

(r). Residual r is reported as a proportion of the maximum dimension of the associated gel image. Our algorithms have been applied using three resolution levels on images that are $\sim 850 \times 1000$ pixels in size.

Fig. 8 shows gel registration accuracy using various distance measures. The best results are obtained using a combination of all three distance measures (SC, FC, and IC). We obtain accurate gel registrations which are robust even at the maximum amount of applied shape deformation. SoftAssign, which has been applied with annealing at a single resolution using Euclidian distance, gives poor accuracy and robustness compared to the best gel-specific method.

Fig. 9 shows gel registration accuracy for a commonly used image-based B-spline registration [2]. Point matching results are repeated for visual comparison. Image based registration is significantly less robust than point matching as deformation increases. These results are due to a combination of the local refinement nature of image-based search and the weakness of some image data. When multiresolution methods are applied, weak spots can be effectively removed from the image data at low resolutions. This creates image regions that contain almost no information and their alignment cannot be refined using an image similarity measure such as normalized cross correlation. However, upon increasing resolution, weak, small diameter protein spots may become discernable in these regions. Local search methods will fail if there is not some overlap of the corresponding weak spots from the previous low resolution registration and they will not be brought into alignment. Point matching algorithms have the capability to match these features

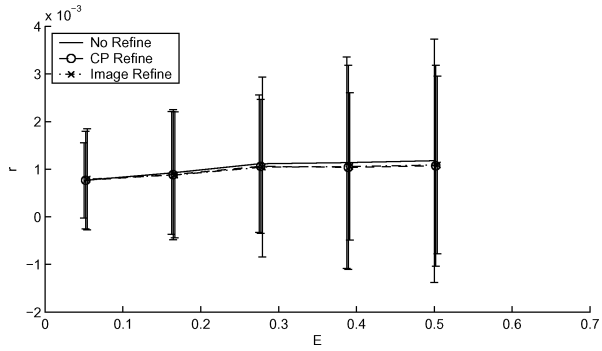


Fig. 10. Effects of local refinement on gel registration accuracy.

even when they are not partially overlapping as they do not explicitly minimize an image-based similarity measure. It may be possible to introduce a semi-global method of image search to improve the performance of image-based registration methods although this has not been done here.

Image-based methods commonly use a regular grid of control points that do not necessarily lie on any particular image features. However, point matching methods benefit from focusing the estimated transformation on features of interest. In our implementation, spline control points lie on protein spot features which are the target of our evaluation measure. This may contribute to the improvement in accuracy gained by our point matching method.

Additionally, the image-based registration method we have evaluated takes no account of outliers in the image intensity residual measure caused not by misregistration, but by genuine spot differences. It may be possible to improve the performance of the method by using a image residual robust to outliers. However, we believe such a method would be unable to distinguish these two types of outlier residuals (based solely on residual intensity) and improvements would be unlikely. Perhaps this could be circumvented by the production of a gel-specific image-difference measure, but this is beyond the scope of this evaluation.

Image-based methods can, however, be used to refine the results of our point matching method. We have evaluated two methods: simple cross-correlation control point optimisation and full B-spline image registration applied after each iteration of point matching. Fig. 10 shows that both methods produce equivalent results, slightly improving on point matching alone. This result demonstrates that image-based registration is unable to significantly improve on a given initialisation using only local search in gel images. Optimizing the location of the spline control points alone gives an equivalent improvement, indicating that little improvement has been achieved away from the control point locations.

IV. Conclusion

We have presented a method of aligning 2-DE gels using point matching that is accurate and robust to large image distortions and large percentages of unmatchable spots. This has been achieved by separately evaluating the components of the point-matching algorithm into distance measure, correspondence estimation and transformation calculation. We have shown that

judicious choice of distance metric and the use of appropriate robust estimation of transform parameters allows a relatively simple algorithmic framework to be used without sacrificing accuracy or robustness of matching. A summary of algorithms best suited for use on data with given characteristics, as determined by our evaluation, is given in Table IV.

The gel-registration algorithm that arose from this evaluation uses a novel combination of Euclidean and shape-context features in an iterative M-estimation algorithm, which corresponds to the assumptions of medium to large deformation with significant numbers of outliers in Table IV. We have shown that it is possible to include inverse-consistency constraints into the M-estimation loop. While improving inverse-consistency of matches, these constraints have no significant effect on matching accuracy, either positively or negatively. However, using these constraints can ensure that matching results are independent of the order of the images. Robustness to both large deformations and the presence of outlier (unmatchable) points in both reference and target images is extremely important in gel-registration. For this image-matching application we have included further image-based features in the distance measure. We have called these *image context* and *feature context*, and they have also been shown to further improve the accuracy and robustness of matching significantly over the point-based scheme, resulting in no loss of matching accuracy in the face of very large distortions. Adding a local refinement step based on image intensities produces a slight, but measurable, improvement in alignment accuracy. Our algorithms have been shown to out-perform image-based registration. The high accuracy and robustness of the system shows promise for use in practical gel alignment situations.

ACKNOWLEDGMENT

The authors would like to thank to R. P. Tonge and AstraZeneca for supplying the 2-DE gel images used in this study, as well as P. Kleinschmidt, Universität Passau, for supplying the BGM code used in this work.

REFERENCES

[1] T. Voss and P. Haberl, "Observations on the reproducability and matching efficiency of two-dimensional electrophoresis gels: Consequences for comprehensive data analysis," *Electrophoresis*, vol. 21, pp. 3345–3350, 2000.

[2] D. Rueckert, M. J. Clarkson, D. L. G. Hill, and D. J. Hawkes, "Non-rigid registration using higher-order mutual information," in *Proc. SPIE Medical Imaging: Image Processing*, K. M. Hanson, Ed., San Diego, CA, Jun. 2000, vol. 3979, pp. 438–447.

[3] P. J. Besl and H. D. McKay, "A method for registration of 3-D shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 239–256, Feb. 1992.

[4] S. Granger and X. Pennec, "Multi-scale EM-ICP: A fast and robust approach for surface registration," in *Proc. Eur. Conf. Computer Vision III*, A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, Eds., Copenhagen, Denmark, 2002, no. 2353, pp. 418–432, ser. Lecture Notes in Computer Science.

[5] H. J. Johnson and G. E. Christensen, "Consistent landmark and intensity-based image registration," *IEEE Trans. Med. Imag.*, vol. 21, no. 5, pp. 450–461, May 2002.

[6] A. Rangarajan, H. Chui, and J. S. Duncan, "Rigid point feature registration using mutual information," *Med. Image Anal.*, vol. 4, pp. 1–17, 1999.

[7] Z. Zhang, "Iterative point matching for registration of free-form curves and surfaces," *Int. J. Comput. Vis.*, vol. 13, no. 2, pp. 119–152, 1994.

[8] H. Chui, A. Rangarajan, J. Zhang, and C. M. Leonard, "Unsupervised learning of an atlas from unlabelled point-sets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 160–173, Feb. 2004.

[9] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape context," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, Apr. 2002.

[10] G. Wahba, *Spline Models for Observational Data*. Philadelphia, PA: SIAM, 1990.

[11] C. Twining, S. Marsland, and C. J. Taylor, "Measuring geodesic distances on the space of bounded diffeomorphisms," in *Proc. 13th Brit. Machine Vision Conf.*, D. Marshall and P. L. Rosin, Eds., Cardiff, U.K., Sep. 2002, vol. 2, pp. 847–856.

[12] K. Rohr, M. Fornefett, and H. S. Stiehl, "Spline-based elastic image registration: Integration of landmark errors and orientation attributes," *Comput. Vis. Image Understand.*, vol. 90, pp. 153–168, 2003.

[13] H. Achatz, P. Kleinschmidt, and K. Paparrizos, "A dual forest algorithm for the assignment problem," *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.*, vol. 4, pp. 1–11, 1991.

[14] J. S. Bridle, "Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters," in *Advances in Neural Information Processing Systems*, D. S. Touretsky, Ed. San Mateo, CA: Morgan Kaufmann, 1990, pp. 211–217.

[15] S. Gold, C.-P. Lu, A. Rangarajan, S. Pappu, and E. Mjolsness, "New algorithms for 2D and 3D point matching: Pose estimation and correspondence," in *Advances in Neural Information Processing Systems*, G. Tesauro, D. S. Touretzky, and T. K. Leen, Eds. Cambridge, MA: MIT Press, 1994, pp. 1019–1031.

[16] P. J. Huber, *Robust Statistics*. New York: Wiley, 1981.

[17] H. J. Johnson and G. E. Christensen, "Landmark and intensity-based, consistent thin-plate spline image registration," in *Proc. 17th Int. Conf. Information Processing in Medical Imaging*, M. F. Insana and R. M. Leahy, Eds., Davis, CA, Jun. 18–22, 2001, vol. 2082, pp. 329–343, ser. Lecture Notes in Computer Science.

[18] M. Rogers, J. Graham, and R. P. Tonge, "2-D electrophoresis gel registration using feature matching," in *Proc. IEEE Int. Symp. Biomedical Imaging*, Arlington, VA, Apr. 2004, pp. 1436–1439.

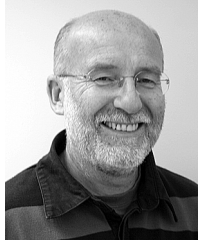
[19] M. J. Brooks, W. Chojnacki, D. Gawley, and A. Van Den Hengel, "What value covariance information in estimating vision parameters?," in *Proc. Int. Conf. Computer Vision*, Vancouver, BC, Canada, 2001, vol. 1, pp. 302–310.

[20] P. J. Rousseeuw, *Robust Regression and Outlier Detection*. New York: Wiley, 1987.

[21] R. Tonge, J. Shaw, B. Middleton, R. Rowlinson, S. Rayner, J. Young, F. Pognan, E. Hawkins, I. Currie, and M. Davison, "Validation and development of flourescence two-dimensional differential gel electrophoresis proteomics technology," *Proteomics*, vol. 1, pp. 377–396, 2001.

Mike Rogers received the B.Sc. degree in artificial intelligence and the M.Sc. degree in cognitive science from The University of Manchester, Manchester, U.K., in 1994 and 1996, respectively, and the Ph.D. degree in computer vision and medical image analysis from the Division of Imaging Science and Biomedical Engineering, University of Manchester, in 2000.

After a postdoctoral period working on the analysis of electrophoresis gels, he is now with Image Metrics, Ltd., Manchester, where he works in the field of computer animation.



Jim Graham (M'85) received the B.Sc. degree in physics from the University of Edinburgh, Edinburgh, U.K., in 1974, and the Ph.D. degree in structural biology from the University of Cambridge, Cambridge, U.K., in 1978.

He joined the University of Manchester, Manchester, U.K., to work on practical applications of image analysis. He is currently a Senior Lecturer in the Division of Imaging Science and Biomedical Engineering, University of Manchester. His research interests are in biological and medical image analysis and computer vision.

36. **A new paradigm for clinical biomarker discovery and screening with mass spectroscopy through biomedical image analysis principles.** H Liao, E. Moschidis, I Riba-Garcia, Y Zhang, R.D. Unwin, J.S. Morris, J. Graham and A.W. Dowsey, *Proceedings of the IEEE International Symposium on Biomedical Imaging Beijing, China. April 2014, , pp 1332-1335.*
doi: 10.1109/ISBI.2014.6868123*.*

A NEW PARADIGM FOR CLINICAL BIOMARKER DISCOVERY AND SCREENING WITH MASS SPECTROMETRY THROUGH BIOMEDICAL IMAGE ANALYSIS PRINCIPLES

Hanqing Liao[1,2], Emmanouil Moschidis[3], Isabel Riba-Garcia[1,2], Yan Zhang[1,2], Richard D. Unwin[2], Jeffrey S. Morris[4], Jim Graham[3], and Andrew W. Dowsey[1,2]

1. Institute of Human Development, University of Manchester, UK
2. Centre for Advanced Discovery and Experimental Therapeutics, University of Manchester and Central Manchester Foundation Trust, Manchester Academic Health Sciences Centre, UK
3. Centre for Imaging Sciences, University of Manchester, UK
4. Department of Biostatistics, UT MD Anderson Cancer Center, Houston, USA

ABSTRACT

Biomarker discovery in amenably sampled body fluids has the potential to empower clinical screening programs for the early detection of disease. Liquid Chromatography interfaced to Mass Spectrometry (LC-MS) has emerged as a central technique for sensitive and automated analysis of proteins and metabolites from these clinical samples. However, the potential of LC-MS as a precise and reliable platform for discovery and screening is dependent on robust, sensitive and specific signal extraction and interpretation. The output of LC-MS is formed as a set of quantifiable images containing thousands of biochemical signals regulated in disease and treatment. We propose to tackle this problem for the first time with a biomedical image analysis paradigm. A novel workflow of image reconstruction, groupwise image registration and Bayesian functional mixed-effects modeling is presented. Poisson counting noise and lognormal biological variation are modeled in the raw image domain, resulting in markedly improved detection limit for differential analysis.

***Index Terms*—**Reconstruction, Image Registration, Functional Mixed Model, Mass Spectrometry, Proteomics

1. INTRODUCTION

The goal of clinical biomarker discovery and screening is to find patterns of proteins and metabolites whose changes in abundance, structure or function robustly discriminate between control and disease. By targeting body fluids (blood, urine or saliva) rather than the pathological tissue, these patterns can be used to directly drive noninvasive routine clinical screening programs as indicators of predisposition, progression, treatment efficacy and early detection of disease before the onset of recognizable symptoms. This screening, and the informatics approach underpinning it, is complementary to and potentially symbiotic with data from macroscopic medical imaging modalities. Whilst body fluids may not be biologically proximal to the disease site, they are often directly affected. However, biomarker discovery has been challenging as human plasma carries 10^5 protein forms, a dynamic range of $\sim 10^{10}$ and significant variation. Systematic workflows have been established (*e.g.* Early Detection Research Network, NCI, USA), but few protein ('*proteomics'*) and metabolite ('*metabolomics*') studies have been successful to date [1].

Since its origins in the late 19th century, mass spectrometry (MS) has become a fundamental method for determining the elemental composition of compounds [2]. Based on this, the life sciences community has adopted MS pervasively since the 1990s, but it is only in the last decade that clinical biomarker discovery and screening has found promise. MS measures the mass-to-charge ratio (*m/z*) of each ionized metabolite or peptide (protein fragment after digestion) contained in a sample mixture, binning each reading to form a histogram '*spectrum'*. While modern instruments achieve very high sensitivity, *m/z* accuracy and resolution, the sample complexity still necessitates pre-fractionation by *retention time* on a liquid chromatography (LC) column. Two-dimensional separation by LC-MS therefore forms an image. Given the format and morphological characteristics of this data, it is perhaps surprising that there has previously been no cross-disciplinary fertilization from the biomedical image analysis field. Fig. 1 illustrates the nature of an LC-MS dataset.

The two cornerstones of computational MS analysis are quantification and identification of the underlying biological signals from the count data. Identification is performed through special acquisition of fragmentation spectra and its pattern matching against empirically curated databases. Identification is not the scope of this paper, except to note that if performed it causes gaps in the LC-MS quantification

This research was facilitated by the Manchester Biomedical Research Centre and Julian Selley, Biological Mass Spectrometry Facility, FLS, University of Manchester. Funding was provided by: BBSRC grant BB/K004158/1 and EPSRC grant EP/E03988X/1 to AWD; and BBSRC grant E14651 and an EPSRC KTA to JG.

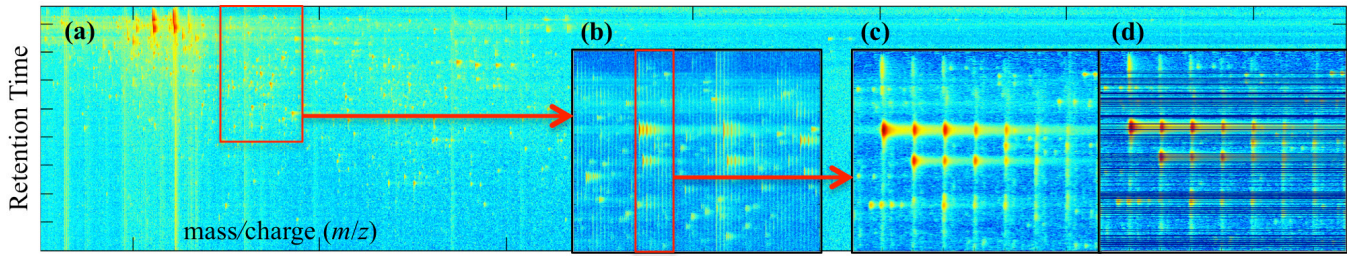


Fig. 1. A typical protein LC-MS dataset, with ion counts in log scale after variance stabilization. **(a)** Illustrates the dynamic range and >100,000 *m/z* datapoints per spectrum. **(b)** Zoomed view shows Poisson noise masking significant dynamic range. **(c)** Zoomed in further, fine details are visible. Each biological signal is exhibited as a series of *isotope* peaks ~1/*z* apart, where *z* is its number of ionic charges. The instrument has a characteristic peak shape, while a periodic background signal is seen. Despite high resolution, overlapping signals are prevalent. **(d)** If identification spectra are acquired in the same run, missing quantification data will be evident (horizontal blue lines).

images, as seen in Fig. 1d. Quantification and difference detection is performed by false-discovery rate controlled statistical testing on corresponding matched peaks across experimental groups of LC-MS datasets, using peak area as a surrogate for biological concentration. With existing tools, the raw data is disregarded after peak detection. While recent peptide modeling methodology [3] has improved sensitivity, substantial detection bias remains as only the strongest signals are reliably detected, quantified and matched between samples. Overlapping signals are either discarded or cannot be quantified accurately. Current algorithms are therefore limited in their ability to derive robust biological data on all detected biochemicals [4].

The proposed workflow consists of three recent image analysis techniques extended to provide a comprehensive biomarker discovery pipeline: (i) Sparse image restoration with a Poisson model for denoising and in-painting of gaps; This stage outputs images for subsequent (ii) smooth deformation non-rigid group-wise registration, which brings strong and weak features into correspondence; (iii) Bayesian multiscale functional mixed-effects modeling to estimate credible intervals and false-discovery rate controlled probabilities for difference detection. The fundamental principle is to retain and model raw image data from start to finish, enabling mining deep below the current detection limit and in complex regions of overlapping signals.

2. METHODS

Conventional processing invariably starts with feature extraction using an implicit Gaussian noise assumption [5]. However, a recent noise analysis [6] has revealed LC-MS data are instead dominated by the effects of the discrete ion event counting process. We therefore perform modeling of each observed LC-MS image $\mathbf{g} = (\mathbf{g}_1,\ldots,\mathbf{g}_n) \in \mathbb{N}$ with the assumption that it represents a sample drawn from a random vector $\mathbf{G} = (\mathbf{G}_1,\ldots,\mathbf{G}_n)$ of n independent Poisson variables:

$$P[G = g \mid \lambda] = \prod_{i=1}^{n} \frac{\lambda_i^{g_i} e^{-\lambda}}{g_i!} \qquad (1)$$

Instead of feature extraction, we preserve all the data for differential analysis by employing *sparse* image restoration by assuming the signals can be represented parsimoniously in an over-complete dictionary. If the L0 pseudo-norm is replaced by a L1 norm, the point estimate still attains sparsity but the minimization becomes convex. Given an appropriate Lagrange multiplier λ, the solution $\mathbf{x}$ with input image $\mathbf{b}$, dictionary $\mathbf{A}$ and Gaussian noise assumption is:

$$\min_{\mathbf{x}} \quad \frac{1}{2}\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1 \qquad (2)$$

For Poisson noise, equation (2) can be utilized by first approximately stabilizing the variance with an *Anscombe* transform, $\mathbf{b} = 2\sqrt{(\mathbf{b}_0+3/8)}$. Nevertheless, the point estimate under Poisson noise is naturally sparsifying due to the distribution's heavy tail. This fact was recently exploited by Shaked *et al.* [7] to provide a sparse version of the seminal Richardson-Lucy iteration for exact Poisson noise handling:

$$\mathbf{x}_{i+1} = \mathbf{A}^* \left\{\frac{\mathbf{b}}{\mathbf{A}\mathbf{x}_i}\right\} \frac{\mathbf{x}_i}{\mathbf{A}^*\{\mathbf{1}\}+\lambda} \qquad (3)$$

Since a positive valued dictionary is necessary to ensure positivity of the restored image, we employ a complete set of separable 2D multiscale cubic B-spline basis functions as our sparse domain. This models signal structure and enables multiscale in-painting. Because this fixed-point iteration is slow to converge, it was necessary to accelerate the method with Biggs-Andrews vector-extrapolation [8]. To correct the shrinkage bias, after execution with $\lambda > 0$, we rerun the method with $\lambda = 0$ on the remaining non-zero coefficients.

The next step is to bring corresponding biological signals across the images into correspondence by accounting for retention time inhomogeneity. Alignment of equivalent spectra among (potentially large) groups of patient samples can be set as a non-rigid registration problem similar to those encountered in medical imaging. An important difference in this case is that the dimensions of the image (LC and MS) represent different physical quantities on different scales. The displacements are much greater in the LC dimension and variable throughout the image. Feature-based registration is appropriate. However, the optimization is hampered by difficulties: large numbers of features with a very wide dynamic range of intensities, large numbers of similar spectra resulting in multiple local minima for registration, non-rigid retention time deformations and significant numbers of unmatchable features.

Rogers and Graham [9] proposed a point-based registration algorithm in the context of alignment of 2D electrophoresis gels, which share some of these difficulties. Their algorithm was based on robust point-matching using M-estimation and a non-Euclidean distance metric in a multi-resolution framework. Robustness to reference image selection was achieved by registration to an evolving mean. We have adapted this method to deal with the much higher and heterogeneous spatial resolution of the LC-MS images.

In the final stage, to discover statistically significant biological regulation characteristic of potential biomarkers we have adopted a Bayesian Markov-Chain Monte Carlo (MCMC) method [4] for linear mixed-effects modeling on the restored, registered images. In large-scale biomarker discovery, a number of confounding systematic biases ('fixed effects') will be evident, such as the blocking of runs over different days. Additionally, multiple sources of statistical variation ('random effects') will be intermixed, such as when analyzing longitudinal samples from a subject against other subjects. Techniques that consider linear fixed and random effects are termed *linear mixed models*. These can model regression relationships between outcomes and a set of multiple predictors, while accounting for potential correlation among the observations that might be induced by the experimental design. Morris and Carroll introduced an approach extending the linear mixed model to analyze high-dimensional complex functional data, termed the *wavelet functional mixed model* (WFMM). This work was extended to handle image data and other higher dimensional functions in Morris, et al. [4]. For a set of 2D images $Y_i(t_1,t_2)$, $i=1, \ldots, N$, predictors X_{ia}, $a=1 \ldots p$, and random effect predictors Z_{ib}, $b=1 \ldots m$, the functional mixed model is given by:

$$Y_i(t_1,t_2) = \sum_{a=1}^{p} X_{ia} B_a(t_1,t_2) + \sum_{b=1}^{m} Z_{ib} U_b(t_1,t_2) + E_i(t_1,t_2), \quad (4)$$

Where $B_a(t_1,t_2)$ is a *fixed effect function* that measures the effect of predictor X_{ia} on the image $Y_i(t_1,t_2)$ at position (t_1,t_2). The random effects $U_b(t_1,t_2)$ and residual errors $E_i(t_1,t_2)$ are assumed to be mean-zero Gaussian random variables.

This approach models differences between the images without performing feature extraction, so has the potential to find results that would have been missed by peak detection failing to discern true peaks from noise [4]. The technique has been applied to low-resolution spectra and 2D electrophoresis gels. As LC-MS datasets are significantly larger, images were partitioned and cluster computing adopted. Data is Anscombe and lognormal transformed to stabilize biological variation across samples. The model is sampled with MCMC after each input image is wavelet transformed (2D Daubechies wavelets with 4 vanishing moments). The wavelet domain random variables allow heteroscedasticity both spatially and over scales, whilst the fixed effects use an adaptive spike-slab prior to promote peak-like signals. After applying inverse wavelet transforms to the MCMC samples of this wavelet space model, the result is a full posterior image distribution for each effect.

3. RESULTS

In the controlled validation study, 12 proteins were digested, resulting in thousands of peptide features across each LC-MS image. 4 groups of 3 images were acquired, with 4 proteins held at the same concentration in all groups, while 8 were varied by known but approximate amounts. The experiment was repeated with co-acquisition of identification spectra. The total set of 24 images took 24 hours to acquire on an Agilent 6530 Q-TOF. Datasets were then normalized to a peptide internal standard. As an authentic experiment, this validation data reflects the physical characteristics of production clinical studies.

The datasets were processed with our workflow and the leading methodology of Progenesis LC-MS (Nonlinear Dynamics, UK). For our workflow, each LC-MS run was first processed with the image restoration method. Results are illustrated in Fig. 2a-d for 3 values of the shrinkage parameter λ. The strategy was to provide sufficient L1 regularization to draw out faint biological signals and for robust in-painting. A conservative shrinkage $\lambda = \sqrt{2}$ was found to be suitable to avoid decimating real signal. The same value for λ gave equivalent results in all image regions. Conversely, feature extraction methods based on a Gaussian noise assumption requires a varying threshold across *m/z* due to reduced ion counts as m/z increases (as can be seen in Fig. 1a). This adds further evidence that Poisson noise is the dominant variation in LC-MS [6].

After restoration, the 24-image dataset was groupwise registered, as illustrated in Fig. 2e-f. All images were then manually examined for misregistration. The registration method identified one identification image as an outlier, which was confirmed to be the case and removed from the study. Otherwise, registration was successful between quantification and in-painted identification images, allowing identifications to be propagated to the quantification data.

The registered quantification images were then input into WFMM. The design matrix consisted of three fixed effects representing the log ratio (*i.e.* up-fold or down-fold

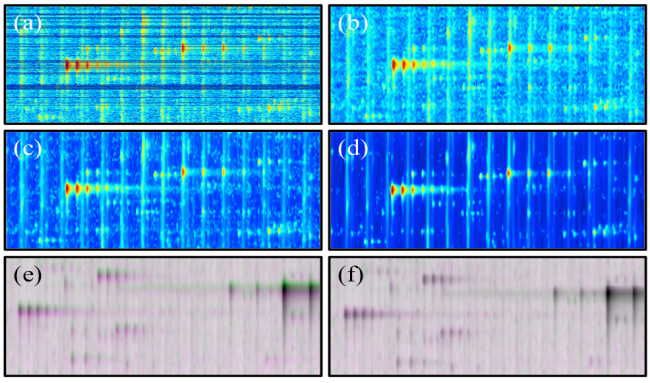


Fig. 2. (a) Raw count data for a 10 *m/z* region. **(b-d)** Denoising and in-painting results with a shrinkage factor of $2^{0.0}$ (b), $2^{0.5}$ (c) and $2^{1.0}$ (d). **(e-f)** A restored region before (e) and after (f) registration, demonstrating alignment of a quantification dataset (magenta) to an in-painted identification dataset (green).

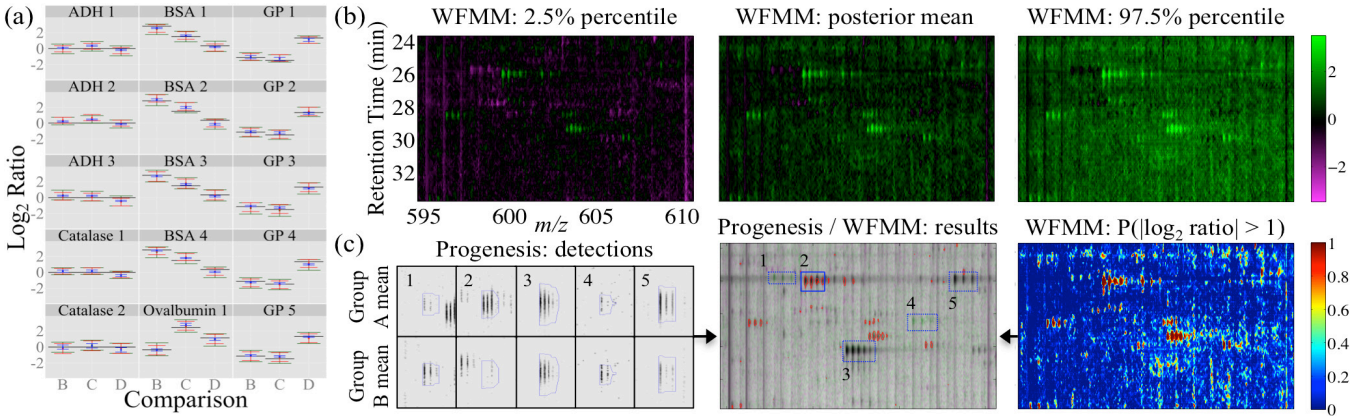


Fig. 3. (a) Mean and 95% credible interval for the ratios between group A and B,C,D for 15 curated peptides. Results are shown for Progenesis (blue), and WFMM with restored (red) and raw (green) input. Black lines denote approximate ground truth. **(b)** WFMM mean and 95% credible interval for the image fixed effect between group A and B. **(c)** Right: For (b), posterior probability of $\log_2$ ratio > 1 (*i.e.* more than a doubling or halving of protein). Middle: From this probability image, contours (red) signifying a false discovery rate < 1% overlaid on mean images of group A (green) and B (magenta). The 5 peptides detected by Progenesis are labelled/boxed, with the single significantly regulated peptide shown in solid blue. Left: Progenesis background subtraction and segmented boundaries for these peptides.

regulation between groups) between the three images in group A and the three images in groups B, C and D.

For objective validation, we examined the derived fold changes for 15 features. Since these were necessarily identified and validated manually, we were limited to intense, isolated features. For suitable comparison, extracted features from Progenesis were modeled with MCMCglmm R package for Bayesian mixed modeling. As illustrated in Fig. 3a, mean fold change estimates were consistent between Progenesis and our workflow, whereas credible intervals were narrower with Progenesis. We postulate that this is due to additional biological knowledge utilized by Progenesis for background subtraction and peptide modeling. Nevertheless, intense features are frequently housekeeping proteins, whereas interesting biomarkers are often expressed only in trace amounts. As shown in Fig. 3b-c, WFMM detects numerous significant changes in regions that Progenesis removes as background. Moreover, regions flagged by WFMM form peak trains characteristic of peptides. Since we do not model this distinctive signal structure, these patterns are unlikely to be false positives.

4. CONCLUSIONS

We have presented a new type of workflow for biomarker discovery and screening in LC-MS that is based upon principles that underpin biomedical image analysis methodology. The raw data is retained and utilized from beginning to end, with differential quantification founded on the proven mixed-effects model, thus enabling the handling of complex experimental designs. Unlike existing methods, it is not reliant on the performance of prior background subtraction and feature extraction routines, and therefore is capable of finding significant changes below current software detection limits. These newly discovered changes are candidate biomarkers for subsequent targeted validation.

Since no prior biological knowledge is utilized, our workflow can be applied generally to all types of proteomics or metabolomics LC-MS experiment. Nevertheless, biological signal structure and correlation could also be modeled directly within our framework to improve sensitivity, which is a direction for future work. As well as comprehensive validation, we are also particularly interested in adapting our workflow to emerging *MS Imaging* technology, where mass spectra are acquired spatially across tissue sections for powerful 'virtual' histology. Moreover, the Bayesian functional mixed model also has a wide range of application in difference detection for other types of medical imaging data, including fMRI, EEG, and DTI.

5. REFERENCES

[1] S. Srivastava, "The early detection research network: 10-year outlook," *Clin. Chem.*, vol. 59, no. 1, pp. 60–67, 2013.

[2] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature*, vol. 422, no. 6928, pp. 198–207, 2003.

[3] B. Renard, M. Kirchner, H. Steen, J. Steen, et al., "NITPICK: peak identification for mass spectrometry data," *BMC Bioinformatics*, vol. 9, no. 1, p. 355, 2008.

[4] J. S. Morris, "Statistical methods for proteomic biomarker discovery based on feature extraction or functional modeling approaches," *Stat. Interface*, vol. 5, no. 1, pp. 117–136, 2012.

[5] A. W. Dowsey, J. A. English, F. Lisacek, J. S. Morris, et al., "Image analysis tools and emerging algorithms for expression proteomics," *Proteomics*, vol. 10, no. 23, pp. 4226–4257, 2010.

[6] P. Du, G. Stolovitzky, P. Horvatovich, R. Bischoff, J. Lim, et al., "A noise model for mass spectrometry based proteomics," *Bioinformatics*, vol. 24, no. 8, pp. 1070–1077, 2008.

[7] E. Shaked, S. Dolui, and O. Michailovich, "Regularized Richardson-Lucy algorithm for reconstruction of Poissonian medical images," in *Proc. ISBI: From Nano to Macro*, 2011.

[8] D. S. Biggs and M. Andrews, "Acceleration of Iterative Image Restoration Algorithms," *Appl. Opt.*, vol. 36, no. 8, 1997.

[9] M. Rogers and J. Graham, "Robust and Accurate Registration of 2-D Electrophoresis Gels Using Point-Matching," *IEEE Trans. Image Process.*, vol. 16, no. 3, pp. 624–635, 2007.

Applications of Image Analysis: Assessing Bone Quality

37. **Detecting reduced bone mineral density from dental panoramic radiographs using statistical shape models.** P.D. Allen, J. Graham, D.J.J. Farnell, E. Harrison, R. Jacobs, K. Karayianni, C. Lindh, P.F. van der Stelt, K. Horner and H. Devlin, *IEEE Transactions on Information Technology in Biomedicine* ***11****(6): 601-610, 2007.* Doi: 10.1109/TITB.2006.888704

Detecting Reduced Bone Mineral Density From Dental Radiographs Using Statistical Shape Models

P. Danny Allen, Jim Graham, *Member, IEEE*, Damian J. J. Farnell, Elizabeth J. Harrison, Reinhilde Jacobs, Kety Nicopolou-Karayianni, Christina Lindh, Paul F. van der Stelt, Keith Horner, and Hugh Devlin

***Abstract*—We describe a novel method of estimating reduced bone mineral density (BMD) from dental panoramic tomograms (DPTs), which show the entire mandible. Careful expert width measurement of the inferior mandibular cortex has been shown to be predictive of BMD in hip and spine osteopenia and osteoporosis. We have implemented a method of automatic measurement of the width by active shape model search, using as training data 132 DPTs of female subjects whose BMD has been established by dual-energy X-ray absorptiometry. We demonstrate that widths measured after fully automatic search are significantly correlated with BMD, and exhibit less variability than manual measurements made by different experts. The correlation is highest towards the lateral region of the mandible, in a position different from that previously employed for manual width measurement. An receiver-operator characterstic (ROC) analysis for identifying osteopenia ($T < -1$: BMD more than one standard deviation below that of young healthy females) gives an area under curve (AUC) value of 0.64. Using a minimal interaction to initiate active shape model (ASM) search, the measurement can be made at the optimum region of the mandible, resulting in an AUC value of 0.71. Using an independent test set, AUC for detection of osteoporosis ($T < -2.5$) is 0.81.**

***Index Terms*—Active shape model (ASM), bone mineral density (BMD), dental panoramic tomogram (DPT), inferior mandibular cortex (IMC), osteopenia, osteoporosis, segmentation.**

I. Introduction

OSTEOPOROSIS is a general loss of bone mineral density and can lead to an increased risk of fracture. Based on factors such as previous fracture, family history, and height loss. Patients deemed to be at risk are referred for bone mineral density (BMD) assessment using dual-energy X-ray absorptiometry (DXA). However, there has recently been great interest among dental researchers in the possibility of identifying those at risk of reduced BMD from dental radiographs since mandibular BMD is related to systemic BMD [1].

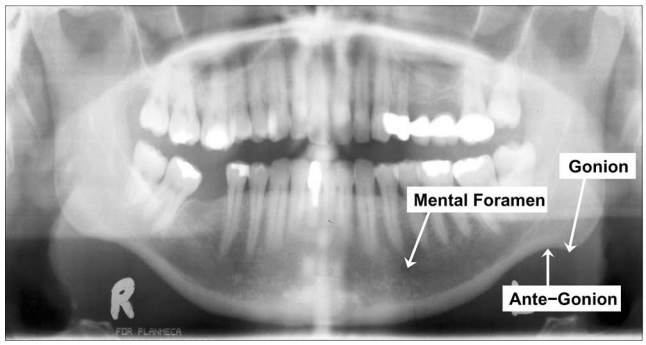


Fig. 1. Example of a DPT of a normal (nonosteoporotic) patient. The positions of anatomical points key to manual annotation are shown.

Fig. 2. A portion of the dental tomogram shown in Fig. 1 showing the appearance of the mandibular cortex in a normal (nonosteoporotic) patient.

Fig. 1 shows an example of a dental panoramic tomogram (DPT) of a normal patient and Fig. 2 shows a close up of the right mandible. Fig. 3 shows a schematic of Fig. 2—the cortical region in this diagram is referred to as the inferior mandibular cortex (IMC). There is evidence that the thickness of this cortex is correlated with systemic BMD and hence causes osteoporosis [2]. Fig. 4 shows the equivalent view of Fig. 2 for a patient with osteoporosis—the mandibular cortex is much harder to perceive visually as it is both thinner and less distinct from the mandible as a whole. In particular, the thickness of the mandibular cortex at a point closest to the mental foramen, referred to as the mental index (MI), (Fig. 3) has been found to be the best indicator of low BMD compared to the equivalent indices at the gonion (GI) and the antegonion (AI) (Fig. 1) [2].

There is considerable room for subjectivity in the precise placement of the MI measurement; the mental foramen is a very indistinct feature, and the endosteal border can become very

Manuscript received March 17, 2006; revised October 31, 2006. This work was supported by a Research and Technological Development project grant from the European Commission Fifth Framework Programme "Quality of Life and Management of Living Resources" under grant (QLK6-2002-02243).

P. D. Allen, J. Graham, and E. J. Harrison are with the Division of Imaging Science and Biomedical Engineering, University of Manchester, Manchester M13 9PL, U.K. (e-mail: philip.allen@manchester.ac.uk).

D. J. J. Farnell was with the University of Liverpool, Liverpool L693BX, U.K. He is now with the School of Medicine, University of Manchester, Manchester M15 6FH, U.K. (e-mail: d_j_j_farnell@yahoo.co.uk).

R. Jacobs is with the Oral Imaging Centre, Katholieke Universiteit Leuven, Leuven B-3000, Belgium (e-mail: reinhilde.jacobs@uz.kuleuven.ac.be).

K. Nicopolou-Karayianni is with the Dental School, University of Athens, Athens 15784, Greece (e-mail: ketykara@otenet.gr).

C. Lindh is with the Faculty of Odontology, Malmö University, Malmö SE-20506, Sweden (e-mail: Christina.Lindh@od.mah.se).

P. F. van der Stelt is with the Academic Centre for Dentistry, Amsterdam 1066 EA, The Netherlands. (e-mail: P.vd.Stelt@acta.nl).

K. Horner and H. Devlin are with the School of Dentistry, University of Manchester, Manchester M15 6FH, U.K. (e-mail: hugh.devlin@manchester.ac.uk).

Digital Object Identifier 10.1109/TITB.2006.888704

1089-7771/$25.00 © 2007 IEEE

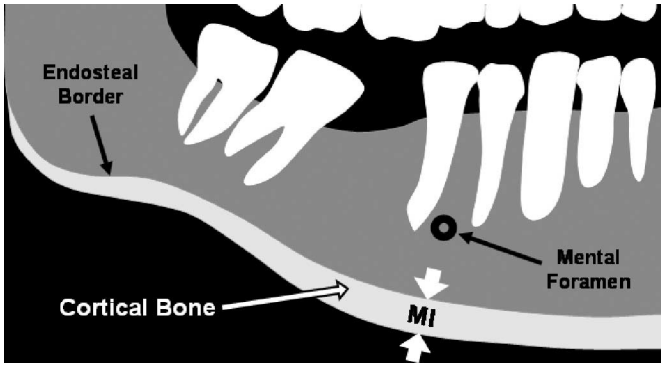


Fig. 3. Schematic diagram of the dental tomogram shown in Fig. 2, showing the point at which the inferior mandibular cortex thickness is measured by dentists (mandibular index MI).

Fig. 4. Portion of a DPT showing the same region of the mandible as in Fig. 2, but for a patient with osteoporosis. Note the thinning of the inferior mandibular cortex compared with Fig. 2.

indistinct in cases of osteoporosis (Fig. 4). These factors do not pose significant problems for an expert radiologist. However, for general dental practitioners (GDP), they lead to considerable variability in MI measurement, even with individual training, and so routine assessment of low BMD risk from dental radiographs by GDPs is not practical [4].

Dentists use a large number of radiographs, accounting for 32% of all medical radiological examinations in the U.K. [5], opening the possibility of obtaining valuable medical information about patients' osteoporotic status from a routine radiological examination. Here, we describe an automatic method of measuring radiographic indices using computer image analysis that is sensitive to mandibular BMD, and hence, systemic osteoporosis.

Our approach is to use an active shape model (ASM) method [7] to locate the upper and lower borders of the inferior mandibular cortex, and hence measure its thickness.

II. DATA

The patient data set had been collected for a previous study [3] and consisted of 132 consecutive female patients aged between 45–55 who attended the University Dental Hospital of Manchester for routine dental treatment.

A. Radiographic Examination

All of the patients received a radiological examination of the mandible using a DPT. All radiographs were performed using either a Cranex DC-3 unit (Soredex Orion Corporation, Finland) or a Planmeca PM 2002C unit (Planmeca, Finland) using the same film/cassette combination. The films were digitized using a Kodak LS85 digitizer (Eastman Kodak, Rochester, NY) at a resolution of 25.64 pixels/mm.

B. BMD Assessment

One hundred and twenty six females had central DXA of the proximal femur and lumbar spine on the GE Lunar DPX-L (GE Lunar Corporation, Madison, Wisconsin). BMD measurements at each site were compared to the manufacturers reference data to give a T-score value, which is the number of standard deviations the BMD measure lies from the sex-matched young adult mean value. Using the World Health Organisation criteria, patients are defined as osteopenic if their T-score value is between -1 and -2.5 and osteoporotic if their T-score value is less than -2.5. In this study, patients were categorized by the lowest T-score value at either the total hip or lumbar spine (L1-L4). Of the 126 patients with BMD measurements, 79 were normal, 42 were osteopenic, and five were osteoporotic.

III. ASM Method

A. Point Distribution Model (PDM)

The ASM method has been extensively documented already elsewhere [7]–[9], and only a brief description will be given here. At its core is a PDM that describes the principal modes of variation of a set of landmark points used to describe the object of interest. The model is "trained" using points placed on a training set of example shapes, usually manually (see Section III-B), at anatomically consistent locations around the border of the object (see for example, Fig. 1). The points are concatenated into a single shape vector $\mathbf{x} = (x_1, y_1, x_2, y_2, \ldots, x_n, y_n)$ for each of the training examples, after alignment to a common coordinate frame, where n is the number of points. New example shapes can be generated from a principal component analysis of the covariance matrix generated from the shape vectors, thus

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{Pb} \tag{1}$$

where $\bar{\mathbf{x}}$ is the mean shape. $\mathbf{P}$ is the $n \times t$ matrix of the t most significant eigenvectors $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_t$ of the covariance matrix, and $\mathbf{b}$ is a vector of parameters $b_1, b_2, \ldots b_t$ describing the weights assigned to each eigenvector to describe a particular shape. Equation (1) shows that $\mathbf{b}$ is equivalent to $\mathbf{x}$ as a shape description. Each eigenvector $\mathbf{P}_i$ corresponds to a "mode of variation" in the observed shapes in the training set. Varying the values of $b_1, b_2, \ldots, b_t$ allows us to generate shapes within the observed range. For example Fig. 5 shows the effect of varying b_1 by $\pm 3\sigma$ around its mean value, while keeping $b_2, \ldots, b_t$ at their mean values. This mode of variation principally represents the range from narrow to broad mandible shape. Other modes represent different characteristics of the observed shape.

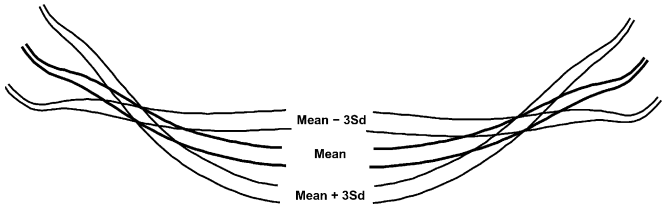


Fig. 5. Effect of varying the first mode of variation by ±3sd on the PDM of the mandible.

B. Manual Annotation

To build a PDM of the IMC of the mandible, the set of training examples have to be manually annotated. This was done by two experts using a custom-written graphical user interface. The positions of the upper and lower border of the mandibular cortex were marked at two key anatomical landmarks: the points closest to the mental foramen, and the points at the ante-gonion (see Fig. 1), henceforth, referred to as the MF and AG points. Between these a number of equally spaced points were placed to define the upper and lower borders of the mandibular cortex.

Between the left and right MF points, the shadow of the spine is unavoidably superimposed on the center of the image of the mandible resulting in poor definition of the cortex. Lateral to the AG points there can also be superimposition of shadows of the opposite side of the mandible, making the endosteal border indistinguishable and the lower mandibular border difficult to discern from other structures. Thus, the best region from which to measure cortical thickness is between the AG and MF points, and it is from these that the PDM used for search (Section III-C) was built (though the example in Fig. 5 is from a PDM including points beyond the AG and MF points to make visual interpretation easier).

The model was built using 200 points in all by interpolating between the AG and MF points with 50 equally spaced points along the upper and lower margins of the mandibular cortex on each side defined by the mean manually marked points. The landmark points used in model construction were the mean of the two sets of manual points.

C. Search

PDMs may be used in image search as ASMs [7]. ASMs have been shown to be capable of robust location of image structure in the presence of confusing image features such as those that occur in DPTs. In this application, search is conducted by iterative local refinement. Starting from an appropriate shape and pose (typically the model average position—see Section III-D), sample profiles are constructed normal to the boundary at each landmark point. The position of maximum gradient along the profile is chosen as the updated position of the landmark. If the shape of the model were modified directly using the new point positions, the result would not, in general, be a legal example of the modeled shape. By using $\delta\mathbf{x}$—the vector of displacements of the landmark points, we can impose the shape constraints on the model by rearranging (1) to give

$$\delta\mathbf{b} = \mathbf{P}^T\,\delta\mathbf{x}. \tag{2}$$

This results in a set of shape parameters $\mathbf{b}$, and so constrains the new example to shapes that can be generated by the PDM. However, though this restricts variation in shape to the axes dictated by the eigenvectors of the PDM, it does not limit how *far* along each of these axes the new shape lies. This means that whilst the PDM cannot generate *any* possible shape (like a circle or a star), it can produce examples of mandibles with features exaggerated beyond anything that are likely to be found.

Realistic limits on how far along each of the shape axes to go are determined from the training set. For each example shape in the training set there is a shape vector $\mathbf{b}$, and so, each b_i has a variance determined by its eigenvalue λ_i. We would therefore expect the sum

$$\sum_{i=1}^{t} \frac{b_i^2}{\lambda_i} \tag{3}$$

to follow a χ^2 distribution. Thus, by setting a limit on this sum, and using the area under the χ^2 distribution below this limit, we can retain a desired percentage of the variance observed in the training set for new example shapes. The value of this limit is discussed below in Section III-D.

The above process is iterated until changes in landmark positions are sufficiently small, the precise threshold depending on the application. In this case, ten iterations were found to be sufficient to reduce changes in position to approximately 0.2 pixels (0.026 mm).

The abundance of confounding structure in the images results in the location of incorrect edges during ASM search. While the imposition of shape constraints reduces the effects of such erroneous detections, both speed and accuracy of search can be affected by the detection of "outlier" points. Rogers and Graham [11] have shown that robust estimation of model parameters can lead to much more accurate fits in these circumstances. Here, we use a version of ASM search that uses M-estimators [10] to fit the model parameters in (2), using the method described in [11]. Briefly, the process of estimating the parameters $\mathbf{b}$ for a given shape $\mathbf{x}$ proceeds by minimizing the residuals $\mathbf{r} = (\mathbf{x} - \mathbf{x}_0)$ where $\mathbf{x}_0$ is the current set of model points. In the M-estimator method a set of weights ω are calculated based on the standard deviation of the residuals σ, and thus

$$\omega_i = \begin{cases} 1, & r_i < \sigma \\ \sigma/|r_i|, & \sigma \le r_i < 3\sigma \\ 0, & r_i \ge 3\sigma \end{cases} \tag{4}$$

These weights are then used to determine the influence of each point on the estimation of the model parameters.

Since the inferior border of the IMC is far more clearly defined than the superior border (Fig. 2), the search is divided into two phases. The first phase uses a model built from the points on the inferior border only to locate that edge, defining the overall shape and pose of the mandible. This result is used as the initialization of phase 2, which is a search using the complete model of the IMC to obtain the positions of both inferior and superior borders.

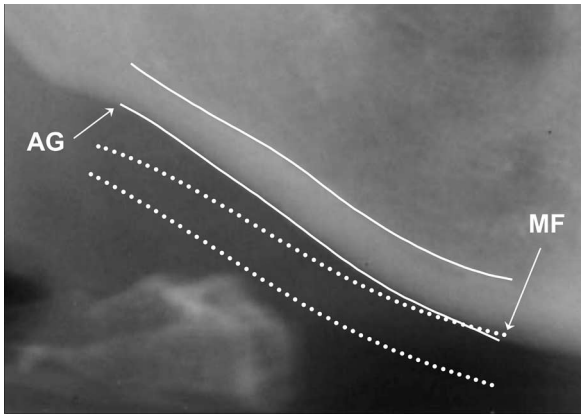


Fig. 6. Example of the results of a model fit (UFit) showing the initial instance of the model, i.e., the start point of the search (*dotted line*), and the results of fitting the model to the data (*solid line*).

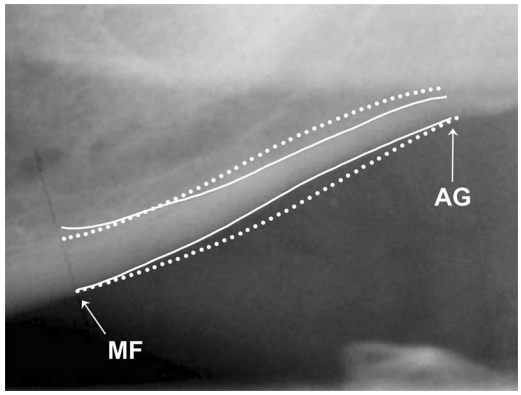


Fig. 7. An example of the results of a model fit (4PFit) showing the initial instance of the model, i.e., the start point of the search (*dotted line*), and the results of fitting the model to the data (*solid line*).

D. Experimental Procedure

To test the ability of an ASM search to segment an unseen example, i.e., one not used in the training set, a leave-one-out methodology is employed. Here, the model is trained on all examples except the one to be tested, and this is repeated for all examples in the data set.

Two versions of the ASM search described in Section III-C were tested experimentally on the data set.

The first was a free search without any manual initialization points. The ASM search for the inferior border of the IMC (phase 1) was initialized from the mean position and pose of the training data. For some images, the correct shape and pose are some distance from this starting point. ASM search uses a multiresolution coarse-to-fine search strategy in such circumstances [7], and that was employed in this case. The results of this search were then used to initialize a full endosteal border ASM search by warping the mean example of the full endosteal PDM such that its lower edge matched the results of the lower edge ASM fit. We refer to fits determined this way as "unconstrained fits" or "UFits." An example of a UFit search result showing start condition and final results is shown in Fig. 6.

The second version used four manually defined reference points on the lower mandible edge at the left and right AG and MF as starting points. To start the phase 1 search, the mean example of the lower mandible border PDM was stretched and positioned such that its AG and MF points matched the manually placed start points. An edge-based ASM search was then initiated, making no further reference to the manual points during the search. The full endosteal border ASM search was then initiated from the results of the phase 1 search in the same way, as described in the unconstrained fits.

The use of this straightforward interaction allowed us to decouple the effects of location and shape in ASM search. Starting the search so close to the true position guarantees that the search will finish up with the correct pose. The quality of ASM fit is determined solely by the ability of the PDM to represent the variation in shape that occurs among the images. We refer to fits determined this way as "constrained fits" or "4PFits." An example of the results of a 4PFit showing the initial start condition and final search result is shown in Fig. 7.

There are a number of parameters, which need to be set in an ASM search such as sample profile length, degree of PDM shape constraint (3), number of resolution levels, etc., and the optimum values for these were found empirically.

For the unconstrained multiresolution fit, a shape constraint of 99% (see Section III-C) was required to provide sufficient flexibility to accommodate the variation in shapes while retaining sufficient shape constraint to avoid unfeasible matches arising from the spurious edge features close to the mandible. However, for the full resolution fitting of the complete mandibular cortex model in either the 4PFit method or the UFit method, a constraint of 100% was necessary for the model to be able to describe the fine detail of the endosteal border accurately and give the best sensitivity to bone mineral density.

This is effectively a removal of model parameter constraint since the tail of the χ^2 distribution goes on to infinity, however, this still restricts the shapes to those possible along the axes of variation within the PDM (see Section III-C above). This high degree of model flexibility was possible since the second phase search started very close to the correct position, and so, only the mandibular cortical edges would be within reach of the search profiles of the models.

E. Image Resolution

The panoramic dental tomograms were scanned from film at a resolution of 25.64 pixels/mm. At this resolution, the film grain is visible, contributing a source of noise in the images, which was found to interfere with ASM search. To overcome this, a degree of smoothing was necessary. Dental panoramic radiography in digital format is becoming increasingly used; these images typically have a resolution of 8.8 pixels/mm, and hence, it is appropriate to evaluate the effectiveness of the method for segmenting images at the current digital resolution. Experiments over a range of subsampled resolutions on the data set showed that reducing the resolution by Gaussian smoothing and subsampling, to that of the digital radiographs had little effect on the model fit accuracy, as measured by the point-to-point difference,

TABLE I
MODEL FIT ACCURACY RESULTS

Comparison	Region	Point Difference (mm)		Cortical Thickness (mm)		
		point-to-point	point-to-curve	bias	lower lim	upper lim
Manual 1-2	AG-MF	2.45 (2.45)	0.31 (0.33)	-0.02	-0.77	0.72
	MF	2.19 (3.00)	0.38 (0.38)	0.13	-1.09	1.36
Manual-UFit	AG-MF	5.73 (4.57)	0.49 (1.58)	-0.30	-1.01	0.40
Manual-4PFit	AG-MF	0.59 (0.54)	0.31 (0.40)	-0.25	-0.79	0.28
	MF	0.71 (0.70)	0.45 (0.56)	-0.31	-2.03	1.41
Fit1-Fit2	AG-MF	2.31 (2.44)	0.14 (0.24)	-0.04	-0.21	0.29
	MF	2.23 (2.99)	0.27 (0.41)	-0.08	-1.32	1.16

Point differences are presented as mean value (standard deviation).

and no effect on the sensitivity to reduced BMD, which is indicated by ROC analysis. Therefore, the results presented in the following sections are based on a 30% reduced resolution of 7.69 pixels/mm, which is approximately equivalent to digital radiographs.

IV. RESULTS

In common with many studies in medical image analysis, we define "accuracy" to mean conformity with expert medical annotation. To compare the model fits with the manual annotation we use the mean point-to-point, and the mean point-to-curve difference between the manually placed points and those resulting from the model fit in order to estimate the accuracy of the model fit. Since our goal is to measure mandibular cortical thickness, we also compare the measurements of thickness derived from model fits with those from manual annotation. The thickness is measured as the distance between corresponding points on the lower and upper border of the mandibular cortex. The comparison is done using a Bland and Altman plot [12], where the difference between two sets of measurements are plotted against their mean. From this analysis, the *bias* is measured as the mean difference between the two sets of measurements, and the limits of agreement are the mean difference ± 1.96 standard deviations.

Ultimately, we wish to test the sensitivity of the derived measurements to osteoporosis and this is done by calculating the correlation coefficient between the parameter in question and BMD, and by plotting an ROC curve [13]. The area under the curve (AUC) can be used to quantify the overall diagnostic efficacy of the parameter in question—ranging from 0.5 (no better than chance) to 1.0 (perfect discrimination).

A. Fit Accuracy

The accuracy results are summarized in Table I. Four comparisons are made.

1) *Manual 1–2*—comparison of the manual annotation of the two observers.
2) *Manual-UFit*—comparison of the mean of the manual points with that of the unconstrained model fit.
3) *Manual-4PFit*—comparison of the mean of the two sets of manual points against the results of the four-point initialized model fit (4PFit).
4) *Fit1-Fit2*—the 4PFit involves user interaction, and hence, there is a certain degree of subjectivity invlolved in the exact placement of the four initialization points. To estimate the magnitude of this effect, we perform two 4PFits, each initialized by a different observer.

Each of the above comparisons were made for the whole region of the mandible annotated, (i.e., the AG-MF region), and separately the MF points, as these are the points used in manual measurement. The point-to-point and point-to-curve differences are presented as "mean value (standard deviation)." For comparison of cortical thickness measurements using the Bland–Altman plots, the bias is the mean of the differences between the two sets of measurements being compared, and the limits of agreement are the mean difference $\pm 1.96\sigma$ [12]. As an example, Fig. 10 shows the Bland–Altman plot for the Mannual 1–2 comparison.

For the Manual-UFit comparison, the point-to-point differences are large—more than twice that of the manual inter-observer reliability. This is because the unconstrained fits are able to successfully find the correct location and shape of the mandible, but may not find the correct scale. Because, the mandible exhibits a strong grey-level edge along its lower border, there is strong evidence in the image for the position of a point orthogonal to the mandibular edge, but there are no features such as edges to characterize the position of a particular point *along* the edge of the mandible, i.e., its correct mediolateral position with respect to the AG and MF landmarks. This is borne out in the difference between point-to-point and point-to-curve differences when comparing manual annotations. This means that once the lower mandible ASM has adhered to the mandible edge, there is no motivation in the search mechanism to stretch or contract to the correct scale. This suggests that an unconstrained ASM fit will accurately measure a portion of the cortical thickness, but that the exact anatomical region of the mandibular cortex that is being measured cannot be guaranteed. This is demonstrated in Fig. 8 where the results of an unconstrained fit are plotted against the target image, along with the user defined AG and MF points. The upper and lower borders of the IMC has been correctly located, but the points are displaced laterally with respect to the AG and MF points. Fig. 9 shows the equivalent image for a 4PFit—here anatomical correspondence is always guaranteed by the initialisation points.

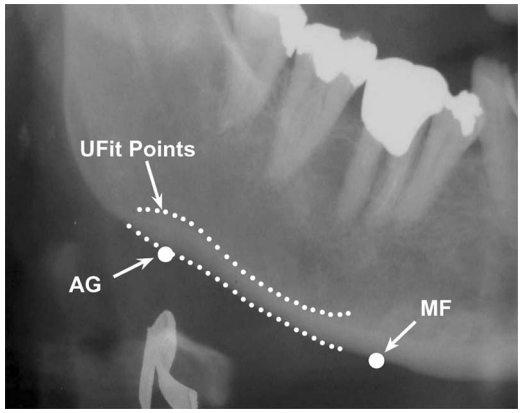


Fig. 8. Example of the results of a fully automatic model fit. Though the lower upper and lower bounds of the inferior mandibular cortex are accurately delineated, anatomical correspondence with respect to the AG and MF points cannot be guaranteed.

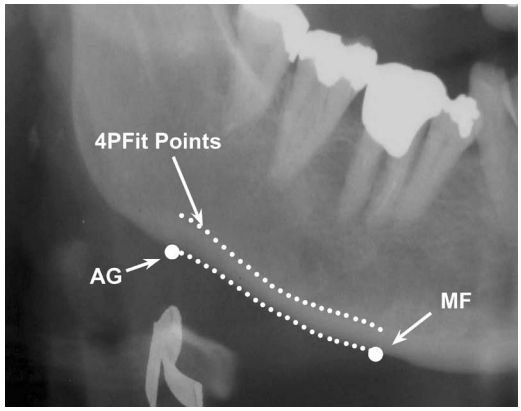


Fig. 9. Example of the results of a model fit using four-point manual initialisation on the same patient as shown in Fig. 8 - here the anatomical correspondence is correct.

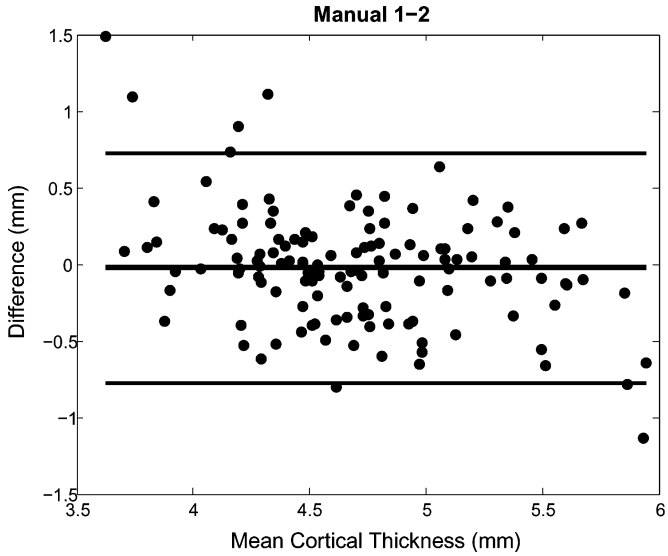


Fig. 10. Bland–Altman plot comparing the mean cortical width (AG-MF region) measured manually by two observers. $Bias = -0.02$ mm, Limits of agreement $= -0.77$–0.72 mm.

TABLE II
CORRELATION COEFFICIENTS BETWEEN CORTICAL THICKNESS MEASUREMENTS AND BMD

Parameter	Spine BMD		Hip BMD		Min T-score	
	r	p	r	p	r	p
UFit	**0.21***	0.009	**0.19**	0.016	**0.24***	0.003
Manual AG-MF	**0.21***	0.009	**0.19**	0.016	**0.24***	0.003
Manual MF	0.08	0.187	0.13	0.073	0.13	0.073
4PFit AG-MF	**0.23***	0.005	**0.25***	0.002	**0.27***	0.001
4PFit MF	0.10	0.133	0.09	0.158	0.11	0.110

Figures in *bold* exceed the p = 0.05 threshold and figures with a "*" exceed the p = 0.01 threshold.

Thus, the point-to-curve differences and bias and limits of agreement for the AG-MF region are only slightly higher than those of the Manual-Fit and the Manual 1–2 comparison, but the search result is more accurate than the point-to-point differences would suggest. Because the location along the mandible edge cannot be guaranteed in the unconstrained fit, the results for the MF are of little meaning and so are not included in this case.

For the Manual-4PFit, the point-to-point differences are much lower than inter-observer (Manual 1–2) equivalent. This is because between two observers there is much greater subjectivity in the position of the points *along* the mandibular border than there is in their distance *from* the border. Hence, the point-to-curve differences for Manual 1–2 comparison are much lower than the point-to-point differences, and are comparable with those of the Manual-4PFit comparison.

For the Manual-4PFit comparison, the bias in the cortical thickness measurement is larger than the manual inter-observer bias suggesting a systematic difference between the maximum gray-level gradient, and the edge perceived by the human observers. For the AG-MF region, the limits of agreement are slightly lower than the Manual 1–2 comparison, but considerably larger for the MF points. Closer inspection of the model fits suggests these larger limits of agreement are due to poor ASM location of the upper MF points, most probably due to image noise in this region caused by the shadow of the spine.

For the Fit1-Fit2 comparison, the point-to-point differences are similar to those of the Manual 1-2 comparison, since subjectivity in point position along the mandibular border has been reintroduced by the use of two sets of initialization points. However, the point-to-curve difference and the bias and limits of agreement for the AG-MF region are much lower than the other comparisons (Table I).

B. Sensitivity To Reduced BMD

1) BMD Correlations: Table II shows the correlation coefficients (Pearson's) between the cortical thickness measurements and the BMD values for the hip, spine, and the minimum BMD T-score of the two sites (see Section II-B). Figures in *bold* exceed the $p = 0.05$ threshold, and figures with a "*" exceed the $p = 0.01$ threshold.

None of the cortical thickness measurements derived from the MF points show significant correlation, whereas, the AG-MF measurements do. For the hip and spine BMDs, the 4PFit

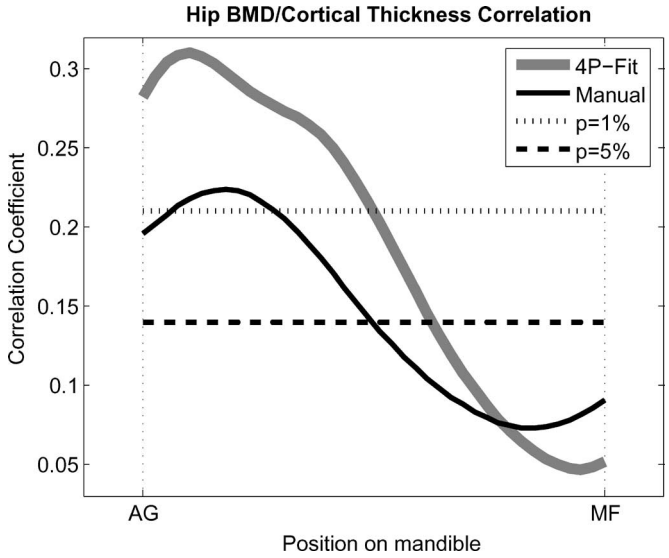


Fig. 11. Correlation coefficient between cortical thickness and Hip BMD plotted as a function of position along mandible. For each of the 100-point positions along the mandible, a mean of ten adjacent cortical thickness measurements are plotted.

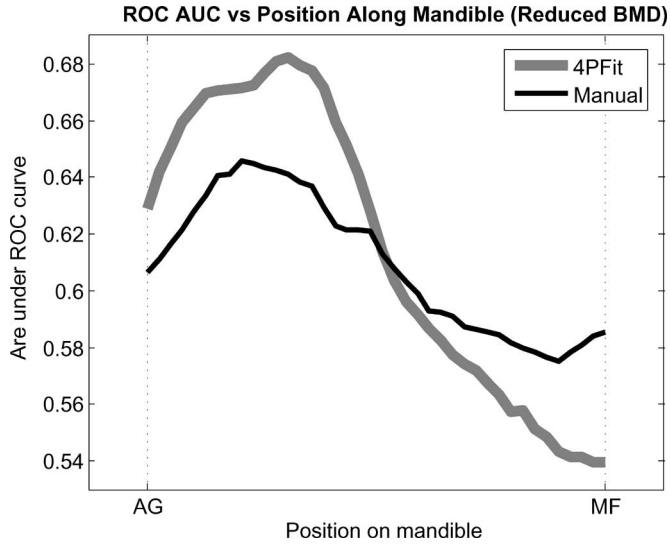


Fig. 12. Area under the ROC curve for reduced BMD as a function of position along mandible from the AG to MF points. The results of manual annotation and initialized model fit (4PFit) are compared. For each of the 100-point positions along the mandible, a mean of ten adjacent cortical thickness measurements are plotted.

performs better than the manual measurements, or the UFit measurements.

Fig. 11 demonstrates how the correlation between cortical thickness and BMD of the hip varies with position along the mandible for the manual and 4PFit results. The cortical thickness measurements show greatest sensitivity to BMD in the lateral halves of the mandible, and least around the MF points. This would suggest that the MF points are not the optimal place to measure cortical thickness in the detection of osteoporosis. A similar pattern is observed for the spine BMD, and the minimum T-score of the two sites.

2) Roc Analysis: Because, the number of osteoporotic patients is so small (5) in this data set, it is difficult to generate a meaningful ROC curve using osteoporosis as the categorical variable, and so, in the following analysis, the osteoporotic and osteopenic patients are combined into a "reduced BMD" group of 47.

Fig. 12 shows the relationship between position along the mandible and the area under the ROC curve derived from the cortical thickness at that position. The overall pattern is similar to the results for direct BMD correlation (Fig. 11) in that the sensitivity of cortical thickness measurement to reduced BMD are optimal in the lateral halves of the mandible. Fig. 13 shows the ROC curves obtained using the cortical thickness averaged over this optimum region of the mandible for both manual and 4PFit points yielding an area under the curve (AUC) of 0.66 and 0.71, respectively. Analysis of the original films corresponding to this data set using the traditional visual methods yielded an AUC of 0.63 [3].

Fig. 14 shows the resulting ROC curve for reduced BMD using the mean cortical thickness obtained from the unconstrained model fit as the discriminating parameter. Compared to the equivalent from the four-point initialized model fit, the two curves are indistinguishable ($p = 0.60$).

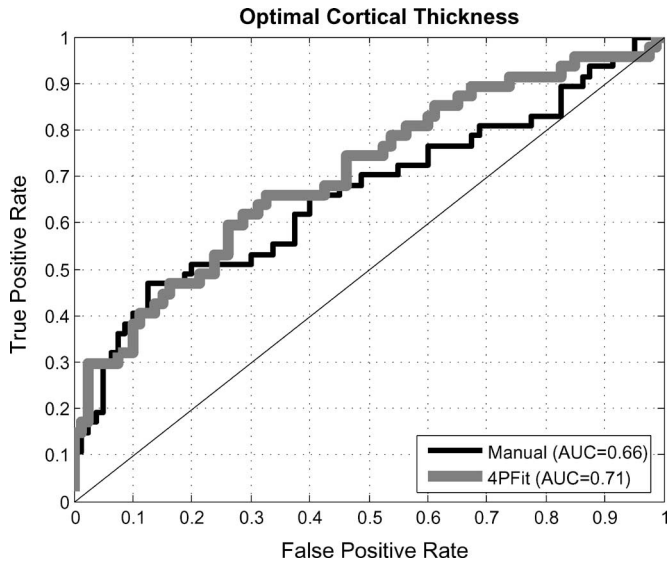


Fig. 13. ROC curve for reduced BMD using the cortical thickness from the optimal region suggested by Fig. 12 as the discriminating parameter. The results of manual annotation and initialised model fit (4PFit) are compared.

V. Discussion

One problem that all studies of this kind face is the lack of absolute ground-truth against which to test the proposed measurement technique. Within the limits of this study, the average results from two expert manual observers is the only reference we have against which to test the ASM model fit in its ability to measure cortical width. Therefore, all measurements of fit accuracy presented depend on the accuracy of the manual measurements. The sensitivity of the ASM method to BMD and its ability to detect osteopenia, however, can be compared

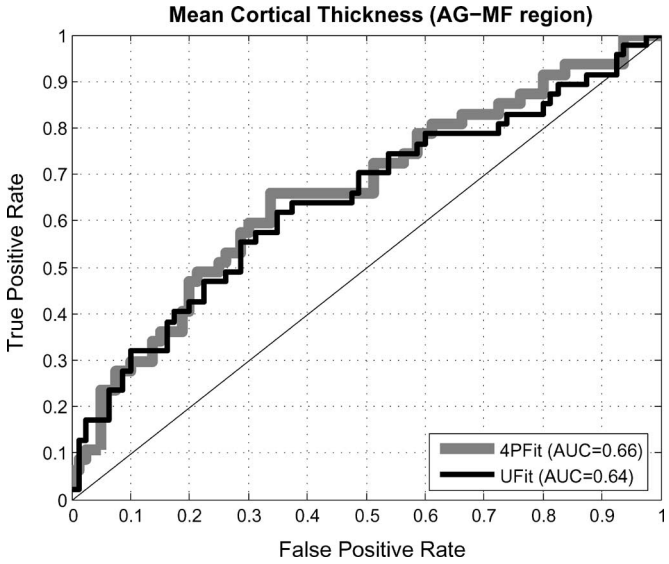


Fig. 14. ROC curve for Osteopenia using the mean cortical thickness (AG-MF region) as the discriminating parameter. Initialized model fit (4PFit) and the unconstrained fit (UFit) are compared.

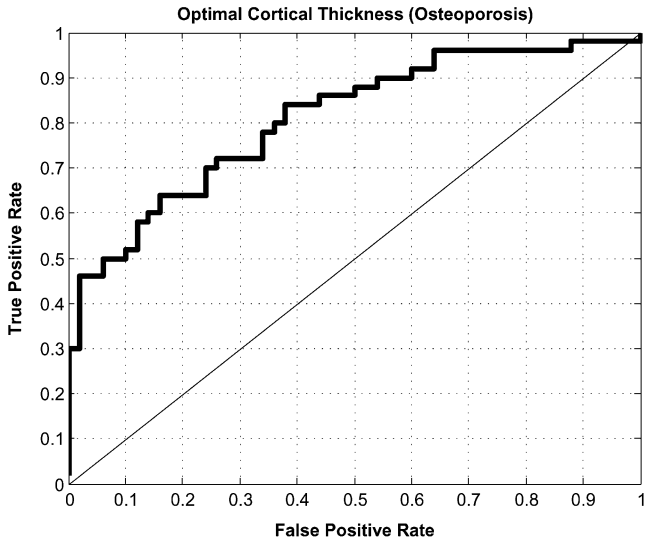


Fig. 15. ROC curve for Osteoporosis using the optimal cortical thickness measured from a 4PFit as the discriminating parameter ($AUC = 0.81$). Here, a data set of 50 osteoporotics and 50 nonosteoporotics is used.

directly with the objective DXA measurements, independently of the subjectivity of expert image interpretation.

We can conclude from the above that it is possible to accurately measure the width of the inferior mandibular cortex in panoramic dental tomograms using an edge-based ASM method. For these measurements to have an exact anatomical correspondence, four manually placed initialization points are required. This is a reasonable level of interaction, since only the lower mandible edge need be identified—a clearly visible feature in all patients—and the points only need to be placed close to the border, not exactly on it, since the ASM search will locate the exact position of the local edge anyway.

Correlation of the cortical thickness with the BMD measured from the spine or hip was highest for the lateral portion of the AG-MF region of the mandible. The results of the 4PFit were an improvement on the manual measurements when compared with both hip and spine BMD, and the fully automatic unconstrained model fit yielded correlations equivalent to the manual results. This indicates that even in the cases where there is reduced anatomical correspondence between the model position and the mandible, width measurements still produce useful information on BMD.

The ROC analysis appears to confirm the conclusion that the lateral portion of the AG-MF region of the mandible is the optimum area from which cortical thickness is measured. This effect is more pronounced for the model fit results than the manual measurements, as model fitting in the MF region is relatively poor due to noise mostly from the shadow of the spine.

It is conceivable that this observation reflects real physiological effects. For example, the cortical bone in this region medial to the antegonial point AG may be more sensitive to the systemic effects of skeletal osteoporosis. In other regions such as that adjacent to the mental foramen, the local musculature may preserve bone due to local functional stimulation.

The low number of osteoporotic patients in this data set make it difficult to derive meaningfull ROC analysis for detection of osteoporosis, as the statistical confidence intervals tend to be large. Development of the model fitting method was performed on this data set, because it represents a realistic sample of routine panoramic dental tomograms from patients likely to benefit from low BMD screening.

Fig. 15 shows an ROC curve for detecting osteoporosis ($T < -2.5$), generated using DPTs from 50 osteoporotic and 50 nonosteoporotic individuals, who did not contribute to the training set. This allows us to conduct a limited experiment on truly "unseen" images, and to make a preliminary evaluation of the diagnostic efficacy in detecting osteoporosis. The resulting ROC curve has a larger area ($\mathrm{AUC} = 0.81$) than those shown above since the more severe condition of *osteoporosis* is being used as the discriminating factor rather than *osteopenia*. Traditional measurement of the cortical thickness by five experts from the original films yields an AUC of between 0.61 and 0.68 for the same 100 individuals. Traditional manual analysis of data from 653 subjects yielded an AUC range of 0.71–0.78 [16].

As this research was intended to develop a clinical tool for diagnosis, it is worth considering from a dentist's perspective, and in the context of everyday practice. Most dentists currently use radiographs on film. There has, however, been a steady proliferation of digital radiology in dentistry over the last 15 years, and it is appropriate to develop computed methods of image analysis for dental use.

It should be remembered that the manual measurements made for this study were the result of *expert* annotation and that the direct measurement techniques that have received previous research attention [14] require time and care to give a result. Furthermore, the repeatability of such measurements may not be acceptable.

Osteoporosis diagnosis is not part of everyday dental practice, and hence, any involvement in this task should be facilitated for

the dentist. Measurement of cortical width will only be practical if it is fully automatic or very nearly so. The limited and straightforward interaction described here may be sufficiently unobtrusive to be practical. However, the results of automatic search indicate that useful measurement can be made without the involvement of dentist. The improved specificity and sensitivity arising from being able to make measurements at anatomically precise locations holds out the possibility of improved diagnostic performance. The ASM method does of course require training by experts, however this has already been done in this study and is not required for further applications of the method. The application of the trained model to the unseen dataset described earlier demonstrates this.

There have been calls [15] for improving access to dual-energy X-ray absorptiometry for individuals at risk of osteoporosis. Dentists are in a unique position to carry out fortuitous identification of patients at risk of osteoporosis and make a contribution to general healthcare. This research offers the potential to facilitate this process. Further work is in progress, collecting DPT's and DXA measurements from a large patient sample to establish the diagnostic validity of our technique and the diagnostic threshold appropriate for clinical practice.

REFERENCES

[1] K. Horner, H. Devlin, C. W. Alsop, I. M. Hodgkinson, and J. E. Adams, "Mandibular Bone Mineral Density as a Predictor of Skeletal Osteoporosis," *Brit. J. Radiol.*, vol. 69, no. 827, pp. 1019–1025, 1996.

[2] H. Devlin and K. Horner, "Mandibular radiomorphometric indices in the diagnosis of reduced skeletal bone mineral density," *Osteoporos Int.*, vol. 13, no. 5, pp. 373–378, May 2002.

[3] K. Horner, H. Devlin, and L. Harvey, "Detecting patients with low skeletal bone mass," *J. Dent.*, vol. 30, no. 4, pp. 171–175, May 2002.

[4] C. V. Devlin, K. Horner, and H. Devlin, "Variability in measurement of radiomorphometric indices by general dental practitioners," *Dentomaxillofacial Radiol.*, vol. 30, no. 2, pp. 120–125, 2001.

[5] R. J. Tanner, B. F. Wall, P. C. Shrimpton, and D. Hart, *et al.*, *Frequency of Medical and Dental X-Ray examinations in the UK1997/98/ NRPB-R320*. Philadelphia, PA: Chilton/NRPB, 2001.

[6] J. A. Kansis, "Diagnosis of osteoporosis and assessment of fracture risk," *Lancet*, vol. 359, pp. 1929–1936, 2002.

[7] T. F. Cootes, C. J. Taylor, D. H. Cooper, and and J. Graham, "Active Shape Models–Their Training and Application," *Comput. Vis. Image Understand.*, vol. 61, no. 1, pp. 38–59, 1995.

[8] G. Behiels, F. Maes, D. Vandermeulen, and P. Suetens, "Evaluation of image features and search strategies for segmentation of bone structures in radiographs using active shape models," *Med. Image Anal.*, vol. 6, no. 1, pp. 47–62, 2002.

[9] T. F. Cootes, A. Hill, C. J. Taylor, and J. Haslam, "Use of active shape models for locating structures in medical images," *Image Vis. Comput.*, vol. 12, no. 6, pp. 355–365, 1994.

[10] P. J. Huber, *Robust Statistics*. New York: Wiley, 1981.

[11] M. Rogers and J. Graham, "Robust active shape model search," *Proc. Eur. Conf. Comput. Vision 2002, Lecture Notes Comput. Sci.,* vol. 2353, pp. 517–530.

[12] J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *Lancet*, pp. 307–310, Feb. 8, 1986.

[13] J. A. Hanley and J. M. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, pp. 29–36, 1982.

[14] S. C. White, "Oral Radiographic Predictors of osteoporosis," *Dentomaxillofacial Radiol*, vol. 31, pp. 84–92, 2002.

[15] J. E. Compston, "Action plan for the prevention of osteoporotic fractures in the European community," *Osteoporosis Int.*, vol. 15, pp. 259–262, 2004.

[16] K. Karayianni, K. Horner, A. Mitsea, L. Berkas, M. Mastoris, R. Jacobs, C. Lindh, P. F. van der Stelt, E. Harrison, J. E. Adams, S. Pavitt, and and H Devlin, "Accuracy in osteoporosis diagnosis of a combination of mandibular cortical width measurement on dental panoramic radiographs and a clinical risk index (OSIRIS): The OSTEODENT project," *Bone*, vol. 40, no. 1, pp. 223–229, 2006.

P. Danny Allen received the B.Sc. degree in physics with astrophysics and the Ph.D. degree in astrophysics from the University of Leeds, Leeds, U.K., in 1991 and 2007, respectively.

He is currently a Research Associate in the Division of Imaging Science and Biomedical Engineering at the University of Manchester, Manchester, U.K. His current research interests include the practical application of computer image analysis to medicine.

Jim Graham (M'01) received the B.Sc. degree in physics from the University of Edinburgh, Edinburgh, Scotland, in 1974, and the Ph.D. degree in structural biology from the University of Cambridge, Cambridge, U.K., in 1978.

After that, he joined the University of Manchester, Manchester, U.K., to work on practical applications of image analysis. He is currently a Senior Lecturer in the Division of Imaging Science and Biomedical Engineering, University of Manchester. His current research interests include biological and medical image analysis, and computer vision.

Damian J. J. Farnell received the B.Sc. degree in mathematical physics and the Ph.D. degree in theoretical physics from the University of Manchester Institute for Science and Technology (UMIST), Manchester, U.K., in 1991 and 1994, respectively.

In 2002, he joined the University of Manchester, Manchester, where he is currently a Statistician in the School of Medicine and works on medical image analysis. From 2003 to 2006, he was a Lecturer in ophthalmic imaging at the University of Liverpool, Liverpool, U.K. His current research interests include medical image analysis and numerical biosimulation.

Elizabeth J. Harrison received the B.Sc. degree in physics from the University of Newcastle, Newcastle upon Tyne, U.K., in 1995, and the Ph.D. degree in bone densitometry, in 2003.

She is currently a Research Associate in the Division of Imaging Science and Biomedical Engineering at the University of Manchester, Manchester, U.K. Her current research interests include technical and practical aspects of bone densitometry technologies.

Reinhilde Jacobs received the Ph.D. degree in dentistry from the Katholieke Universiteit Leuven, Leuven, Belgium, in 1996, and the Masters degree in dental radiology from the University of London, London, U.K., in 2002.

She is currently a Professor at the School of Dentistry of the Katholieke Universiteit Leuven. Her current research interest include oral imaging with a specific focus on imaging in relation to bone, and implantation.

Kety Nicopolou-Karayianni received the Dental degree from the University of Thessalonica, Thessalonica, Greece, in 1975, and the Ph.D. degree from the University of Athens, Athens, Greece, in 1985.

She is currently a Professor in the Department of Oral Radiology and Diagnosis at the School of Dentistry, University of Athens. Her current research interests include the assessment of diagnostic performance of X-ray imaging systems.

Christina Lindh received the DDS degree from the Lund University, Lund, Sweden, in 1974 and the doctoral degree (Odont Dr) at the Faculty of Odontology, Lund University, Sweden, in 1996.

She is currently a Senior Lecturer and an Assocoiate Professor at the Department of Oral Radiology, Malmö University, Malmö, Sweden. Her current research interests include imaging of jaw bone tissue with a focus on osteoporosis and implantology.

Paul F. van der Stelt received the Graduation degree in dentistry, in 1974, and the Ph.D. degree in radiology, in 1979.

He is a Professor of oral and maxillofacial radiology at the Academic Centre for Dentistry Amsterdam (ACTA), Amsterdam, The Netherlands. His current research interests include digital radiology and the implementation of image processing in radiodiagnosis.

Keith Horner received the B.ChD. degree (Hons.) from Leeds University, Leeds, U.K., in 1981, the M.Sc. degree in experimental oral pathology from the London University, London, U.K., in 1985, and the Ph.D. degree from Manchester University, Manchester, in 1997.

He is currently a Professor of oral and maxillofacial imaging in the School of Dentistry, University of Manchester. His current research interests include bone quality, osteoporosis, and evidence-based use of imaging in dentistry.

Hugh Devlin received the Bachelors degree in dental surgery from the University of Bristol, Bristol, U.K., in 1976, and the Ph.D. degree from the University of Manchester, Manchester, U.K., in 1988, on osteoporosis-related work.

He is a Reader at the Dental School, University of Manchester. His current research interests include image analysis and statistics, and specifically, the application of dental diagnostic methods in detecting osteoporosis.

38. **Automated osteoporosis risk assessment by dentists: a new pathway to diagnosis.** H. Devlin, P.D. Allen, J. Graham, R. Jacobs, K. Karayianni, C. Lindh, P.F. van der Stelt, E. Harrison, J.E. Adams, S. Pavitt and K. Horner, *Bone **40**: 835-442, 2007.* doi: 10.1016/j.bone.2006.10.024

ELSEVIER

Bone 40 (2007) 835–842

BONE

www.elsevier.com/locate/bone

Automated osteoporosis risk assessment by dentists: A new pathway to diagnosis

H. Devlin [a,*], P.D. Allen [b], J. Graham [b], R. Jacobs [d], K. Karayianni [c], C. Lindh [e], P.F. van der Stelt [f], E. Harrison [b], J.E. Adams [b], S. Pavitt [a], K. Horner [a]

[a] *School of Dentistry, University Dental Hospital, Higher Cambridge Street, Manchester, M15 6FH, UK*
[b] *Imaging Science and Biomedical Engineering, University of Manchester, UK*
[c] *Dental School, University of Athens, Greece*
[d] *Oral Imaging Centre, School of Dentistry, Oral Pathology and Maxillofacial Surgery, Katholieke Universiteit Leuven, Belgium*
[e] *Faculty of Odontology, Malmő University, Sweden*
[f] *Academic Centre for Dentistry, Amsterdam, The Netherlands*

Received 28 May 2006; revised 29 September 2006; accepted 29 October 2006
Available online 22 December 2006

Abstract

General dental practitioners use a vast amount of panoramic radiography in their routine clinical work, but valuable information about patients' osteoporotic status is not collected. There are many reasons for this, but one of the prime reasons must be the disruption involved in clinical routine with lengthy manual radiographic assessment. We have developed computer software, based on active shape modeling that will automatically detect the mandibular cortex on panoramic radiographs, and then measure its width. Automatic or semi-automatic measurement of the cortical width will indicate the osteoporotic risk of the patient. The aim of our work was to assess the computer search technique's ability to measure the mandibular cortical width and to assess its potential for detection of osteoporosis of the hip, spine and femoral neck.

Mandibular cortical width was measured using the manually initialized (semi-automatic) method and, when assessed for diagnosing osteoporosis at one of the three measurement sites, gave an area under the ROC curve (A_z)=0.816 (95% CI=0.784 to 0.845) and for the automatically initialized searches, A_z=0.759 (95% CI=0.724 to 0.791). The difference between areas=0.057 (95% Confidence interval=0.025 to 0.089), $p<0.0001$. For diagnosing osteoporosis at the femoral neck, mandibular cortical width derived from the manually initialized fit gave an area under the ROC curve (A_z)=0.835 (95% CI=0.805 to 0.863) and for the automatically initialized searches A_z=0.805 (95% CI=0.773 to 0.835). The difference in A_z values between active shape modeling search methods=0.030 (95% CI=−0.010 to 0.070), and this was not significant, p=0.138.

We concluded that measurement of mandibular cortical width using active shape modeling is capable of diagnosing skeletal osteoporosis with good diagnostic ability and repeatability.
© 2006 Elsevier Inc. All rights reserved.

Keywords: Osteoporosis; Active shape modeling; Risk assessment; Mandible; Radiography

Introduction

Mandibular cortical width on dental panoramic radiographs is significantly correlated with bone mineral density at the hip [1], lumbar spine [2] and forearm [3], the most common sites of fracture related to osteoporosis in post-menopausal women. Measuring mandibular cortical width could be used for diagnosis as a screening tool for osteoporosis. Taguchi et al. [2] found that mandibular cortical thickness was related to the bone mineral density of the third lumbar vertebra. Devlin and Horner [3] found that mandibular cortical width had moderate accuracy when used to diagnose skeletal osteopenia. Subsequent work [4] advised that a cortical thickness of less than 3 mm in the mental foramen region should be a trigger for referral for dual energy X-ray absorptiometry (DXA).

While DXA facilities are often limited, millions of dental panoramic radiographs are taken every year across Europe. A recent study based in the United Kingdom, showed that 61% of

* Corresponding author. Fax: +44 161 275 6480.
E-mail address: Hugh.Devlin@manchester.ac.uk (H. Devlin).

8756-3282/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.bone.2006.10.024

general dental practitioners used panoramic radiography equipment [5]. Measurements of mandibular cortical width from them may prove to be a cost-effective, efficient triage method of selecting those patients at high-risk of osteoporosis [4].

One important barrier to using cortical width measurements in primary dental care is the significant observer variability in measurement that is not improved by individualized instruction [6]. Furthermore, manual measurement of cortical width may be seen as a time-consuming interruption by the dentist in their busy schedule.

We have developed computer software, based on active shape modeling [7], that will automatically detect the mandibular cortex on panoramic radiographs and then measure its width. Active shape modeling is a technique widely used in computer vision to detect shapes and analyze them and has been used successfully to replicate the shape of vertebrae [8] and accurately detect the edge shape of bone in digitized radiographs [9]. Once the mandibular cortex has been detected using the active shape model, multiple measurements of width and further analysis of the endosteal border become possible with minimal user interaction.

In 2003, the 3-year OSTEODENT project was commenced, consisting of collaboration by five European centers to investigate the role of dental radiographs in the diagnosis of osteoporosis. The overriding aim of this project was to identify the most valid and effective radiographic index, or combination of radiographic and clinical indices, for the diagnosis of osteoporosis applicable for use by dentists. The aim of the work reported here was to assess the computer search technique's ability to measure the mandibular cortical width and to assess its potential for detection of osteoporosis of the hip, spine and femoral neck.

Method

Ethical approval was given for the recruitment of female subjects (aged 45–70 years) following their informed consent. The study was open to all female patients in this age-group, except those who suspected that they might be pregnant. No one was excluded from recruitment based on race or pre-existing medical condition such as secondary osteoporosis. With Ethics Committee approval, those who have previously had a bone density scan performed and were identified as having a below average bone density were recruited into the study. Recruitment of osteoporotic individuals was encouraged to provide a sufficient sample size with narrow confidence intervals around both sensitivity and specificity values of the diagnostic tests.

We compared the diagnostic ability of clinical risk indices with that of the computer radiographic measurements. Two well established indices were chosen, that of the National Osteoporosis Foundation (NOF) index [10] and the Osteoporosis Risk Assessment Index (ORAI) [11]. The NOF index scores 1 point for each of the following: patient is >65 years, weight <57.6 kg, maternal/paternal history of fracture, current cigarette smoking, and a personal history of fracture. ORAI used the following subject scoring system: age >75 years (+15), age 65–74 years (+9), age 55–64 (+5), body weight <60 kg (+9), body weight 60–70 kg (+3), estrogen therapy (+2). The total score for each patient was calculated for the two indices.

Central dual energy X-ray absorptiometry (DXA)

Dual energy X-ray absorptiometry (DXA) scans were performed on the Hologic QDR 4500, Hologic Discovery (Hologic Inc., Bedford, Massachusetts, USA) and the GE Lunar Prodigy (GE Lunar Corporation, Madison, Wisconsin) at four centers throughout Europe. The four centers used were located in Leuven (Belgium), Athens (Greece), Manchester (UK), and Malmo (Sweden) and ambulant female patients were recruited from the area surrounding these centers. Shewart's rules were used to monitor quality assurance throughout the study period [12].

The European spine phantom was used to standardize measurements between different manufacturers using the method described by Pearson and colleagues [13]. *T* and *Z* scores were calculated using Hologic reference data for the lumbar spine and NHANES reference data for the proximal femur [14].

Patients were diagnosed as osteoporotic according to the World Health Organization (WHO) criteria, i.e. those with a bone mineral density *T*-score value 2.5 S.D. or more below the mean value of the young sex matched reference population.

Radiography

The subjects received a dental panoramic radiograph examination while biting on a plastic block in the left premolar region. The plastic block contained a spherical, steel ball bearing (3.175 mm diameter), which was used to compensate for the image magnification.

Digital and conventionally processed dental panoramic radiography machines were used. Leuven (Belgium) and Malmo (Sweden) used a Cranex III (Soredex, IL, USA) whereas Athens (Greece) and Manchester (UK) used a Planmeca (Planmeca USA, Roselle, IL, USA). The imaging parameters also varied but typically were 70 kV at 8 mA for 15 s. In Leuven, ADC Solo (Afga, Mortsel, Belgium) was used as the photostimulable phosphor plate system for image capture and digital read out, but other centers used a conventional film/cassette.

Point distribution model

The radiographs were digitized using a Kodak LS85 digitizer (Eastman Kodak, Rochester, NY) at a resolution of 25.64 pixels/mm. Using a previous training set of 132 DPR images, a point distribution model (PDM) [7] of the inferior mandibular cortex was created by manual annotation of the endosteal and periosteal borders. Two experts performed this task independently using a graphical user interface, outlining the inferior mandibular cortex by placing equally spaced points on the computer images between the mental foramen and antegonial region. The mean point position of both experts was used to define the shape of the cortex, and the PDM built using 200 points interpolating between the manually placed points. The point distribution model was used to search for and identify the inferior mandibular cortex. Principal component analysis applied to the covariance matrix of the point position allows the main "modes of variation" of the target shape to be determined. The point distribution model captured the principal modes of variation of the shape of the inferior mandibular cortex. These modes of variation were manipulated in an image search program to find the region of the image whose shape was contained within the observed shape distribution of the training data, and which provided the best fit to the expected image appearance. The point distribution model was then used to locate the inferior mandibular cortex in the OSTEODENT

sample, a set of dental panoramic radiographs which had not been included in the training set. A search strategy was used in which, from a given starting shape and pose, the point distribution model iteratively deformed in an attempt to align the points with the strongest image gradient (edge) found within a predefined search area around each point. This combination of point distribution model and search mechanism is known as an active shape model.

The initial stage of the active shape model search was to locate just the lower border of the inferior mandibular cortex using a point distribution model built only from points corresponding to the lower border. This is because the periosteal border of the inferior mandibular cortex is a much more clearly defined feature than the upper endosteal border. The final stage of the active shape model search used a point distribution model of the endosteal and periosteal borders of the inferior mandibular cortex.

The points defining the inferior mandibular cortex were positioned on a border which was well defined in a direction perpendicular to the edge, but not along it. In other words whilst we can be reasonably confident that the endosteal and periosteal borders of the inferior mandibular cortex were correctly defined in the active shape model search, the correct anatomical positioning of the points with respect to the mental foramen and antegonial regions could not be guaranteed.

Thus, two search strategies were investigated: a constrained search using manually placed initialization points, and an unconstrained search with no manual initialization. In the constrained search, only four points were planted by the user on the periosteal border of the inferior mandibular cortex at the left and right mental foramina and antegonial regions. The point distribution model of the mean shape was then warped so that its mental foramina and antegonial points matched those placed by the user, and this was used as the search start point with no further reference to the initialization points being used during the search. Accurate placement of the initialization points by the user with respect to the mandibular cortex was not required since the border was located by the active shape model search. The unconstrained, active shape model search was completely automatic and started from the mean shape and pose found in the training set—in this case a multi-resolution coarse-to-fine search strategy was required for robust search results. The endosteal border was not defined by the user in either the constrained or unconstrained searches, but was located by the active shape model search.

After convergence of the search, the mandibular cortical width was measured at a series of contiguous locations along the lower border of the mandible between the antegonion and the mental foramina for all subjects. The measured width at each location was averaged over an interval of approximately 10% of the length of the lateral cortical border. For each of these locations the correlation was calculated between the measured cortical width and the skeletal bone mineral density ground-truth. Fig. 1(a) shows the correlation values as a function of location along the mandible. Statistically significant correlations were found in the lateral region of the inferior mandibular cortex, the highest correlation occurring at 10–20% of the distance from the antegonion to the mental foramen. Fig. 1(b) shows a similar curve plotting the A_z values arising from ROC analysis. Maximum sensitivity is obtained for measurements made in the same region. The "optimal measurement region" defined by this measurement is indicated in Fig. 2. Results reported from the manually-initialized and fully automatic searches all refer to measurements made in this region.

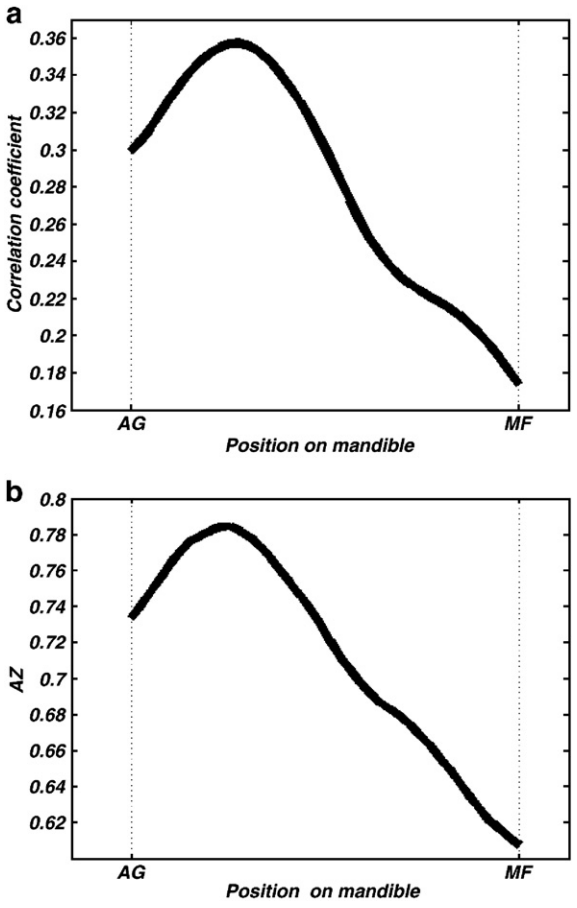


Fig. 1. From the manually initialized ASM fit, the mandibular cortical thickness was calculated between the antegonial and mental regions. The cortical width measurements from the lateral region of the mandible were (a) most highly correlated with the minimum *T*-score derived from all three BMD measurement sites or (b) provided the highest A_z values in ROC analysis. In (a), the correlation coefficient significance levels are 0.098 and 0.069 at the $p=0.01$ and $p=0.05$ levels, respectively; with all our data points comfortably exceeding the 1% threshold. AG = antegonial position, MF = mental foramen position.

The width of the ball bearing image was used to scale the linear cortical width measurements in the constrained and unconstrained fits. On each of the digital images the position of the ball bearing was marked manually. The image was then cropped around this position to a size larger than the expected size of the ball bearing. A canny edge detector [15] was then used to detect the edges in this cropped region and a Hough transform [16] used to isolate those edges belonging to an ellipsoidal object. From the detected elliptical image, the dimensions of the ball bearing were then calculated. The measured dimensions of the ball-bearing were used to scale the width measurements made at different centers.

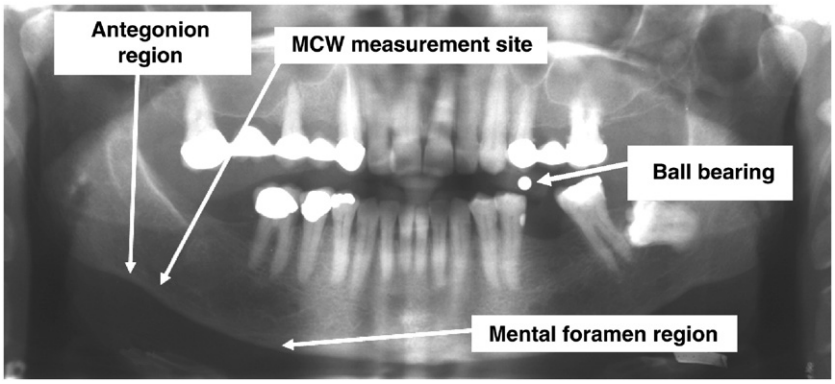


Fig. 2. Dental panoramic radiograph (DPR) showing antegonial and mental regions, between which the measurements of mandibular cortical width were made. The ball-bearing used for inter-center calibration is visible on the left side of the mandible.

Statistical analysis

ROC curve analysis was used to measure the diagnostic abilities of the computed measurements of cortical width in diagnosis of osteoporosis. In this respect, separate analyses were performed for a diagnosis of osteoporosis at any measured site (lumbar spine, femoral neck or total hip) and at the femoral neck alone. The areas under ROC curves (A_z) were calculated using the Medcalc® software program (MedCalc Software, Mariakerke, Belgium).

Repeatability of automated methods

It was not possible in our study to measure repeatability of the cortical width measurements on radiographs obtained from multiple exposures of the same patient for ethical reasons. The robustness of the technique was tested by repeated measurement of the same radiographic sample set.

Automatically initialized search

The same model applied to the same digital image produced identical searches each time, so the reproducibility error was zero in this case.

Manually initialized search

There is a source of variability in manually initialized search, as four initialization points need to be specified interactively. The variability arising from doing this was simulated by perturbing each of the four initialization points by a set distance in a random direction and repeated for 10 searches. The size of the perturbation was calculated from the two sets of manual mark-ups and corresponded to the mean distance between the two sets of four points placed manually at the mental foramen and antegonial positions. The mean within subject variance was calculated for the 10 searches using one-way ANOVA. The repeatability is the difference between two measurements for the same subject and is expected to be less than 2.77 times the within-subject standard deviation for 95% of pairs of observations.

Results

671 subjects were recruited in total. 10 subjects were eliminated from the study because 8 were outside the age range of the inclusion criteria (45–70 years) and bone mineral density data was incomplete on 2. The radiographs of a further 9 subjects were unsuitable for further analysis due to unacceptable quality, accidental loss, data corruption or the absence of a ball bearing image and were eliminated from the study. Of the remaining 652 subjects that formed the study population, 140 had osteoporosis at one of the three measurement sites and 65 had osteoporosis at the hip. The mean age of the subjects was 54.9 years (S.D. = 6.10).

For the manually and automatically initialized searches, there were significant differences between the mandibular cortical widths of normal subjects and those with osteoporosis at one of the three measurements sites (Table 1; using Mann–Whitney U test, $p < 0.0001$ for both searches).

There were significant correlations between the mandibular cortical widths derived from both computer image searches and bone mineral density at the total hip, spine and femoral neck (Table 2).

ROC curves were plotted of cortical width derived from manually initialized searches for diagnosing osteoporosis at one of the three sites (Fig. 3) or at the femoral neck (Fig. 4). The manually initialized searches gave higher A_z values for the ROC curves than the automatically initialized searches (Figs. 5 and 6).

Table 1
Mean mandibular cortical width (MCW) and standard deviation (S.D.) derived from automatically and manually initialized searches for osteoporotic and normal individuals

	N	Mean MCW	S.D.
Manually initialized fit	Normal	3.747	0.596
	osteoporosis at any of three sites	3.031	0.552
Automatically initialized fit	Normal	3.778	0.681
	osteoporosis at any of three sites	3.117	0.682

Table 2
Mandibular cortical width derived from automatically and manually initialized search correlated with bone mineral density measured at the total hip, spine and femoral neck using dual X-ray energy absorptiometry

	Spearman's rho	Manually initialized search	Automatically initialized search
Automatically initialized search	Correlation coefficient	0.722	1.000
	Sig. (2-tailed)	$p<0.001$	
Total hip	Correlation coefficient	0.399	0.328
	Sig. (2-tailed)	$p<0.001$	$p<0.001$
Femoral neck	Correlation coefficient	0.460	0.376
	Sig. (2-tailed)	$p<0.001$	$p<0.001$
Lumbar spine	Correlation coefficient	0.433	0.359
	Sig. (2-tailed)	$p<0.001$	$p<0.001$

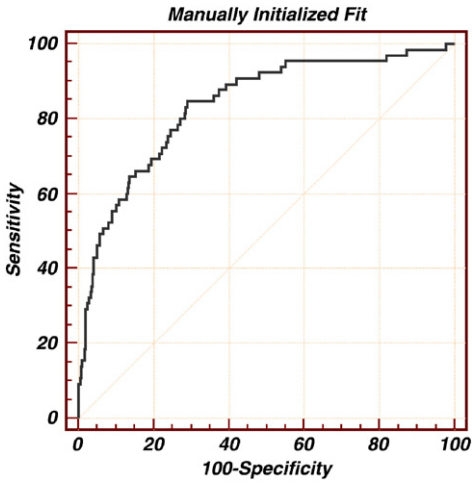


Fig. 4. Receiver Operating Characteristic curve for the measurements of cortical width obtained by the manually initialized method in the diagnosis of osteoporosis at the femoral neck. A_z=0.835 (95% CI=0.805 to 0.863).

Comparison of methods

The difference in area (A_z) under the ROC curves was used to identify the most effective method for diagnosis of osteoporosis in any site. Mandibular cortical width derived from the manually initialized fits gave an area under the ROC curve (A_z) of 0.816 (95% CI=0.784 to 0.845) and for the automatically initialized searches, A_z was 0.759 (95% CI=0.724 to 0.791). The manually initialized search model had a significantly greater A_z than the automatically initialized search model (A_z difference=0.057; 95% CI 0.025 to 0.089, p=0.0001).

For diagnosis of osteoporosis at the femoral neck, mandibular cortical width derived from the manually initialized fit gave an area under the ROC curve (A_z)=0.835 (95% CI=0.805 to 0.863) and for the automatically initialized searches A_z=0.805 (95% CI=0.773 to 0.835). There was no significant difference in A_z between the two methods (A_z difference=0.03, 95% CI −0.009 to 0.070, p=0.135).

There was a significant correlation between manually initialized and automatically initialized search models (Spearman's rho=0.722, $p<0.0001$, 95% CI=0.683 to 0.757). A Passing & Bablok plot [17] was used to compare the manually initialized and automatic search results (Fig. 7). The 95% confidence intervals of the slope and intercept were used to determine significant differences from 1 and 0, respectively.

Manual Search (Y)=0.30+0.90 Automatic Search (X).

Intercept=0.30 (95% CI: 0.13 to 0.47) and Slope=0.90 (95% CI: 0.85 to 0.95).

There was no significant deviation in linearity between the automatic or manually initialized search methods ($p>0.10$, using the Cusum test).

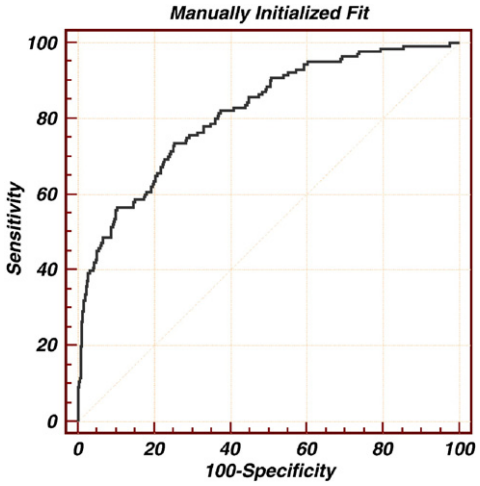


Fig. 3. Receiver Operating Characteristic curve for the measurements of cortical width obtained by the manually initialized method in the diagnosis of osteoporosis at any site (lumbar spine, femoral neck or total hip). A_z=0.816 (95% CI=0.784 to 0.845).

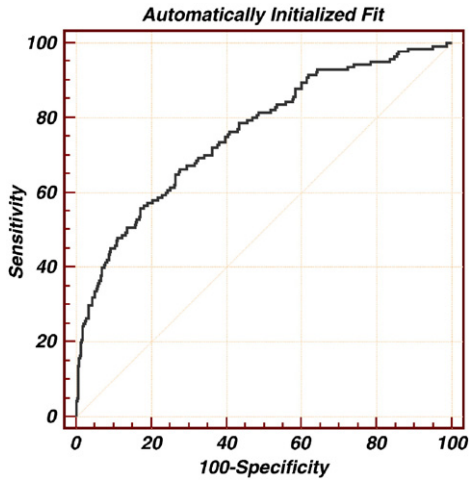


Fig. 5. Receiver Operating Characteristic curve for the measurements of cortical width obtained by the automatically initialized method in the diagnosis of osteoporosis at any site (lumbar spine, femoral neck or total hip). A_z=0.759 (95% CI=0.724 to 0.791).

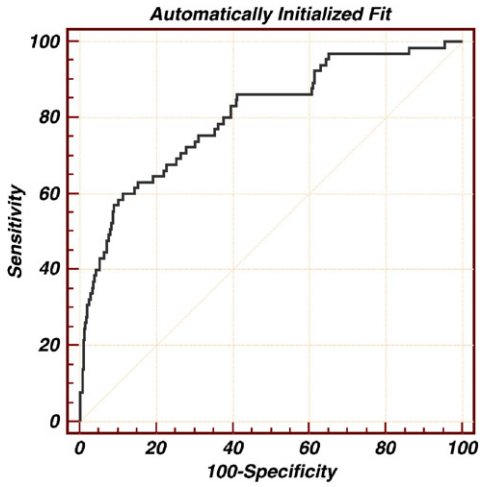


Fig. 6. Receiver Operating Characteristic curve for the measurements of cortical width obtained by the automatically initialized method in the diagnosis of osteoporosis at the femoral neck. A_z=0.805 (95% CI=0.773 to 0.835).

Clinical indices

Table 3 summarizes the area under the ROC curve for the radiographic measurements and the clinical indices in diagnosing osteoporosis. ORAI performed significantly better than the NOF index at detecting osteoporosis at the femoral neck (p=0.001) and at any one of the three measurement sites (p<0.001). For detecting osteoporosis at the femoral neck there was no significant difference in A_z values between the manually initialized fit and ORAI (p=0.431), and between the automatically initialized fit and ORAI (p=0.109). For detecting osteoporosis at one of the three sites, there was no difference in A_z values between the manually initialized fit and ORAI (p=0.641) and between the automatically initialized fit and ORAI (p=0.135).

Numbers of patients detected

If a <3 mm threshold was applied to the manually initialized search data, then 119 patients would have "failed" the radiographic test and been referred out of a total population of 652. Of these 119 referred patients, 72 were found to have osteoporosis at one site as measured using DXA. The probability that the referred patients had osteoporosis, given that they had failed the radiography test, was therefore 60.5%. This contrasts with the prior probability of osteoporosis in our study population of 21.5%. Using the automatically initialized search data with a cut-off threshold of <3 mm, 126 patients would have been referred with osteoporosis, of whom 67 were found to have osteoporosis using DXA. The probability that any of these referred patients had osteoporosis, given that they had failed the radiographic test was 53.2%, over double the prior probability of osteoporosis. A high threshold of <4 mm with the manually initialized search gave an excellent sensitivity of 96.4% (but poor specificity of 29.5%).

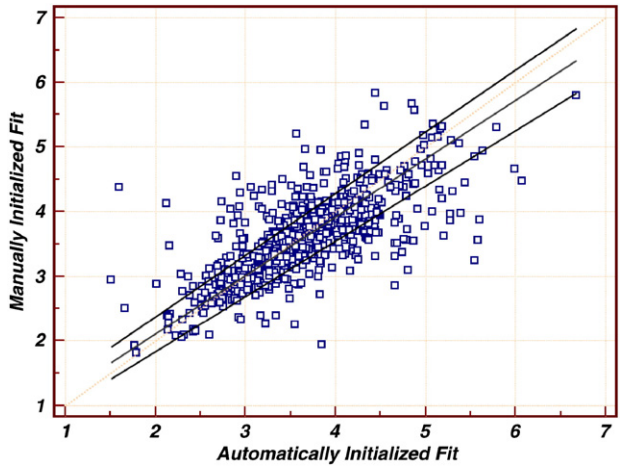


Fig. 7. A Passing & Bablok plot to compare the manually initialized and automatic search results. Linear regression line (with 95% CI).

Table 3
Summary table comparing radiographic assessment by computer with the clinical indices National Osteoporosis Foundation (NOF) index and the Osteoporosis Risk Assessment Index (ORAI)

	AUC for osteoporosis at the femoral neck A_z	AUC for osteoporosis at any of three sites A_z
Manually initialized fit	0.835	0.816
Automatically initialized fit	0.805	0.759
ORAI	0.861	0.803
NOF	0.732	0.671

AUC=area under the ROC curve.

Repeatability of manually initialized search

For all 652 subjects, the mean mandibular cortical width was 3.59 mm (range 1.81 to 5.83 mm). The mean within-subject variance for the subsequent 10 manually initialized searches (corrected for image magnification) was 0.062 mm, giving a measurement error of 0.25 mm and repeatability of 0.69 mm. This error range indicates that the difference between two measurements for the same subject is expected to be less than 0.69 mm for 95% of pairs of observations.

Discussion

Mass screening for the detection of osteoporosis is not recommended as cost-effective. Instead a cheaper method of detecting those at high risk of osteoporotic fracture is desirable. Such a method should require minimal input from the clinician to be cost-effective. Active shape modeling (ASM) search has been used for robust location of anatomical features in a number of medical imaging studies [18]. Of particular interest to the current study is the location and measurement of the shape of the femur in radiographs [19], and detecting the edges of bone in digitized radiographs [9]. In the present study, the automated image analysis software performed in an equivalent manner to that reported previously with manual measurements by experts [20]. Manually initialized ASM measurement of cortical width produced an A_z=0.816 in ROC curve analysis that was high for detecting osteoporosis at either the total hip, femoral neck and lumbar spine and for the automatically initialized searches,

A_z=0.759. Taguchi et al. [20] found that the A_z for their expert manual radiographic measurement was 0.802 (95% CI, 0.705 to 0.899) in detecting osteoporosis. With our image analysis software, we anticipate that the automated computer measurement would alert the examining dentist to the patient's thin mandibular cortex. The dentist could follow up by an examination of the patient's risk factors (low body mass index, age, steroid use etc) and then consider the advisability of further referral for central dual energy X-ray absorptiometry. Our analytical software has a comparable diagnostic ability in detecting osteoporosis to clinical indices such as ORAI, as given by the area under the ROC curve.

Using the manually initialized fit for predicting osteoporosis at the hip, A_z was 0.835, and for the automatically initialized fit was 0.805 indicating these provided good diagnostic tests in predicting osteoporosis at this site.

The manually initialized fit used defined points on the dental panoramic radiograph in the medio-lateral direction placed by the observer, which limited the search along the mandible. The automatically initialized fit was unable to use any edge features on the image with which to establish correct anatomical placement of the points with respect to the antegonion and mental foramen regions. In other words, there were no features along the edge of the mandible which could be used to define the position of the points in the fully automated search. The mean values for cortical width measured using the automatically initialized search were greater than for the manually initialized search (Fig. 7 and Table 1), resulting in reduced sensitivity. This resulted in consistently greater cortical width values for the automatically initialized fit than the manually initialized fit in Fig. 6, with a slope significantly different from unity. The mean values for cortical width measured using the automatically initialized search were greater than for the manually initialized search (Table 1). The manually initialized search strategy provided better A_z values than the automatic search. At the appropriate operating point, both sensitivity (true positive fraction of those with osteoporosis) and specificity (true negative rate) were improved, with corresponding reductions in false positives and false negatives. For maximum diagnostic accuracy, some minimal observer interaction is therefore necessary.

Given the less than perfect diagnostic accuracy of the cortical width measurements, the dental panoramic radiograph would not be taken for osteoporosis screening *per se*, but some unrelated dental investigation. In addition, the patient's case history and medical data must be considered before undertaking further referral and investigation. Radiographic measurements cannot be used as the sole basis for referral. With these limitations, our computer methods of osteoporosis triage could be cost-effective as the assessment is performed automatically on digital films, with minimal intervention from the dentist. Setting the threshold for further investigation to achieve a low percentage of false positive diagnoses would be possible by considering the high specificity end of the ROC curve. Our future investigations will consider the precise threshold required to minimize false positive diagnosis but yet still provide a good diagnostic test. Manual measurements of mandibular cortical width by general dental practitioners have poor repeatability [6], but this problem would be eliminated using the computer-based technology we have described.

The repeatability of the manually initialized search was better than that previously described for the manual measurements of different experts. Devlin and Horner [3] found that the limits of agreement, which indicate the interval in which 95% of measurement differences lie, were +1.32 mm for manual measurements of cortical width. Larger limits of agreement were found when the manual cortical width measurements were made by general dental practitioners [6].

We have previously recommended that patients with a mandibular cortical width of <3 mm should be referred for further DXA investigation. That recommendation was on the basis of our manual measurements of cortical width [4]. The higher sensitivity and specificity and better reproducibility that arises from the digital analysis would, of course, result in an improved analysis at this threshold. Of more importance is to consider the proportion of women labeled "osteoporotic" by this test who are not truly osteoporotic (as measured by DXA). Taguchi et al. [20] found that 60% of their patients who had a cortical width of less than 3 mm were osteoporotic. Threshold values for manual measurements on dental radiographs should be chosen which balance sensitivity and specificity for the prevailing health care systems. Using a high threshold for mandibular cortical width, such as 4.785 mm below which patients are classed as osteoporotic [21], will result in excellent sensitivity, but poor specificity. In the environment seen in many European countries of inadequate availability of DXA [22,23], a low sensitivity/high specificity strategy may be more appropriate, at least where DXA availability is less than the minimum recommended 8 units per million population.

Digital analysis increases the diagnostic yield of radiographs [24]. Gregory et al. [19] developed an active shape model of the femur. They found that the gross morphology of the femur could be used to identify patients who may develop a hip fracture in the future. Despite differences in the positioning of patients and femur in their study, as well as variable magnification of the images, they found that their active shape model was more robust than other methods. We have compensated for magnification errors by asking the patient to bite on a plastic block incorporating a ball bearing of 3.175 mm diameter during the radiographic exposure. Magnification in the DPR is about 20–36% for most machines. Another computer-aided diagnostic technique [25] also used panoramic radiographs to provide osteoporotic pre-screening, but required considerable operator input e.g. to correctly identify the mental foramen. This is a potential weakness in view of the non-visibility of the mental foramen in a minority of patients [26] and the low intra-examiner agreement in localizing the mental foramen on DPRs [27]. Arifin et al. [25] described a semi-interactive method based on image processing in which the cortex was distinguished by thresholding, high-pass filtering and morphological operations to enhance the image. They showed that this method delivered measurements of similar diagnostic value to manual measurement. Their measurements were restricted to the region of the mental foramen, and required

input from an expert user in defining the position of the mental foramen.

In conclusion, active shape modeling is capable of automatic mandibular morphometry of dental panoramic radiographs, producing a diagnostic test of skeletal osteoporosis that is comparable to that achieved by manual measurement. Our technique does have some shortcomings and sources of error, but it requires minimal interaction by the clinician and provides automated warning if the patient is at high risk of osteoporosis. Future work will use further automated image analysis of the morphological features of the cortex to improve the diagnostic accuracy of our methodology. In particular, we believe our methodology has further potential for development where automated detection of low bone density may be beneficial such as implantology, and assessment of the effect of osteoporosis on fracture healing, tooth loss and periodontitis [28].

Acknowledgments

This work was supported by a research and technological development project grant from the European Commission Fifth Framework Program 'Quality of Life and Management of Living Resources' (QLK6-2002-02243; 'OSTEODENT').

References

[1] White SC, Taguchi A, Kao D, Wu S, Service SK, Yoon D, et al. Clinical and panoramic predictors of femur bone mineral density. Osteoporos Int 2005;16:339–46.
[2] Taguchi A, Tanimoto K, Suei Y, Ohama K, Wada T. Relationship between the mandibular and lumbar vertebral bone mineral density at different postmenopausal stages. Dentomaxillofac Radiol 1996;25:130–5.
[3] Devlin H, Horner K. Mandibular radiomorphometric indices in the diagnosis of reduced skeletal bone mineral density. Osteoporos Int 2002;13:373–8.
[4] Horner K, Devlin H, Harvey L. Detecting patients with low skeletal bone mass. J Dent 2002;30:171–5.
[5] Tugnait ACV, Hirschmann PN. Radiographic equipment and techniques used in general dental practice: a survey of general dental practitioners in England and Wales. J Dent 2003;31:197–203.
[6] Devlin CV, Horner K, Devlin H. Variability in measurement of radiomorphometric indices by general dental practitioners. Dentomaxillofac Radiol 2001;30:120–5.
[7] Cootes TF, Cooper DH, Graham J. Active shape models—Their training and application. Comput Vis Image Underst 1995;61:38–59.
[8] Smyth PP, Adams JE. Vertebral shape: automatic measurement with active shape models. Radiology 1999;211:571–8.
[9] Behiels G, Vandermeulen D, Suetens P. Evaluation of image features and search strategies for segmentation of bone structures in radiographs using active shape models. Med Image Anal 2002;6:47–62.
[10] Cadarette SM, Jaglal SB, Murray TM, McIsaac WJ, Joseph L, Brown JP. Evaluation of decision rules for referring women for bone densitometry by dual-energy X-ray absorptiometry. JAMA 2001;286:57–63.
[11] Richy F, Gourlay M, Ross PD, Sen SS, Radican L, De Ceulaer F, et al. Validation and comparative evaluation of the osteoporosis self-assessment tool (OST) in a Caucasian population from Belgium. Q J Med 2004;97:39–46.
[12] Orwoll ES, Oviatt SK. Longitudinal precision of dual-energy X-ray absorptiometry in a multicenter study. The Nafarelin/Bone Study Group. J Bone Miner Res 1991;6:191–7.
[13] Pearson J, Dequeker J, Henley M, Bright J, Reeve J, Kalender W, et al. European semi-Anthropomorphic spine phantom for the calibration of bone densitometers—Assessment of precision, stability and accuracy—The European Quantitation of Osteoporosis Study Group. Osteoporos Int 1995;5:174–84.
[14] Looker A, Wahner H, Dunn W, Calvo M, Harris T, Heyse S, et al. Updated data on proximal femur bone mineral levels of US adults. Osteoporos Int 1998;8:468–90.
[15] Canny J. A computational approach to edge-detection. IEEE Trans PAMI 1986;8:679–98.
[16] Illingworth J, Kittler J. A survey of the Hough transform. Comput Vis Graph Image Process 1988;44:87–116.
[17] Passing H, Bablok W. A new biometrical procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison studies in clinical chemistry: I. J Clin Chem Clin Biochem 1983;21:709–20.
[18] Rogers M. Robust active shape model search. In: Heyden ASG, Nielsen M, Johansen P, editors. Computer Vision—ECCV 2002: 7th European Conference on Computer Vision, Copenhagen, Denmark, vol. 2353; 2002. p. 517–30.
[19] Gregory JS, Stewart A, Undrill PE, Reid DM, Aspden RM. A method for assessment of the shape of the proximal femur and its relationship to osteoporotic hip fracture. Osteoporos Int 2004;15:5–11.
[20] Taguchi A, Ohtsuka M, Nakamoto T, Tanimoto K. Screening for osteoporosis by dental panoramic radiographs. Clin Calcium 2006;16:67–73.
[21] White SC, Atchison KA, Gornbein JA, Nattiv A, Paganini-Hill A, Service SK, et al. Change in mandibular trabecular pattern and hip fracture rate in elderly women. Dentomaxillofac Radiol 2005;34:168–74.
[22] Compston J, Papapoulos SE, Blanchard F. Report on osteoporosis in the European Community: current status and recommendations for the future. Working Party from European Union Member States. Osteoporos Int 1998;8:531–4.
[23] Compston J. Action plan for the prevention of osteoporotic fractures in the European Community. Osteoporos Int 2004;15:259–62.
[24] Geraets WG, Van der Stelt PF, Lips P, Van Ginkel FC. The radiographic trabecular pattern of hips in patients with hip fractures and in elderly control subjects. Bone 1988;22:165–73.
[25] Arifin AZ, Taguchi A, Nakamoto T, Ohtsuka M, Tanimoto K. Computer-aided system for measuring the mandibular cortical width on panoramic radiographs in osteoporosis diagnosis. Proc. SPIE Med Imaging, vol. 5747. Image Processing Conference; 2005. p. 813–21.
[26] Yosue T, Brooks SL. The appearance of mental foramina on panoramic radiographs. I. Evaluation of patients. Oral Surg, Oral Med Oral Pathol 1989;68:360–4.
[27] Sakakura CE, Monteiro Loffredo L de C, Scaf G. Diagnostic agreement of conventional and inverted scanning panoramic radiographs in the detection of the mandibular canal and the mental foramen. J Oral Implant 2004;30:2–6.
[28] Hohlweg-Majert B, Schmelzeisen R, Pfeiffer BM, Schneider E. Significance of osteoporosis in craniomaxillofacial surgery: a review of the literature. Osteoporos Int 2006;17:167–79.

39. **The role of the dental surgeon in detecting osteoporosis: the Osteodent study.** H. Devln, P.D. Allen, J. Graham, R. Jacobs, K. Karayianni, C. Lindh, E. Marjanovic, P.F. van der Stelt, J.E Adams, S. Pavitt and K. Horner, *British Dental Journal, **204**: E16*, *2008*. doi:10.1038/sj.bdj.2008.317

The role of the dental surgeon in detecting osteoporosis: the OSTEODENT study

H. Devlin,[1] P. Allen,[2] J. Graham,[3] R. Jacobs,[4] K. Nicopoulou-Karayianni,[5] C. Lindh,[6] E. Marjanovic,[7] J. Adams,[8] S. Pavitt,[9] P. van der Stelt[10] and K. Horner[11]

IN BRIEF

- Women at high risk of osteoporosis can be identified by dentists using information from panoramic radiographs supplemented by a few clinical questions.
- Dentists may contribute to a woman's general health by facilitating onward referral to medical colleagues.
- The software that carries out the radiographic assessment requires minimal dentist input to work optimally.

RESEARCH

Objective To determine if thinning (<3 mm width) of the lower cortical border of the mandible on dental panoramic radiographs, as well as other clinical risk factors, may provide a useful diagnostic test for osteoporosis in young postmenopausal women. **Design** Six hundred and fifty-two subjects (age range 45-70 years) were involved in this multi-centre, cross-sectional study. **Setting** Patients were recruited from centres in Leuven (Belgium), Athens (Greece), Manchester (UK), and Malmo (Sweden). **Subjects and methods** The subject's age, body weight, whether the patient took hormone replacement therapy or had a history of low trauma fracture were used to form a clinical osteoporosis risk assessment (the OSteoporosis Index of RISk or OSIRIS index). Each patient also received a dental panoramic radiographic examination. **Results** One hundred and forty subjects had osteoporosis involving at least one of the measurement sites (lumbar spine, femoral neck or total hip). Those with osteoporosis tended to have a low OSIRIS score and a thinned cortical mandibular border. The area under the ROC curve for using both cortical width and OSIRIS to predict osteoporosis was 0.90 (95% CI = 0.87 to 0.92). There was a significant improvement in the diagnostic ability of the combined OSIRIS and cortical width test over both tests applied separately ($p < 0.001$). The cost effectiveness of the cortical width and OSIRIS model was improved by using a high specificity threshold rather than high sensitivity. However, this analysis ignores the costs associated with missed cases of osteoporosis. **Conclusion** Dentists have a role to play in the detection and referral of patients at high risk of osteoporosis.

INTRODUCTION

Osteoporosis is a serious disease, but treatment can be instituted when early detection is made possible. Hip fractures, in particular, are associated with significant mortality and morbidity in the elderly,[1] but in one study, less than one fifth (18%) of high risk people had received medical treatment for osteoporosis before the occurrence of hip fracture.[2] The current failure to assess and treat patients at high risk of osteoporosis may be partly due to insufficient resources or time, but a failure of health professionals to identify risk factors and refer the patient for definitive diagnosis using dual-energy X-ray absorptiometry (DXA) is an important contributory factor.

In an attempt to improve this situation, several clinical risk 'tests' have been developed as a means of identifying subjects who would benefit from further investigation. The contribution of clinical risk factors (such as OSIRIS) to the primary prevention of osteoporotic fractures is at present under consideration by the World Health Organisation and the National Institute for Health and Clinical Excellence (NICE). OSIRIS is a weighted combination of those clinical risk factors that are known to independently predict whether a patient has osteoporosis. These indices, however, have not gained universal acceptance as a routine diagnostic test because of their poor specificity in detecting patients at increased risk of osteoporosis.[3,4] Harrison and Adams[5] found that clinical risk indices misclassified unacceptably large numbers of osteoporotic women, with consequent decreased cost effectiveness. Recently, we reported the use of mandibular cortical width measurements on dental panoramic radiographs (DPRs) as an alternative method of identifying patients with osteoporosis.[6,7] The underlying rationale for this is the enormous number of DPRs taken in dental practice. We propose a strategy of the dentist referring individuals with a thin mandibular cortex and other clinical risk factors for further DXA investigation. Detection of a thinned cortex on DPRs using specially developed computer software[8] has been found to be a good predictor of systemic osteoporosis, and because the method is automatic it is also convenient.

This study has two aims. The first was to determine the diagnostic efficacy of combining the OSIRIS clinical index with the cortical width measurement

[1]*School of Dentistry, University of Manchester; [2,3,7,8]Imaging Science and Biomedical Engineering, University of Manchester; [4]Oral Imaging Centre, Katholieke Universiteit, Belgium; [5]Oral Diagnosis and Radiology, Dental School of Athens; [6]Faculty of Odontology, Malmo University, Sweden; [9,11]School of Dentistry, University of Manchester; [10]ACTA, The Netherlands
*Correspondence to: Dr Hugh Devlin
Email: hugh.devlin@manchester.ac.uk

Refereed Paper
Accepted 14 November 2008
DOI: 10.1038/sj.bdj.2008.317
©British Dental Journal 2008

© 2008 Nature Publishing Group

on radiographs, using multivariate statistical analysis. The second aim was to evaluate the cost effectiveness of using this combined test as a basis for further referral for central dual energy X-ray absorptiometry (DXA), using the 90% sensitivity and 90% specificity values as thresholds.

METHODS

This work forms part of the OSTEODENT study, a collaborative project funded by the European Commission Fifth Framework Programme 'Quality of Life and Management of Living Resources'. The methodology of subject recruitment and examination has been fully described previously,[7] and is summarised here. With Ethics Committee approval, female subjects (aged 45-70 years) were recruited consecutively into the study following their informed consent. Subjects were recruited from each centre using publicity material and by word-of-mouth, but this patient group may not be representative of a primary dental care population. The study included all female volunteers and patients in this age group, with subjects excluded only if they suspected that they might be pregnant. All subjects were interviewed and provided information about their age, weight, medication and fracture history. Patients were recruited from centres in Leuven (Belgium), Athens (Greece), Manchester (UK), and Malmo (Sweden). Six hundred and seventy-one subjects were recruited into the study. The bone mineral density of the total hip could not be measured in two subjects and a further eight subjects were found to be aged less than 45 years, so their data were not included in any further analysis. A further nine radiographs were either lost, digitally corrupted or of poor diagnostic quality. The remaining 652 subjects formed the study population and underwent a clinical risk assessment of osteoporosis using the OSIRIS questionnaire (OSteoporosis Index of RISk) and computer cortical width measurement. One hundred and forty subjects had osteoporosis involving at least one measurement site.

Dental radiographs

Each subject underwent a dental panoramic radiographic examination while biting on a spherical, steel ball bearing (3.175 mm diameter), used to calculate the image magnification. The Leuven (Belgium) and Malmo (Sweden) centres used a Cranex III (Soredex, IL, USA) dental panoramic radiography machine whereas Athens (Greece) and Manchester (UK) used a Planmeca (Planmeca USA, Roselle, IL, USA). In Leuven, a photostimulable phosphor plate system for image capture and digital read out was used, but other centres used a conventional film/cassette. Typical imaging parameters for panoramic radiography were 70 kV at 8 mA for 15 s. All of the radiographs were digitised using a Kodak LS85 digitiser (Eastman Kodak, Rochester, NY) at a resolution of 25.64 pixels/mm.

The mandibular cortex was automatically detected on the digitised panoramic radiographs using software based on Active Shape Model search[9], which is a sophisticated computer imaging technique. Its width was measured by the method described by Allen *et al.*[8]

Table 1 Cortical width and OSIRIS data from 652 subjects, 140 of whom were osteoporotic at one of the measurement sites

	Osteoporotic (n = 140)		Normal (n = 512)	
	OSIRIS	Cortical width	OSIRIS	Cortical width
Mean (SD)	-0.24 (2.5)	3.0 (0.6)	3.40 (2.9)	3.7 (0.6)
Maximum	6.6	5.1	14	5.8
Minimum	-5.9	1.8	-4.6	2.2

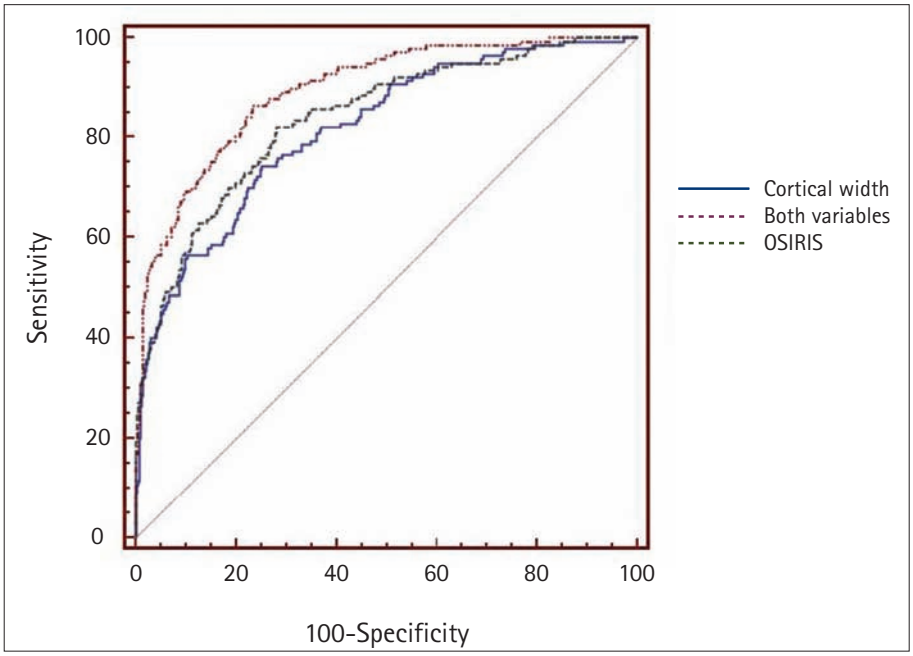


Fig. 1 ROC curve of OSIRIS, cortical width measurements on radiograph, and the effect of combining both variables

DXA Examination

Central DXA of the proximal femur and lumbar spine was performed at each of the four centres. The World Health Organisation (WHO) criteria were used to diagnose osteoporosis, ie using DXA to identify those with a bone mineral density T-score value 2.5 SD or more below the mean value of the young sex matched reference population at any of the lumbar spine, femoral neck or total hip measurement sites. This was used as a 'gold-standard' measure of osteoporosis.

Osteoporosis Index of Risk (OSIRIS)

The Osteoporosis Index of Risk (OSIRIS) is based on four variables:[10] age, body weight, current hormone replacement therapy (HRT) use and history of previous low impact fracture. The index is calculated by adding together:

- Age multiplied by -2 (rounded down to the nearest integer)
- Weight in kg multiplied by 2 (rounded down to the nearest integer)

© 2008 Nature Publishing Group

- +2 if a current user of HRT
- -2 if a history of low trauma fracture.

An OSIRIS score of lower than -3 indicates a high risk of low BMD, between +1 and -3 an intermediate risk and greater than +1.0 a low risk.[11]

Statistical analysis

Student's t-test was used to analyse the significance of the differences between OSIRIS and cortical width values in the osteoporotic and normal individuals.

Discriminant analysis, using the variables cortical width and OSIRIS score, was used to derive the probability of osteoporosis in an individual subject. The model was evaluated using a leave-one-out cross validation strategy to avoid biasing the estimates of discrimination ability. Models were calculated from the entire data set except one, which was used as the test datum. This process was repeated using each of the patients in turn as a test datum. The calculated probability of osteoporosis from each of the experiments was used to generate an ROC curve. The resulting area under the ROC curve was compared with that of OSIRIS and the cortical width measurements. The 90% sensitivity and 90% specificity thresholds were used to calculate the numbers of subjects correctly and incorrectly classified as osteoporotic.

Cost effectiveness

Two strategies were compared to determine those who should receive further investigation using dual energy X-ray absorptiometry. Criteria values for the combined cortical width and clinical analysis used either (a) 90% sensitivity or (b) 90% specificity. In calculating the cost per patient correctly diagnosed with osteoporosis, the costs used for central DXA were £50 per patient and for the OSIRIS index were £5 per patient. Both of these costs estimates have been used recently in other publications by our research group.[5]

RESULTS

The mean difference in mandibular cortical width between osteoporotic and healthy patients (0.718 mm) was highly significant (t = 12.83, p <0.0001) (Table 1). The difference between osteoporotic and normal patients' OSIRIS indices (3.62) was also highly significant (t = 13.54, p <0.0001).

Discriminant analysis was used to obtain a linear combination of weighted average of cortical width and OSIRIS variables that resulted in the best separation between those with and without osteoporosis in our sample. The resulting discriminant score was used to distinguish between the two groups. Wilk's lamba (the ratio of the within-group sum of squares to the total sum of squares) was 0.683 (χ^2 = 247.6, df = 2, p <0.001). Therefore the two groups (those with and without osteoporosis) differed in their mean discriminant score, and 79% of cases were correctly assigned to the groups. Both cortical width and OSIRIS variables contributed equally to the prediction of group membership because they had similar standardised regression coefficients (OSIRIS = 0.682, cortical width = 0.634).

A model derived from a sample will usually fit it better than another sample obtained from the same population. In further leave-one-out cross validation analysis, each case was classified using all the other data to derive the function except that one. A similar but less biased estimate of the correct classification rate of 78.8% of cross-validated grouped cases was obtained.

Using this model, the correlations between the probability of osteoporosis

Table 2 The numbers (and % of the total sample) referred for DXA using a threshold of 90% sensitivity. This guaranteed that the majority of patients with osteoporosis would be referred for further DXA examination, but 164 (56.6%) of the referred patients would have a normal BMD. In total, 178 (or 27.3%) of the 652 patients were misclassified

	No referral for DXA	Refer for DXA	Total
Osteoporosis present	14	126	140
	2.10%	19.30%	
Osteoporosis absent	348	164	512
	53.40%	25.20%	
Total	362	290	652
	55.50%	44.50%	

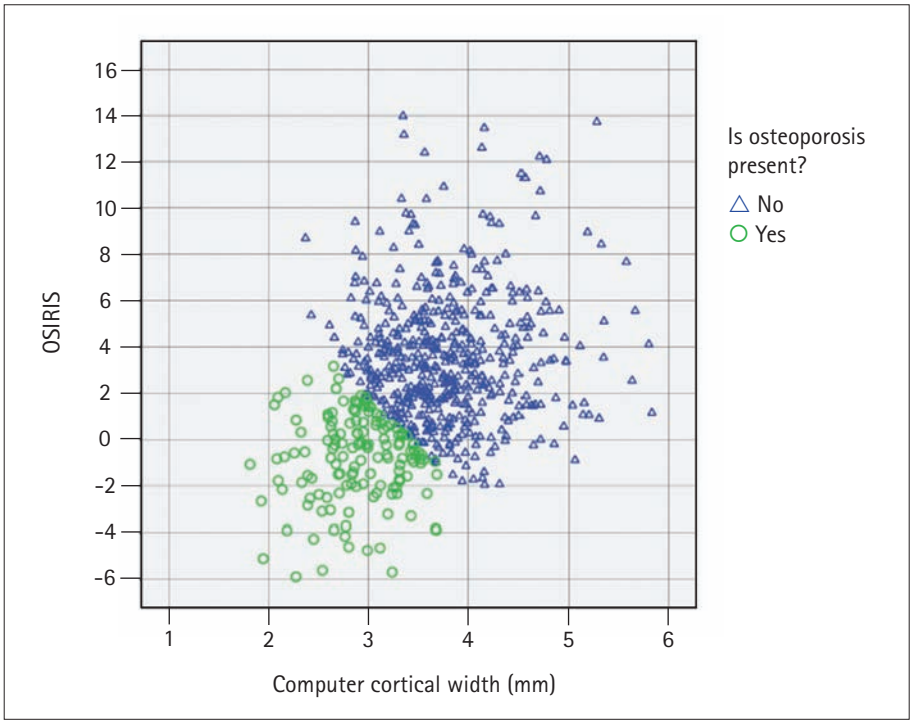


Fig. 2 The variables OSIRIS and cortical width are plotted. The 90% specificity value has been used as a threshold and those who are or are not indicated for referral for central DXA are indicated (ie predicted probability of osteoporosis of either less than or greater than 34.5%)

© 2008 Nature Publishing Group

and the BMD at the lumbar spine, femoral neck and total hip were -0.60, -0.64 and -0.61, respectively. These correlations were highly significant, p <0.01.

The area under the ROC curve (Fig. 1) recorded separately for OSIRIS was 0.84 (95% CI = 0.81 to 0.87) and for the cortical width was 0.82 (95% CI 0.79 to 0.85). The difference between both ROC curve areas was 0.021, which was not significantly different (95% CI = -0.022 to 0.064), p = 0.335. The area under the ROC curve for the predicted probability produced by the linear discriminant analysis in the cross validation experiment (cortical width and OSIRIS) was 0.90 (95% CI = 0.87 to 0.92). There was a significant improvement in the diagnostic ability of the combined OSIRIS and cortical width test over both tests applied separately (p <0.001).

Using the combined OSIRIS and cortical width data, an 'Osteodent Index' was calculated giving the risk of osteoporosis. An operating point on the ROC curve with a specificity value of 90% (95% CI = 87.1 to 92.5) and corresponding value of sensitivity of 69% (95% CI = 60.2 to 76.1) was selected. By using a high specificity value, at the expense of sensitivity, the minimum number of patients would be sent for unnecessary further investigations. Using this criterion value gave a test with a positive likelihood ratio of 6.9 and negative likelihood ratio of 0.35. The diagnostic odds ratio, the ratio of positive likelihood ratio divided by negative likelihood ratio, was 19.7.

The OSIRIS values were plotted against cortical width for the sample of 652 patients. Using the 90% specificity threshold for the combined variable, the sample was divided into those predicted as being at either high or low risk of osteoporosis. Figure 2 shows the scatterplot of OSIRIS index and cortical width, with assignment to either high or low risk of osteoporosis. Our previous work[12] has shown that the optimal decision boundary of whether to further refer patients lies at a 3 mm cortical width. Figure 2 shows that in a patient with a 3 mm cortical width, only when the OSIRIS value is greater than 1.83 is referral not indicated.

There is some overlap of osteoporotic and non-osteoporotic OSIRIS and cortical width values (Fig. 3). Those patients with osteoporosis tend to be grouped towards the lower values of both parameters.

Table 2 shows the false positive and false negative assignments arising from a referral decision at 90% sensitivity for the combined cortical width and OSIRIS data. The corresponding specificity was 68% and the diagnostic odds ratio was 18.73. While using a high sensitivity ensures that only 10% of osteoporotic patients would fail to be referred for further investigation, 56.6% of those referred would have a normal BMD. The cost of this strategy was £141 per osteoporotic patient diagnosed.

An alternative decision strategy is to adopt a high specificity operating point (Table 3). Using a threshold of 90% specificity, results in a referral of only 50 (10%) of those with a normal BMD, but with the disadvantage that 44 out of the 140 osteoporotic patients (31.4%) would be missed. The cost of this strategy was £110 per osteoporotic patient diagnosed.

Table 3 The numbers (and % of the total sample) referred for DXA using a threshold of 90% specificity. Only 50 (or 10%) of patients with normal BMD would be referred for further DXA examination. In total, 94 (or 14.4%) of the 652 subjects were misclassified

	No referral for DXA	Refer for DXA	Total
Osteoporosis present	44	96	140
	6.70%	14.70%	
Osteoporosis absent	462	50	512
	70.90%	7.70%	
Total	506	146	652
	77.60%	22.40%	

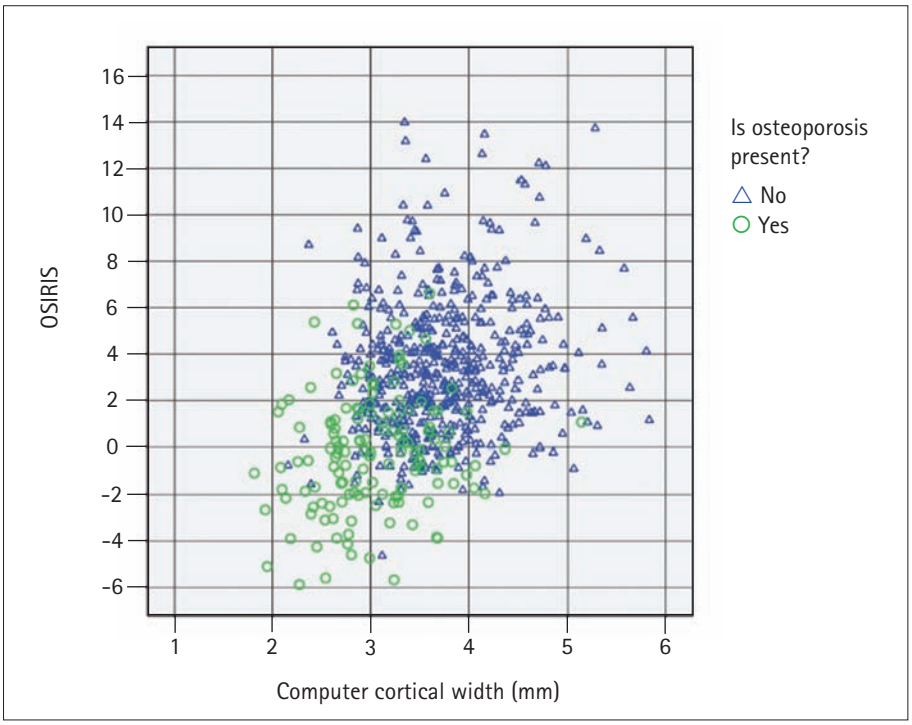


Fig. 3 Cortical width measurements plotted against OSIRIS, with each point represented as either osteoporotic or normal according to DXA. While osteoporotic subjects tend to group towards low values of both parameters, there is extensive overlap between the two groups

DISCUSSION

In this study, we have described a case-finding strategy, where a combination of a clinical index (OSIRIS) and automatically measured width of mandibular cortex, is a technique with good diagnostic accuracy in predicting low bone mineral density at the hip or spine. The diagnostic odds ratio, the ratio of positive likelihood ratio divided by negative likelihood ratio, measures the performance of a test and a value above 20

© 2008 Nature Publishing Group

indicates a diagnostic test with strong evidence for efficacy.[13] Our test, providing a diagnostic odds ratio of 19.7, falls into this category. Furthermore, the fact that the combined test performed better than either the clinical or the radiological test alone demonstrates that they are not providing the same information, but rather complementary information. It is therefore worthy of further clinical trial. Other case-finding strategies that have combined the information from clinical risk factors and selective use of BMD have also proven to be more successful in identifying high-risk patient groups.[14] Clinical risk factors have been shown to independently predict hip fracture risk,[15] but the balance between a simply administered assessment is often at odds with the requirement for a comprehensive assessment of all possible risk factors. Therefore, we plan to further assess the ability of our diagnostic strategy to predict hip fracture in our target population of young post-menopausal women.

In the United States and other industrialised nations, patients are increasingly aware of the benefits of disease prevention, as well as the long-term cost saving with healthcare.[16] What then could be the clinical role for this form of testing for low bone density? Our detection strategy would be used to select those who would undergo further DXA, a case finding approach that follows the UK Royal College of Physicians Guidelines.[17] Millions of dental radiographs are taken by dentists annually, with dental radiographs accounting for nearly 25% of all medical radiographic exposures.[18] Using dentists to select postmenopausal women at high risk of osteoporosis has the advantage that patients are seen regularly from this age group, and that there is an increasing use of radiographs by dentists for diagnosis.[19]

The diagnostic efficacy of the radiographic test alone makes it clearly unsuitable as a screening test, because it performs no better than the simple clinical risk assessment; it also has greater cost and an associated X-ray exposure. We have previously suggested that dentists should refer patients for DXA opportunistically using DPRs that they have taken for the usual dental purposes. Would the combined test be justifiably used as a screening test, at least in the age group of women examined in this study? The decision partly depends on the cost effectiveness of this strategy.

With wide scale dental radiographic and clinical identification of osteoporotic patients, increased healthcare costs in the short-term would be inevitable because of increased demand for DXA services. Europe is already under-resourced for central DXA machines.[20] Our study cannot predict the numbers of femoral neck fractures prevented if our methodology was introduced, but age and a history of previous fracture (which contribute to the OSIRIS index) and a low femoral neck BMD are clinical risk factors which play a role in femoral fracture risk.[21,22]

In the UK, the National Osteoporosis Society has recommended that postmenopausal women given peripheral X-ray absorptiometry be classified into three risk categories.[23] In the first group at high risk of osteoporosis, treatment is recommended, particularly if accompanied by other risk factors. In the second group, the patients are referred for central DXA for further confirmation, and in the third group no additional action is recommended. In this context, we developed an analogous strategy using the combined cortical width and OSIRIS variables to categorise patients into three groups of differing osteoporosis risk. Two thresholds were chosen based on the 90% sensitivity (the low threshold) and the 90% specificity values (the high threshold), as described by Harrison and Adams.[5] The 90% sensitivity value was the predicted probability value of osteoporosis represented by the 10^{th} percentile and the 90% specificity value was the predicted value of non-osteoporosis represented by the 90^{th} percentile. However, the value of this whole approach is limited by the medical side effects and the high cost of providing long-term drug therapy for those in the highest risk group when they do not have osteoporosis.

We therefore compared using either threshold values of 90% sensitivity or 90% specificity for the combined cortical width and OSIRIS variable as two strategies of referral for DXA for those patients considered 'at risk' of osteoporosis. Using a test with 90% specificity (£110 per diagnosed osteoporotic patient) provided a more cost effective option than using 90% sensitivity (£141 per osteoporotic patient). This is due to the large number of patients with normal bone mineral density that were referred unnecessarily for DXA, and therefore the comparatively low yield of osteoporotic patients. There was a cost of £233 per diagnosed osteoporotic patient if all patients in the study received central DXA. This analysis ignores the costs associated with the undiagnosed osteoporotic patients as these were difficult to define in this study. The authors hope that some of these individuals would be identified through further opportunistic testing using either our methodology on a future occasion or other techniques, such as dual energy X-ray absorptiometry or quantitative ultrasound. We also hope to use our test methodology and any subsequent patient treatment to examine the incremental cost-effectiveness ratio per quality adjusted life year. In addition, if the net benefit of our methodology is to be assessed, the distribution of risk assessment cost over the population must be calculated.

Case-finding strategies are prone to operator variability and error. For example, evidence from chest radiography taken in an emergency hospital department has shown that only 25% of patients with radiologically evident vertebral fractures received a diagnosis of osteoporosis or any treatment.[24] Our own research using observer measurements of cortical width has demonstrated that the weakness lies in observer variability. One can postulate that dentists in a primary care setting would be inaccurate in making some measurements by hand.[7,25] Our methodology involves a computer-measured mandibular cortical width to make the initial diagnosis of a high risk of osteoporosis, and following consultation with the patient, the dentist can either provide follow-up clinical questions, such as an OSIRIS clinical risk index, or refer to a specialist.

The cost effectiveness of our proposed methodology is dependent on the prevalence of osteoporosis in our study population (21.5%) being comparable to that

© 2008 Nature Publishing Group

of the UK population. In a study based in Hull, UK, the prevalence of osteoporosis of the hip and spine in general practice was similar (24%), but their study population consisted of women in their seventh decade.[26] The mean age of our study sample was 55 years (sd = 6.1), and the age range (45-70 years) was chosen to test an asymptomatic population which was more representative of young postmenopausal women attending a general dental practice. Our osteoporotic screening method could be made more cost effective if restricted to an elderly population; fewer misdiagnoses are then likely because the incidence of vertebral and hip fractures increases exponentially with age.[27]

Poor radiographic technique is common in general dental practice and may limit the usefulness of our technique. For example in a study by Rushton *et al.*,[28] the image of the lower border of the mandible was at least partially absent in 9% of panoramic radiographs. Using digital panoramic radiographs will approximately halve the reject rate of films as about half of faults are due to chemical processing of film.[28] Patient positioning faults could be reduced by using better positioning aids, further training of dentists with an emphasis on quality assurance, and using only suitably qualified personnel to take radiographs.[28]

In conclusion, our methodology used computer software to detect and analyse the mandibular cortical width and when combined with clinical risk indices data detected patients at early, high risk of osteoporosis.

This work was supported by a research and technological development project grant from the European Commission Fifth Framework Programme 'Quality of Life and Management of Living Resources' (QLK6-2002-02243; 'OSTEODENT').

1. Dharmarajan T S, Banik P. Hip fracture. Risk factors, preoperative assessment, and postoperative management. *Postgrad Med* 2006; **119:** 31-38.
2. Peng E W, Elnikety S, Hatrick N C. Preventing fragility hip fracture in high risk groups: an opportunity missed. *Postgrad Med J* 2006; **82:** 528-531.
3. Ben Sedrine W, Broers P, Devogelaer J-P, Depresseux G *et al.* Interest of a prescreening questionnaire to reduce the cost of bone densitometry. *Osteoporos Int* 2002; **13:** 434-442.
4. Pongchaiyakul C, Nguyen N D, Eisman J A, Nguyen T V. Clinical risk indices, prediction of osteoporosis, and prevention of fractures: diagnostic consequences and costs. *Osteoporos Int* 2005; **16:** 1444-1450.
5. Harrison E J, Adams J E. Application of a triage approach to peripheral bone densitometry reduces the requirement for central DXA but is not cost effective. *Calcif Tissue Int* 2006; **79:** 199-206.
6. Devlin H, Allen P D, Graham J, Jacobs R *et al.* Automated osteoporosis risk assessment by dentists: a new pathway to diagnosis. *Bone* 2007; **40:** 835-842.
7. Karayianni K, Horner K, Mitsea A, Bourkas L *et al.* Accuracy in osteoporosis diagnosis of a combination of mandibular cortical width measurement on dental panoramic radiographs and a clinical risk index (OSIRIS): the OSTEODENT project. *Bone* 2007; **40:** 223-229.
8. Allen P D, Graham J, Farnell D J J, Harrison E J *et al.* Detecting reduced bone mineral density from dental radiographs using statistical shape models. *IEEE Trans. IT in Biomed* 2007; **11:** 601-609
9. Cootes T F, Taylor C J, Cooper D H, Graham J. Active shape models - their training and application. *Comput Vis Image Underst* 1995; **61:** 38-59.
10. Sedring W B, Chevallier T, Zegels B, Kvasz A *et al.* Development and assessment of the Osteoporosis Index of Risk (OSIRIS) to facilitate selection of women for bone densitometry. *Gynecol Endocrinol* 2002; **16:** 245-250.
11. Richy F, Gourlay M, Ross P D, Sen S S *et al.* Validation and comparative evaluation of the osteoporosis self-assessment tool (OST) in a Caucasian population from Belgium. *Q J Med* 2004; **97:** 39-46.
12. Devlin H, Horner K. Mandibular radiomorphometric indices in the diagnosis of reduced skeletal bone mineral density. *Osteoporos Int* 2002; **13:** 373-378.
13. Deeks J J. Systematic reviews of evaluations of diagnostic and screening tests. *In* Egger M, Smith G D, Altman D G (eds). *Systematic reviews in health care: meta-analysis in context*, 2nd ed. pp 255-256. London: BMJ Publishing Group, 2001.
14. Johansson H, Oden A, Johnell O, Jonsson B *et al.* Optimization of BMD measurements to identify high risk groups for treatment – a test analysis. *J Bone Miner Res* 2004; **19:** 906-913.
15. van Staa T P, Geusens P, Kanis J A, Leufkens H G M *et al.* A simple clinical score for estimating the long-term risk of fracture in post-menopausal women. *Q J Med* 2006; **99:** 673-682.
16. Leigh J P, Hubert H B, Romano P S. Lifestyle risk factors predict healthcare costs in an aging cohort. *Am J Prev Med* 2005; **29:** 379-387.
17. Kayan K, De Takats D, Ashford R, Kanis J A, McCloskey E V. Performance of clinical referral criteria for bone densitometry in patients under 65 years of age assessed by spine bone mineral density. *J Postgrad Med* 2003; **79:** 581-584.
18. Brown J E. Advances in dental imaging. *Primary Dent Care* 2001; **8:** 59-62.
19. Gibbs S J. Biological effects of radiation from dental radiography. Council on Dental Materials, Instruments, and Equipment. *J Am Dent Assoc* 1982; **105:** 275-281.
20. Kanis J A, Johnell O. Requirements for DXA for the management of osteoporosis in Europe. *Osteoporos Int* 2005; **16:** 229-238.
21. Johnell O, Kanis J A, Oden A, Johansson H *et al.* Predictive value of BMD for hip and other fractures. *J Bone Miner Res* 2005; **20:** 1185-1194.
22. Kanis J A. Diagnosis of osteoporosis and assessment of fracture risk. *Lancet* 2002; **359:** 1929-1936.
23. National Osteoporosis Society Position statement on the use of peripheral X-ray absorptiometry in the management of osteoporosis. Bath, England, 2004.
24. Majumdar S R, Kim N, Colman I, Chahal A M *et al.* Incidental vertebral fractures discovered with chest radiography in the emergency department: prevalence, recognition, and osteoporosis management in a cohort of elderly patients. *Arch Intern Med* 2005; **165:** 905-909.
25. Devlin C V, Horner K, Devlin H. Variability in measurement of radiomorphometric indices by general dental practitioners. *Dentomaxillofac Radiol* 2001; **30:** 120-125.
26. Ballard P A, Purdie D W, Langton C M, Steel S A, Mussurakis S. Prevalence of osteoporosis and related risk factors in UK women in the seventh decade: osteoporosis case finding by clinical referral criteria or predictive model? *Osteoporos Int* 1998; **8:** 535-539.
27. Kanis J A, Pitt F A. Epidemiology of osteoporosis. *Bone* 1992; **13 (Suppl 1):** S7-S15.
28. Rushton V E, Horner K, Worthington H V. The quality of panoramic radiographs in a sample of general dental practices. *Br Dent J* 1999; **186:** 630-633.

© 2008 Nature Publishing Group

40. **The relationship between the OSTEODENT index and hip fracture risk assessment using FRAX.** K. Horner, P. Allen, J. Graham, R. Jacobs, S. Boonen, S. Pavit, O. Naeckerts, E. Marjanovic, J.E Adams, K. Karayianni, C. Lindh, P. van der Stelt, H. Devln, *Oral Surgery Oral Medicine Oral Pathology Oral Radiology And Endodontology*. ***110(2)****: 243-249, 2010.* doi: 10.1016/j.tripleo.2010.03.035

The relationship between the OSTEODENT index and hip fracture risk assessment using FRAX

Keith Horner, BChD, MSc, PhD, FDSRCPS Glasg, FRCR, DDR,[a] Philip Allen, BSc, PhD,[b] Jim Graham, BSc, PhD,[b] Reinhilde Jacobs, LDS, MSc, PhD,[c] Steven Boonen, MD, PhD,[d] Susan Pavitt, BSc, PhD,[e] Olivia Nackaerts, MSc, PhD,[c] Elizabeth Marjanovic, BSc, MSc, PhD,[b] Judith E. Adams, MBBS, FRCR, FRCP,[b] Kety Karayianni, DDS, PhD,[f] Christina Lindh, DDS, Odont Dr,[g] Paul van der Stelt, DDS, PhD,[h] and Hugh Devlin, BDS, BSc, MSc, PhD,[a] Manchester and Leeds, UK, Leuven, Belgium, Athens, Greece, Malmo, Sweden, and Amsterdam, The Netherlands

UNIVERSITY OF MANCHESTER, KATHOLIEKE UNIVERSITEIT, UNIVERSITY OF LEEDS, UNIVERSITY OF ATHENS, MALMO UNIVERSITY, AND ACADEMIC CENTRE FOR DENTISTRY

Objectives. The OSTEODENT index is a predicted probability of osteoporosis derived from a combination of an automated analysis of a dental panoramic radiograph and clinical information. This index has been proposed as a suitable case-finding tool for identification of subjects with osteoporosis in primary dental care; however, no data exist on the relationship between OSTEODENT index and fracture risk. The aims of this study were to assess the relationship between the OSTEODENT index and hip fracture risk as determined by FRAX and to compare the performance of the OSTEODENT index and FRAX (without femoral BMD data), in determining the need for intervention as recommended in UK national treatment guidance.

Study design. The study was a retrospective analysis of data from 339 female subjects (mean age 55.3 years), from 2 centers: Manchester (UK) and Leuven (Belgium). Clinical information and femoral neck BMD were available for FRAX, and dental panoramic radiographic data and clinical information were available to calculate the OSTEODENT index. Subjects were classified into "treat" or "lifestyle advice and reassurance" categories using the National Osteoporosis Guideline Group (NOGG) threshold.

Results. The OSTEODENT index result was significantly related to the 10-year probability of hip fracture derived from the reference standard FRAX tool (Rs = 0.67, $P < .0001$); 84 patients (24.8%) were allocated to the "treat" category on the basis of FRAX and the UK national guidance. Using this "treatment/no treatment" classification as the reference standard, ROC analysis showed no significant difference between areas under the curves for the OSTEODENT index (0.815) and the 10-year probability of hip fracture derived from the FRAX index without BMD (0.825) when used as tests for determining therapeutic intervention.

Conclusion. The results suggest that the OSTEODENT index has value in prediction of hip fracture risk. Prospective trials are needed to confirm this finding and to examine the feasibility for its use in primary dental care. **(Oral Surg Oral Med Oral Pathol Oral Radiol Endod 2010;110:243-249)**

A major challenge in managing osteoporosis is the difficulty in identifying affected individuals before the condition is established and fracture has occurred.[1-3]

Dental radiological examination can be used for osteoporosis risk assessment and a substantial number of patients undergo dental panoramic radiographic examinations each year.

In Manchester, we acknowledge the support of the University of Manchester, Manchester Academic Health Science Centre, and the NIHR Biomedical Research Centre. S.B. is senior clinical investigator of the Fund for Scientific Research, Flanders, Belgium (F.W.O.-Vlaanderen) and holder of the Leuven University Chair in Metabolic Bone Diseases. This work was supported by a research and technological development project grant from the European Commission Fifth Framework Programme "Quality of Life and Management of Living Resources" (QLK6-2002-02243; "OSTEODENT").

[a]School of Dentistry, University of Manchester, Manchester Academic Health Science Centre, Manchester, UK.
[b]Imaging Science and Biomedical Engineering, Manchester Academic Health Sciences Centre, University of Manchester, UK.
[c]Oral Imaging Centre, School of Dentistry, Oral Pathology and Maxillofacial Surgery, Katholieke Universiteit, Leuven, Belgium.
[d]Division of Gerontology and Geriatrics & Center for Musculoskeletal Research, Department of Experimental Medicine, Katholieke Universiteit Leuven, Belgium.
[e]Comprehensive Health Research Division, Clinical Trials Research Unit (CTRU), Leeds Institute of Molecular Medicine, University of Leeds, UK.
[f]Dental School, University of Athens, Greece.
[g]Faculty of Odontology, Malmo University, Sweden.
[h]Academic Centre for Dentistry, Amsterdam, The Netherlands.

Received for publication Feb 2, 2010; returned for revision Mar 19, 2010; accepted for publication Mar 19, 2010.
1079-2104/$ - see front matter
© 2010 Mosby, Inc. All rights reserved.
doi:10.1016/j.tripleo.2010.03.035

Diagnosis of osteoporosis relies on measurement of "areal" bone mineral density[4] (g/cm^2; BMD_a) at the hip and spine by dual energy x-ray absorptiometry (DXA), with classification being based on standard deviation (SD) scores (T score at or below –2.5).[1] BMD_a is, however, only one risk factor for fracture, with a doubling of risk for each SD reduction in BMD_a.[5,6] Furthermore, some of the clinical risk factors used as indicators for BMD_a measurement are themselves associated with a fracture risk greater than can be accounted for by BMD_a alone.[7] Factors such as a low body mass index, low milk consumption, lack of sunlight exposure, and low physical activity account for about half of hip fractures, and which are reversible by the patient.[8] In recent years, therefore, emphasis has shifted away from diagnosis of osteoporosis by BMD_a toward assessment of clinical fracture risk. Early identification of risk factors for osteoporosis and fracture may facilitate access to appropriate medical intervention and fracture risk reduction. Such a "case-finding" approach, based on clinical risk factors, has been recommended rather than population screening. Using meta-analysis techniques applied to studies on population-based cohorts that identified clinical risk factors for fracture, algorithms (FRAX) have been developed that compute age-specific 10-year fracture probabilities.[9] FRAX can be used without the availability of femoral neck BMD_a, or with its incorporation into the algorithm when available.[9-11] In the United Kingdom, a management strategy guideline, based on an individual's estimated fracture risk, has been devised by the National Osteoporosis Guideline Group (NOGG),[11] providing a clear pathway from clinical risk assessment using FRAX through to appropriate guidance on intervention.

Approximately 1 in 3 of all radiological examinations are made by dentists.[12] Dental radiographs invariably show images of the bone of the jaws and there is evidence that jaw BMD_a and radiomorphometric indices correlate significantly with BMD_a of other skeletal sites, including the hip and spine.[13-16] Subsequent work has demonstrated that various radiographic features of mandibular bone on panoramic radiographs have potential value in predicting BMD_a at these important sites of fracture. The OSTEODENT study established that an "automatic" measurement of mandibular cortical bone thickness on panoramic radiographs was a valid test for diagnosis of osteoporosis in women aged 45 to 70 years, with a receiver operating characteristic (ROC) area exceeding 0.80.[17] Subsequent data analysis showed that combining the radiographic mandibular cortical width data with clinical information, in the form of the Osteoporosis Index of Risk (OSIRIS)[18] (age, weight, current estrogen therapy, and history of low trauma fracture), produced a test result (the "OSTEODENT index") that was significantly better for prediction of BMD_a than either test alone (ROC curve area = 0.90).[19] The authors suggested that combining these clinical and radiographic tests had potential to be used in primary dental health care as a case-finding strategy for identification of patients at risk of osteoporosis.

Although the OSTEODENT index, along with other dental radiographic measurements, has been shown to have diagnostic validity in predicting BMD_a and osteoporosis, there is no information on how the index relates to risk of fracture or to patient management decisions. If the OSTEODENT index is to be a tool for case finding in osteoporosis, then its performance should be comparable with that of the FRAX tool (without inclusion of BMD_a data) in assessment of fracture risk and management recommendations.

The aims of this study, therefore, were to (1) assess the relationship between the OSTEODENT index and hip fracture risk as determined by the FRAX tool, and (2) compare the performance of 2 clinical tests, the OSTEODENT index and the FRAX tool (*without* BMD_a data), to determine appropriate intervention as recommended by NOGG.

MATERIALS AND METHODS

This study was carried out by retrospective analysis of patient data from the OSTEODENT study (European Commission Fifth Framework Programme "Quality of Life and Management of Living Resources"; QLK6-2002-02243). The aim of that study was to identify the most valid and effective radiographic index, or combination of radiographic and clinical indices, for the diagnosis of osteoporosis applicable for use by dentists in a primary health care setting. Details of the study have been reported in full previously[20] and a summary of pertinent aspects are reported here.

Women in the age range of 45 to 70 years were invited to participate in the study, recruited at 4 centers: Athens (Greece), Leuven (Belgium), Malmö (Sweden), and Manchester (UK). Local ethical approval for the study was obtained in each recruiting center and informed consent was obtained from all subjects. The racial origin of the patients, their menopausal status, and history of hysterectomy and hormone replacement therapy (HRT), if applicable, were noted. Weight (wt) and height (ht) were measured and body mass index (wt/ht;^{2} kg/m^2) calculated. Information about menopausal status, date of menarche, previous fracture history, tobacco-smoking habits, and alcohol intake were recorded. Recruitment was performed over a 24-month period extending from October 2003 to September 2005. For the study reported here, subjects from only 2 of the recruiting centers (Leuven and Manchester) are included in the analysis. Subjects from Athens and

Malmö were excluded because no record of glucocorticoid use (required for use of the FRAX tool) existed.

Bone densitometry

DXA of the left hip was carried out on each subject to determine femoral neck (FN) BMD_a. Scans were performed by experienced radiographers on the Hologic QDR 4500 (Hologic Inc., Bedford, MA) in Leuven and the Hologic Discovery (Hologic Inc.) in Manchester. T-scores were calculated using National Health and Nutrition Examination Survey (NHANES) reference data.[21] The European spine phantom (ESP)[22] was used to standardize measurements between different manufacturers using the method described by Pearson and colleagues.[23] Standardization of BMD_a measurements was performed by one experienced scientist (E.M.). Scans and results from the 2 centers were reviewed and confirmed for good quality by one clinical radiologist (J.E.A.) with expertise in this field. Shewarts rules were used to monitor quality assurance throughout the study period.[24]

Clinical evaluation

All subjects were interviewed using a standard questionnaire covering a wide range of medical, social, and family history, including age, sex, weight, height, previous fractures, parental hip fracture, current smoking, glucocorticoid use, rheumatoid arthritis, secondary osteoporosis, alcohol intake, and current HRT use.

Radiography

In Manchester, dental panoramic radiography was performed on each subject using a Planmeca PM2002CC (Planmeca Oy, Helsinki, Finland). In Leuven, radiography was carried out using a Cranex 3DC (Soredex, Tuusula, Finland). The imaging parameters varied according to equipment difference and patient variation, but typically were 70 kV (constant potential) at 8 mA for 15 seconds. In Leuven, ADC Solo (Afga, Mortsel, Belgium) was used as the photostimulable phosphor plate system for image capture and read out, whereas Manchester used a conventional film/cassette combination. Procedures for radiographic quality assurance and magnification control have been reported in detail previously.[20] Hard-copy radiographs were digitized using a Kodak LS85 digitizer (Eastman Kodak, Rochester, NY) at a resolution of 25.64 pixels/mm.

Reference standard: NOGG intervention category derived from fracture risk using FRAX

The clinical information collected from subjects was used to calculate the 10-year probability of hip fracture using the FRAX fracture risk assessment tool developed by the World Health Organization (WHO) (http://www.shef.ac.uk/FRAX/, accessed between October 19, 2008, and January 20, 2009). The FRAX tool for the United Kingdom was used for both Manchester and Leuven subjects. FRAX incorporates the following clinical risk factors: age, sex, weight, height, previous fracture, parental hip fracture, current smoking, glucocorticoid use, rheumatoid arthritis, secondary osteoporosis, and alcohol intake exceeding 3 units per day. In the absence of BMD data, subjects are classified into low risk (requiring reassurance), intermediate risk (requiring BMD_a to be ascertained), and high risk (requiring treatment be considered either with or without further BMD_a assessment).

We entered the femoral neck BMD and clinical data into the FRAX tool and, using the NOGG management guidance link (http://www.shef.ac.uk/NOGG) from the FRAX Web site, we calculated the 10-year hip fracture probability. Using the reference standard given there, subjects were classified into either "treat" or "lifestyle advice and reassurance" categories.

Radiographic/clinical test: OSTEODENT index assessment

The OSTEODENT index is the predicted probability of osteoporosis based on a combination of automatic measurement of mandibular cortical width on dental panoramic radiographs and OSIRIS,[18] calculated by a computer program.[17] Briefly, the mandibular cortex was automatically detected on digitized panoramic radiographs using software based on an Active Shape Model search.[25,26] The clinical data required to calculate OSIRIS (age, body weight, current HRT use, and history of previous low-impact fracture) was then entered and combined with the radiographic data, producing the predicted probability (%) of osteoporosis. Each subject's radiographic and clinical information was entered to give her OSTEODENT index.

Clinical test: 10-year probability of hip fracture derived from FRAX without FN BMD_a

Clinical information of subjects was used to calculate the 10-year probability of hip fracture using the FRAX tool (without inclusion of FN BMD_a).

Statistical analysis

The relationship between the OSTEODENT index and the 10-year probability of hip fracture derived from the FRAX tool (with FN BMD_a) was calculated using Spearman's rank correlation. Similarly, the relationship between the 10-year probability of hip fracture derived from the FRAX tool (without FN BMD_a) and the 10-year probability of hip fracture derived from the FRAX tool (with FN BMD_a) was calculated.

The reference standard decision of "treat" or "lifestyle advice and reassurance" derived from 10-year hip fracture risk probability was calculated using FRAX (with FN BMD_a) and NOGG management guidance link from the FRAX Web site and was used as a "gold reference standard" in our calculations. The abilities of FRAX (without FN BMD_a) and the OSTEODENT index in categorizing subjects into the NOGG "treat" category were compared using ROC curve analysis. The areas under the ROC curves (AUC) and analysis for significant statistical differences between AUCs were calculated using the MedCalc programme (MedCalc Software bvba, Mariakerke, Belgium).

RESULTS

A total of 351 women were recruited to the study. It was not possible to record the mandibular cortical width for 7 subjects because these films were unusable for reasons including damage, accidental loss, lack of a ball-bearing image, and unacceptable image quality. Data on hip fracture probability (with FN BMD_a) was not available for a further 5 subjects. Clinical information required to calculate FRAX was lacking in 3 subjects for whether they had a confirmed diagnosis of rheumatoid arthritis, the number of alcohol units consumed per day, and their exposure to oral glucocorticoids. Dentate and edentulous patients were included, as we have shown previously that the dental state of the patient was not significant in predicting mandibular bone mineral density.[27] Several subjects had more than one missing item of data and complete datasets were available for 339 women. The mean age of this population was 55.3 years (SD = 6.32). From this study population, 62 (18%) were classified as having osteoporosis at the femoral neck.

The OSTEODENT index results (predicted probability of having osteoporosis) in the study population ranged from 0% to 99.4%. The data were not normally distributed, with most people having a low score and with a median result of 17.95%. Similarly, most subjects had low 10-year hip fracture probability using FRAX (with or without FN BMD_a). For FRAX (without FN BMD_a), median probability was 0.7%, with a range from 0.1% to 19.0%, whereas for FRAX (with FN BMD_a) the median was 0.6% and the range extended from 0% to 49.0%.

Relationship between the OSTEODENT index and 10-year probability of hip fracture

A significant relationship was demonstrated between the OSTEODENT index and the 10-year probability of hip fracture derived from the reference standard FRAX tool (with FN BMD_a), $R_s = 0.67$, $P < .0001$, indicating a strong relationship between the 2 variables (Fig. 1, *A*).

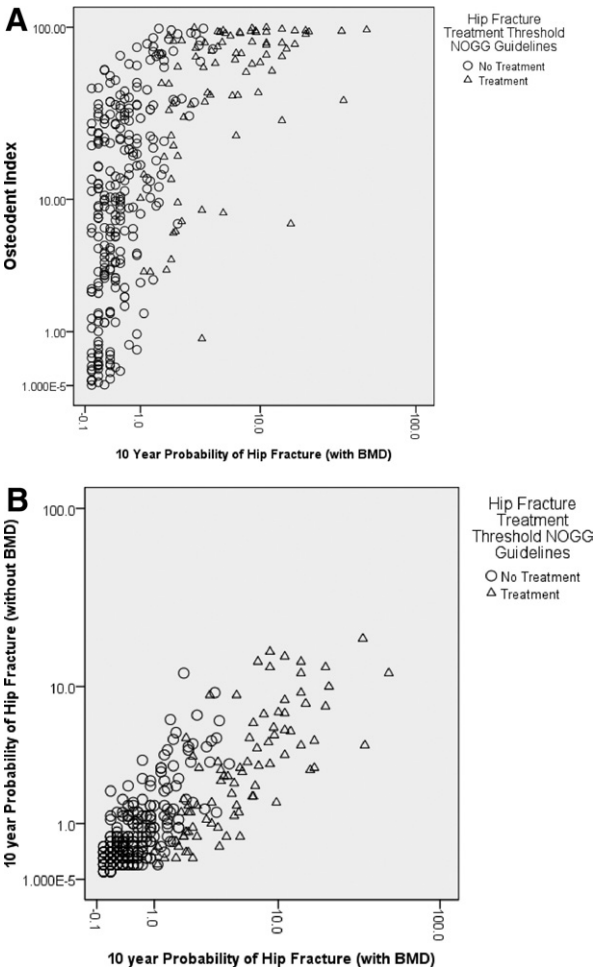


Fig. 1. Scatter plots showing **(A)** the relationship between the OSTEODENT index (%) and FRAX (with FN BMD_a) (10-year probability of hip fracture) and **(B)** the relationship between FRAX (without FN BMD_a) and FRAX (with FN BMD_a) for the 339 subjects in the study. Logarithmic axes are used for both plots.

The relationship between the 10-year probabilities of hip fracture calculated using FRAX (without FN BMD_a) and FRAX (with FN BMD_a) was $R_s = 0.77$ ($P < .0001$; Fig. 1, *B*). The 2 independent correlation coefficients were significantly different from each other ($z = -2.72$, $P = .007$). All the indices were positively skewed. Most subjects had a low 10-year probability of hip fracture (Table I), measured using any of the indices, and the correlation coefficients may therefore have been influenced by outliers.[28]

Eighty-four patients (24.8%) were considered as requiring treatment by NOGG subsequent to their FRAX (with FN BMD_a) assessment. Using this as the reference standard, ROC curve analysis (Fig. 2) showed that the AUC for the OSTEODENT index used as a means

Table I. The value of the different indices below which the stated percentage of subjects were classified

Percentiles	*OSTEODENT index, %*	*FRAX (without BMD_a), %*	*FRAX (with FN BMD_a), %*
Median	17.9	0.7	0.6
20	3.3	0.3	0.2
30	6.4	0.4	0.3
40	10.2	0.5	0.4
60	28.2	1.0	1.0
70	37.9	1.4	1.7
80	62.6	2.6	3.0
90	83.3	4.7	6.8

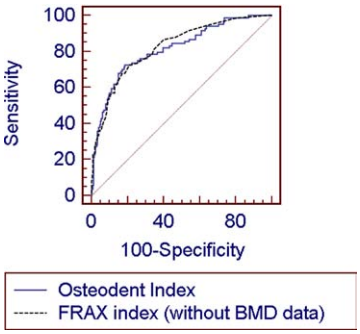


Fig. 2. Receiver operating characteristic curves for the OSTEODENT index and FRAX (without FN BMD_a), demonstrating their value as diagnostic tests for allocating subjects into the NOGG "treat" category. There was no significant difference between areas under the curves.

of determining treatment need was 0.815 (SE = 0.030) and that for the 10-year probability of hip fracture derived from the FRAX index without BMD was 0.825 (SE = 0.029). There was no significant difference in AUC between these 2 curves (z = 0.36, P = .72).

DISCUSSION

Regular visits by patients to the dentist for check-ups and treatment provide an opportunity to address issues that may be of indirect relevance to dental health, but of great importance to general well-being. Thus, many dentists, at least in the United Kingdom, advise on smoking cessation,[29] perform blood pressure checks,[30] and carry out oral cancer screening.[31] Bone health is of immediate importance to dentists, not least in the context of implant and periodontal therapies. The inclusion of opportunistic osteoporosis case-finding by dentists has, therefore, a reasonable prospect of adoption if a suitable test is available and if cost-effectiveness can be demonstrated. The OSTEODENT software, based on radiographic data supplemented by some simple clinical information, has been shown to have good diagnostic validity for identification of women with low BMD_a,[17,19] but there was no information available about its relationship with fracture risk. This study aimed to address this deficiency.

The study design was retrospective, being a reevaluation of data obtained from a previous study, designed before the publication of the WHO FRAX fracture risk assessment tool. Such a design invariably introduces limitations. The number of women included was limited to those from only 2 of the original 4 recruiting centers because all items of data needed for FRAX had not been collected consistently. It should be recognized that the study did not measure actual fracture incidence in a longitudinal cohort study as the number of subjects was too small to provide statistically significant fracture data. Nonetheless, our study provides useful information that may help determine whether a larger, prospective study of the OSTEODENT index and fracture prediction is indicated.

The OSTEODENT index was significantly correlated with 10-year probability of hip fracture derived from FRAX (with FN BMD_a). The OSTEODENT index has previously been shown to have high diagnostic value in prediction of low BMD_a and osteoporosis. As BMD is a risk factor for hip fracture, such a finding is not surprising. The useful information, however, was the indication of the strength of the relationship between the OSTEODENT index and 10-year probability of hip fracture calculated using FRAX (with FN BMD_a). This significant association (R_s = 0.67) was weaker than that (R_s = 0.77) between FRAX (without FN BMD_a) and FRAX (with FN BMD_a), although the stronger relationship with the latter is not surprising in view of the shared elements contributing to their calculation.

Demonstration of a significant association between 2 variables does not address the potential value of the OSTEODENT index as a clinical case-finding test. Calculation of sensitivity and specificity and/or the use of ROC curve analysis do provide such information, but require a reference standard with which the test can be compared. In this study, in the absence of any "true" individual fracture data, we used the 10-year probability of hip fracture obtained from FRAX (with FN BMD_a). To perform ROC analysis, these reference standard data had to be dichotomized and this was achieved using the intervention threshold recommended by NOGG, which can be justified, as the FRAX Web site links directly through to the NOGG management recommendation. The FRAX tool provides an estimate of 10-year probability of both hip and "major" (clinical spine, forearm, hip, or shoulder) fracture. Similarly, when FN BMD_a is included, FRAX leads to NOGG guidance on intervention for both frac-

ture probabilities. In this study, we chose to consider only hip, rather than major, fracture probabilities because the impact of hip fracture on the affected individual is greater.

Our ROC results indicate that the ability of the OSTEODENT index to predict patient management decisions according to NOGG is comparable with that of FRAX without FN BMD_a. The OSTEODENT index is a combination of clinical and radiographic information, whereas FRAX without FN BMD_a is derived from clinical data alone, although there are more clinical items considered by FRAX than by the OSTEODENT index. It seems possible that the radiographic information provided by OSTEODENT is providing some indicator of BMD_a status that compensates for the smaller number of clinical data items included. As such, it is interesting to consider whether adapting the OSTEODENT software to include the FRAX tool factors, rather than the more limited OSIRIS factors, might improve performance as a predictive tool. This should be considered in future research, although the advantage of the clinical data collection used in OSTEODENT (age, weight, current HRT, and history of low trauma fracture) is that it is quick and easy to collect and more feasible for application in dental practice.

Although there are limitations in this study, the OSTEODENT index has potential as a case-finding tool for osteoporosis and as an indicator of hip fracture risk, and a larger prospective trial in primary care seems justified. Such a study should also address the important issues of stakeholder acceptability, possible interprofessional barriers, and an essential economic evaluation.

REFERENCES

1. World Health Organization Study Group. Assessment of fracture risk and its application to screening for postmenopausal osteoporosis. Report of a WHO Study Group. Geneva: World Health Organization; 1994. WHO Technical Report Series, No. 843.
2. European Commission. Report on osteoporosis in the European Community. Strasbourg: European Community; 1998.
3. Royal College of Physicians. Osteoporosis: clinical guidelines for the prevention and treatment. London: Royal College of Physicians; 1999.
4. Engelke K, Glüer CC. Quality and performance measures in bone densitometry: part 1: errors and diagnosis. Osteoporos Int 2006;17:1283-92.
5. Marshall D, Johnell O, Wedell JJ. Meta-analysis of how well measures of bone mineral density predict occurrence of osteoporotic fractures. Br Med J 1996;312:1254-9.
6. Johnell O, Kanis JA, Oden A, Johansson H, De Laet C, Delmas P, et al. Predictive value of bone mineral density for hip and other fractures. J Bone Miner Res 2005;20:1185-94.
7. Kanis JA. Diagnosis of osteoporosis and assessment of fracture risk, Lancet 2002;359:1929-36.
8. Johnell O, Gullberg B, Kanis JA, Allander E, Elffors L, Dequeker J, et al Risk factors for hip fracture in European women: the MEDOS Study. Mediterranean Osteoporosis Study. Bone Miner Res 1995;10:1802-15.
9. Kanis JA, Oden A, Johansson, Borgström F, Ström O, McCloskey E. FRAX and its applications to clinical practice. Bone 2009;44:734-43.
10. Kanis JA, Johnell O, Oden A, Johansson H, McCloskey E. FRAX™ and the assessment of fracture probability in men and women from the UK. Osteoporos Int 2008;19:385-97.
11. Kanis JA, McCloskey EV, Johansson H, Strom O, Borgstrom F, Oden A. Case finding for the management of osteoporosis with FRAX®—assessment and intervention thresholds for the UK. Osteoporos Int 2008;19:1395-1408.
12. Tanner RJ, Wall BF, Shrimpton PC, Hart D, Bungay DR. Frequency of medical and dental x-ray examinations in the UK, 1997-98. National Radiological Protection Board, London, UK: 2000. NRPB-R320.
13. Horner K, Devlin H, Alsop CW, Hodgkinson IM, Adams JE. Mandibular bone mineral density as a predictor of skeletal osteoporosis. Br J Radiol 1996;69:1019-25.
14. Taguchi A, Tanimoto K, Suei Y, Ohama K, Wada T. Relationship between the mandibular and lumbar vertebral bone mineral density at different postmenopausal stages. Dentomaxillofac Radiol 1996;25:130-5.
15. Devlin H, Horner K. Mandibular radiomorphometric indices in the diagnosis of reduced skeletal bone mineral density. Osteoporos Int 2002;13:373-8.
16. White SC, Taguchi A, Kao D, Wu S, Service SK, Yoon D, et al. Clinical andpanoramic predictors of femur bone mineral density. Osteoporos Int 2005;16:339-46.
17. Devlin H, Allen PD, Graham J, Jacobs R, Karayianni K, Lindh C, et al. Automated osteoporosis risk assessment by dentists: a new pathway to diagnosis. Bone 2007;40:835-42.
18. Sedrine WB, Chevallier T, Zegels B, Kvasz A, Micheletti MC, Gelas B, et al. Development and assessment of the Osteoporosis Index of Risk (OSIRIS) to facilitate selection of women for bone densitometry. Gynecol Endrocrinol 2002;16:245-50.
19. Devlin H, Allen P, Graham J, Jacobs R, Nicopoulou-Karayianni K, Lindh C, et al. The role of the dental surgeon in detecting osteoporosis: the OSTEODENT study. Br Dent J 2008;204(10):E16.
20. Karayianni K, Horner K, Mitsea A, Berkas L, Mastoris M, Jacobs R, et al. Accuracy in osteoporosis diagnosis of a combination of mandibular cortical width measurement on dental panoramic radiographs and a clinical risk index (OSIRIS): the OSTEODENT project. Bone 2007;40:223-9.
21. Looker A, Wahner H, Dunn W, Calvo M, Harris T, Heyse SP, et al. Updated data on proximal femur bone mineral levels of US adults. Osteoporos Int 1998;8:468-89.
22. Kalender WA, Felsenberg D, Genant HK, Fischer M, Dequeker J, Reeve J. The European Spine Phantom—a tool for standardization and quality control in spinal bone mineral measurements by DXA and QCT. Eur J Radiol 1995;20:83-92.
23. Pearson J, Dequeker J, Henley M, Bright J, Reeve J, Kalender W, et al. European semi-anthropomorphic spine phantom for the calibration of bone densitometers: assessment of precision, stability and accuracy. The European Quantitation of Osteoporosis Study Group. Osteoporos Int 1995;5:174-84.
24. Orwoll ES, Oviatt SK. Longitudinal precision of dual-energy x-ray absorptiometry in a multicenter study. The Nafarelin/Bone Study Group. J Bone Miner Res 1991;6:191-7.
25. Davies RH, Twining CJ, Cootes TF, Waterton JC, Taylor CJ. A minimum description length approach to statistical shape modeling. IEEE Trans Med Imaging 2002;21:525-37.
26. Allen PD, Graham J, Farnell DJ, Harrison EJ, Jacobs R, Nicopolou-Karayianni K, et al. Detecting reduced bone mineral density from dental radiographs using statistical shape models. IEEE Trans Inf Technol Biomed 2007;11:601-10.

27. Devlin H, Horner K. A study to assess the relative influence of age and edentulousness upon mandibular bone mineral density in female subjects. Oral Surg Oral Med Oral Pathol Oral Radiol Endod 2007;104:117-21.
28. Abdullah MB. On a robust correlation coefficient. Statistician 1990;39:455-60.
29. Stacey F, Heasman PA, Heasman L, Hepburn S, McCracken GI, Preshaw PM. Smoking cessation as a dental intervention—views of the profession. Br Dent J 2006;201:109-13.
30. Greenberg BL, Glick M, Goodchild J, Duda PW, Conte NR, Conte M. Screening for cardiovascular risk factors in a dental setting. J Am Dent Assoc 2006;138:798-804.
31. Kujan O, Duxbury AJ, Glenny AM, Thakker NS, Sloan P. Opinions and attitudes of the UK's GDPs and specialists in oral surgery, oral medicine and surgical dentistry on oral cancer screening. Oral Dis 2006;12:194-9.

Reprint requests:

Hugh Devlin, PhD, MSc, BSc, BDS.
School of Dentistry, University Dental Hospital of Manchester
Higher Cambridge Street
Manchester, M15 6FH, United Kingdom
Hugh.Devlin@manchester.ac.uk

41. **Improving the detection of osteoporosis from dental radiographs using active appearance models.** M.G. Roberts, J. Graham and H. Devlin. *Proceedings of the IEEE International Symposium on Biomedical Imaging Rotterdam, The Netherlands. April 2010. W. Niessen and E. Meijering, eds. IEEE. pp 440-443.* doi: 10.1109/ISBI.2010.5490314

IMPROVING THE DETECTION OF OSTEOPOROSIS FROM DENTAL RADIOGRAPHS USING ACTIVE APPEARANCE MODELS

M. G. Roberts[1], *J. Graham*[1], *H. Devlin*[2].

[1] Imaging Science and Biomedical Engineering, School of Cancer and Enabling Sciences, University of Manchester, UK.
[2] School of Dentistry, University of Manchester, UK.

ABSTRACT

We describe improvements to a method of detecting patients at risk of osteoporosis from automatic measurement of the inferior mandibular cortex on panoramic dental tomograms. Previous work had used an Active Shape Model (ASM) to locate the mandibular edges. However, the edge-based ASM has little lateral positioning information and in osteoporotic cases the superior border is often poorly defined. These problems can degrade the accuracy of the measurements of cortical width. We have obtained superior accuracy using an Active Appearance Model incorporating a complex texture model derived from an osteoporotic-enriched training set. This leads to improvements in diagnostic accuracy, when applied to a large dataset of 663 subjects with known Bone Mineral Density.

Index Terms— Osteoporosis, Dental Radiographs, Active Appearance Model, Active Shape Model, Image Segmentation.

1. INTRODUCTION

Osteoporosis is a progressive skeletal disease characterized by low bone mass and structural deterioration of bone tissue, leading to an increased susceptibility to fragility fracture. Osteoporosis is associated with increased morbidity and mortality - 27% of women who sustain a hip fracture die within 1 year. Early detection of osteoporosis can allow therapeutic intervention, but the condition is often undiagnosed. There has been recent interest among dental researchers in identifying those at risk of reduced Bone Mineral Densitiy (BMD) from dental radiographs [1]. Figure 1 shows part of a Panoramic Dental Tomogram (PDT), on which the inferior mandibular cortex is visible. It was reported in [2] that measuring the thickness of the cortical bone using Active Shape Model (ASM) search [3] provides a good diagnostic of low BMD at other skeletal sites. Further technical details on the ASM search are given in [4].

In [4] two ASM procedures were presented: the shape could be manually initialised by an expert practitioner clicking on 4 points (see Figure 1) along the inferior mandible; or a fully automatic search starting from the mean shape could be performed. These were referred to as 4PFit and UFit respectively. As not all dental practitioners are expert in aligning the shape with the 4 (indistinct) landmark points, a fully automatic system is desirable. This may also be useful in large epidemiological studies. However both point-to-line accuracy and ROC curve area were poorer in the UFit case. Closer inspection of UFit solutions revealed some ASM search failures, or gross lateral misalignment, due to the fact that the edge-based ASM has little evidence to use for positioning laterally along the mandible. This may lead to bias in the thickness measurement. Furthermore in osteoporotic cases the superior border is often poorly defined, which can lead to poor positioning of the superior border by the ASM. We have attempted to address these issues. We use more search phases than [4], starting with a more global ASM search with a laterally extended shape model, and concluding with an Active Appearance Model (AAM) [5] using a complex texture model. Because the AAM models the correlation between shape and texture, it may be better suited than an ASM to fit to thin osteoporotic mandibles with poorly defined superior borders. It is inevitable that there will be occasional search failures, and so we have enhanced the method by providing criteria for identifying search failure.

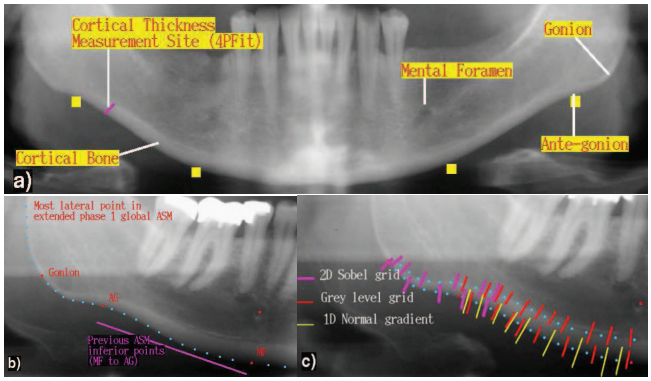


Fig. 1. a) Typical dental panoramic tomogram, with labelled anatomical points. The rectangles indicate the lateral position of the 4 point initialisation. b) the phase 1 ASM shape extent (patient right side); c) the regions of the AAM texture model sample types are indicated.

2. MATERIALS AND METHODS

2.1. Data

We used the training set data already reported in [4], comprising PDTs and BMD measurements from 132 female patients aged 45-55 who attended for routine dental treatment. The

This work was supported by a research grant from the Dunhill Trust and used data previously acquired by a study funded by a technological development project grant from the European Commission (QLK6-2002-02243).

independent test data had been previously collected during the OSTEODENT study [1, 2] and consisted of 663 ambulant female patients of which 140 were osteoporotic. Patients were diagnosed osteoporotic according to the World Health Organization criteria, i.e. those with a bone mineral density T-score value below -2.5, evaluated at 3 skeletal sites (femoral neck, total hip, and lumbar spine). Further details are given in [1, 2]. As the original training set contained very few osteoporotic cases, the training set was enhanced by adding a further 50 osteoporotic or osteopoenic cases taken from the OSTEODENT set (BMD T-score < -2). The modelling was extended by annotating more lateral points past the Gonion (Figure 1). Further intermediate points for texture sampling are interpolated as in [4].

2.2. Segmentation Method

In [4] a two-phase ASM was used with separate models for the left and right halves, built from expertly annotated points lying between the antegonion (AG) to the sub-mental foramen (MF) (Figure 1a,b). Firstly ASMs trained on just the inferior border (Figure 1b) from Ante-Gonion (AG) to sub-Mental Foramen (MF) points are run; followed by ASMs using both borders. We have extended this to use three phases. Firstly a global ASM describing both halves of the cortex is run to locate the inferior border. The model and search are laterally extended beyond the Gonion (GO) (Figure 1b), to seek more lateral positioning information. Starting from the point locations found at the end of the first phase, we run an ASM search for the lower border only, as in phase 1 of [4], but with a laterally extended model defined between MF and GO (rather than AG). Finally we locate the superior border, and refine lateral positioning, by running Active Appearance Model (AAM) search for the two halves. A merged texture model was used for the AAM by concatenating several sub-sample types, some of which are sampled only in specific sub-regions (Figure 1c). From AG to MF grey levels are sampled in rectangles normal to the shape, with a width of three sampling steps (1 step=2pixels). The sampled grey level texture is renormalised onto (-1,1) using the (signed) Geman-McClure robust function [6], so the median is zero, with a robust scale estimate based on the Median Absolute Deviation. In the region from GO to 15% of AG-MF distance past AG, the 2D Sobel gradients in a sampling rectangle of width 5 steps are computed, with the x-axis aligned to the local tangent. Changes in gradient vector components across the rectangle may offer lateral positioning clues where the curvature changes. The gradient vector components are renormalised onto (-1,1) across the whole sample using an orientation-preserving sigmoidal function as in [7]. Finally smoothed 1D normal gradient components are computed along the inferior border from MF to AG with sigmoidal renormalisation.

To evaluate whether the UFit search was successful we calculated two measures. Firstly we calculated the AAM texture model residual sum of squares, with some empirical rescaling factors (derived by bootstrap resampling) to bring to better alignment with a χ^2 distribution as in [8]. We then transform this to a log-probability of being so far into the upper tail of an approximated Gaussian cumulative distribution. The texture measure is the lower of that for the left and right AAM submodels. Image noise can make it difficult to distinguish a failure from a successful fit with high superimposed clutter, so we also use a measure based on shape symmetry. The mandible is reasonably left/right symmetric, but search failures often have one half successfully fitted, whilst the other fits to either shadow artefacts below the mandible, or other edges up near the teeth. Both of these tend to destroy left/right symmetry. We measure shape symmetry by computing the angle of the local shape gradient θ_i at point i, reflecting it in a notional symmetry axis to obtain $\hat{\theta}_i$, and the actual angle at the corresponding point on the other side θ'_i. The notional symmetry axis is derived by fitting both halves to a global shape model, and taking the rotation of the global pose angle from the y-axis. The symmetry measure M_S is a similar log-probability measure as for the texture residuals, given by:

$$M_S = log(1 - F(\frac{z}{\sqrt{2}})) \qquad z = \frac{\sum_{i \in I} \|\hat{\theta}_i - \theta'_i\| - \mu_S}{\sigma_S} \quad (1)$$

where $F(x) = 0.5(1 + erf(x))$, and I contains the indices of a subset of 20 (equispaced) points between MF and AG. The mean and standard deviation μ_S, σ_S are derived from the training set. Cut off thesholds are then derived from the 4PFit distributions of the two measures. These are set to $\mu + 5\sigma$ of either measure, or $\mu + 3\sigma$ in both, where μ is the median, and σ is the robust S_n estimate [9] of standard deviation suited to asymmetric distributions.

2.3. Experimental Procedure

To evaluate segmentation accuracy using the ground-truth in the training set, leave-8-out cross-validation was used with randomised ordering. Leave-8-out was used rather than leave-1-out due to the long training time required for the feature AAMs, and should not lead to significant deterioration given a training set size of 182. To evaluate the diagnosis of osteoporosis, the search algorithms were run on the remaining independent OSTEODENT test data set using ASM and AAM models built with the full training set. The segmentations (via leave-8-out) of the 50 OSTEODENT images used for training were added to the diagnostic test set. The algorithm was run in 4PFit and UFit modes. ASM and AAM typically use a multi-resolution coarse-to-fine search on a Gaussian image pyramid. We used two levels of pyramid for all phases in the 4PFit searches, increasing for UFits to four levels in phase 1, and three levels thereafter. For the 4PFit searches, 10 replications were performed to evaluate precision, with Gaussian random displacements added to the annotated points, based on manual precision figures in [4]. The precision error is calculated as the mean displacement from the mean solution over the ten replications (with 9 degrees of freedom). For comparision we also completed the segmentation of both borders using a final edge-seeking ASM phase as in [4].

To calculate the mandible thickness we fitted Bezier splines to both borders from MF to AG and placed 100 equi-distant points on the inferior spline. The distance from each inferior point to the nearest point on the superior spline was computed, and laterally smoothed in a moving window of semi-width 0.1L, using a Gaussian kernel of standard deviation 0.05L, where L is the total spline distance from MF to AG. Similarly to [2] we optimised the measurement site, as correlation between IMC thickness and BMD varies along the mandible [2]. We selected the inferior point giving maximal ROC curve area in the training set. We independently

optimised the 4PFit and UFit sites, because lateral error in the UFit case may mean the measurements become more unreliable close to the AG, where the cortex rapidly thins. The optimal sites were found to be 0.79L and 0.66L from MF for 4PFit and UFit respectively, though the smoothing window allows for some latitude. We evaluated ROC curves against osteoporosis diagnosed at any of the 3 skeletal sites; and against osteoporosis diagnosed only using Femoral Neck BMD. The latter is likely to be more closely correlated with mandibular BMD, as both sites are predominatly comprised of cortical (not trabecular) bone. Also hip fracture is the most serious consequence of osteoporosis.

3. RESULTS

The point-to-curve error statistics are presented in table 1. For comparison, the mean point-to-curve error in the ASM of [4] was 0.31mm for 4PFit, rising to 0.49mm for UFit. Note that the quoted mean comparison figure from [4] represents a more idealised accuracy, as there were fewer osteoporotics, and no random precision error was added to the initialisation. On the same dataset as [4] the AAM hybrid achieves respective accuracies of 0.24mm (4PFit) and 0.31mm (UFit). Thus the UFit point-to-line accuracy on largely normal cases has improved to be comparable to the previous 4PFit. A more significant improvement in accuracy is in fitting the superior borders of osteoporotic cases, which are the most difficult cases. The ASM accuracy degrades for superior osteoporotic borders in both 4PFit and UFit cases, whereas the AAM accuracy is maintained at similar levels to normal cases for the 4PFit; whilst in the UFit case the upper tail of the error distribution is reduced, and the 75^{th} percentile is halved compared to ASM. But a small number of UFit partial search failures (or poor lateral alignment) mean that the UFit mean error degrades to 0.64mm, though that is still significantly better than for the ASM. For the superior osteoporotic borders, the 98% confidence intervals for the mean error difference (ASM-AAM), derived by factored (e.g. by randomised initialisation) bootstrap resampling, are [1.0,3.1] pixels (4PFit) and [1.2,2.90] pixels (UFit).

The mean precision for the 4PFit was 0.09mm (point-to-line) and 0.76mm (point-to-point), which compare favourably to the respective point initialisation precisions (0.31mm, 2.45mm). The point-to-point error in the UFit case was 4.58mm, an improvement on the 5.73mm in [4], but still significantly larger than the expert manual point-to-point precision of 2.45mm reported in [4]. Figure 2 shows the AAM UFit solution (subset of points illustrated) for an osteoporotic case; note the thinning of the cortex.

The failure criteria identified 58 failure cases out of 663, a total of 9%. Upon visual inspection 8 of these were found to be false positives. We visually examined the further 100 worst ranked image fits, and identified a further 15 failures that were missed with the set thresholds. In clinical application it may therefore be desirable to slightly reduce the thresholds. The current settings give 99% specificity but only 77% sensitivity to failure. In only 5 cases was there failure of both left and right sides.

Figure 3 shows the ROC curves obtained using the mandibular cortical thickness at the optimum position for both 4PFit and UFit estimates. Table 2 compares the ROC areas for the new AAM method with previously published

Table 1. Model Fit Point Errors (mm)

Fit	point-to-curve error (mm)			
	Mean	median	75%-ile	90%-ile
	4PFit			
ASM All Data	0.29	0.16	0.35	0.87
AAM All Data	0.25	0.18	0.34	0.64
ASM Ost Sup	0.53	0.30	0.92	1.67
AAM Ost Sup	0.32	0.24	0.47	0.80
	UFit			
ASM All Data	0.4	0.17	0.37	1.18
AAM All Data	0.36	0.19	0.37	0.78
ASM Ost Sup	0.85	0.34	1.08	1.7
AAM Ost Sup	0.64	0.26	0.55	1.48

Table 2. ROC Curve Areas

	4PFit		UFit	
BMD Site	ASM [2]	AAM	ASM [2]	AAM
Any Site	0.816	0.823	0.759	0.799
Femoral Neck	0.835	0.862	0.805	0.851

Table 3. False Positive Rates for UFit

	FPR at 70% sensitivity		FPR at 80% sensitivity	
BMD Site	ASM [2]	AAM	ASM [2]	AAM
Any Site	32.4%	23.2%	46.6%	37.9%
Femoral Neck	24.2%	13.8%	36.1%	22.5%

Table 4. McNemar test statistics (dN) for False Positive Rate comparisons between ASM and AAM extracted thickness at various sensitivity points on ROC curves.

	dN for sensitivities		
BMD Site	70%	75%	80%
Any Site	10.53	6.26	6.43
Femoral Neck	33.21	25.53	18.72

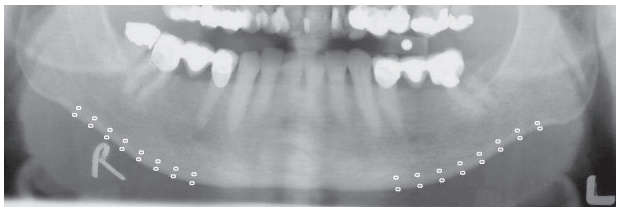


Fig. 2. AAM UFit for osteoporotic patient. Rectangles are drawn at a subset of points from MF to AG. Note the thinning of the cortex.

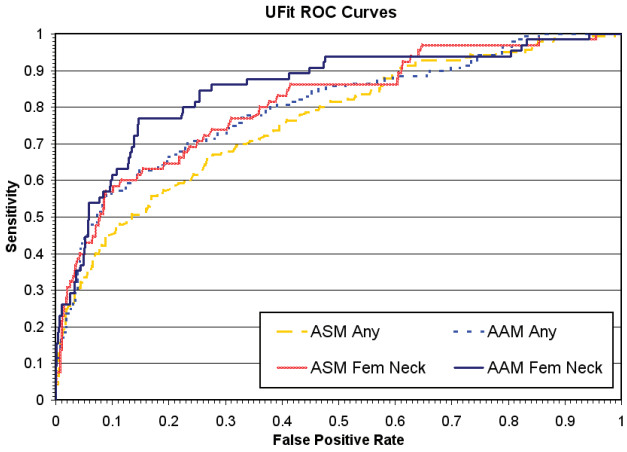


Fig. 3. ROC Curves for UFit comparing AAM segmention with previous ASM. Curves are shown for osteoporosis at any skeletal site (Any in Legend), and at the femoral neck (Fem Neck) in Legend)

ASM results from [2], whilst Table 3 compares false positive rates (FPR) at 70% and 80% sensitivities. The specificities at these (and 75%) sensitivities were compared using the McNemar test [10] (Table 4) All test statistics were far in excess of the 99% point of the χ^2_1 distribution (6.63) for femoral neck diagnoses, and when comparing osteoporosis at any site the lowest McNemar test statistic was 6.26 (for 75% sensitivity), which is still significant for $p = 0.025$. This confirms that there is a real reduction in false positive rate at likely operating points of equivalent sensitivity.

4. DISCUSSION

The hybrid ASM/AAM results in an improvement in fitting accuracy measured by point-to-line distance, particularly for the superior edges in osteoporotic cases. However despite including a larger span of the overall mandible, and attempting to implicitly incorporate curvature features in the AAM, lateral positioning errors remain quite large for UFits. This may be because the AAM update is too local, the AG inflection is sometimes very subtle, and also the areas of significant curvature around the gonion are often corrupted by shadow artefacts. We are currently investigating hybrid schemes including a more global search using connected graphs of feature patches.

Because of limited correlation between mandible thickness and BMD at other skeletal sites, improvements in diagnostic accuracy for 4PFit are slight. The UFit accuracy improvement results in a more substantial gain in diagnostic accuracy, which brings the fully automatic method close to the previous manually initialised ASM, if low BMD at any skeletal site is the required condition. The AAM Ufit reduces FPR at 70% sensitivity from 32% to 23% (similar to 4PFit ASM). An even more substantial reduction in FPR occurs if we take Femoral Neck BMD as the gold standard, for which at 80% sensitivity the FPR reduces from 36% to 22.5%, better than the 4PFit ASM FPR of 27%. This larger improvement against osteoporosis at the femoral neck may indicate that there is a fundamentally better correlation between the two bone sites, which is physically plausible as both are predominantly cortical (rather than trabecular) bone.

In summary we have improved the fully automatic point-to-line fitting accuracy by use of a hybrid ASM/AAM search. In fully automatic UFit mode the method diagnoses osteoporosis at the femoral neck with a sensitivity of 80% and specificity 77.5%. As most dental practitioners are not expert in this method of assessment, the more automatic the method, the more likely it is to be adopted. Nevertheless any automatic method is prone to occasional failure, and so we have supplemented the fit with failure conditions, in the event of which the dentist could revert to a 4-point manual initialisation of the shape.

Mandibular thickness is not the only clue present in the image. Osteoporotic mandibles also tend to exhibit "holes" - small dark regions where cortical bone is largely absent; conversely there may be additional superior edges ("residues") where there is still some partial remnants of where the endosteal border used to be. Future work will therefore also use morphological image analysis to improve diagnostic accuracy, but this will rely upon an accurate initial segmentation.

5. REFERENCES

[1] K. Karayianni et al., "Accuracy in osteoporosis diagnosis of a combination of mandibular cortical width measurement on dental panoramic radiographs and a clinical risk index (OSIRIS): The OSTEODENT project," *Bone*, vol. 40, no. 1, pp. 223–229, 2007.

[2] H. Devlin, P.D. Allen, J. Graham, and et al., "Automated osteoporosis risk assessment by dentists: a new pathway to diagnosis," *Bone*, vol. 40, pp. 835–842, 2007.

[3] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active Shape Models - Their Training and Application," *Computer Vision and Image Understanding*, vol. 61, no. 1, 1995.

[4] P. D. Allen et al., "Detecting Reduced Bone Mineral Density From Dental Radiographs Using Statistical Shape Models," *IEEE Transactions on Information Technology in Biomedicine*, vol. 6, no. 11, pp. 601–610, 2007.

[5] T F Cootes, G J Edwards, and C J Taylor, "Active appearance models," in *5th European Conference on Computer Vision*. 1998, pp. 484–498, Springer.

[6] S Geman and D McClure, "Statistical methods for tomographic image reconstruction," *Bulletin of the International Statistical Institute*, vol. LII, pp. 4–5, 1997.

[7] I M Scott et al., "Improving active appearance model matching using local image structure," in *18th IPMI Conference*, 2003, pp. 258–269.

[8] M G Roberts et al., "Vertebral shape: Automatic measurement with dynamically sequenced active appearance models," in *8th MICCAI Conference*, 2005, pp. 733–740.

[9] P J Rousseeuw and C Croux, "Alternatives to the median absolute deviation," *J Amer Stat Assn,*, vol. 88, pp. 1273–1283, 1993.

[10] Q McNemar, "Note on the sampling error of the difference between correlated proportions or percentages.," *Psychometrika*, vol. 12, pp. 153–157, 1947.

42. **Changes in mandibular cortical width measurements with age in men and women.** M. Roberts, J. Yuan, J. Graham, R. Jacobs and H. Devlin, *Osteoporosis International, 22: 1915-1925 (2011).* doi 10.1007/s001 98-010-1401-3.

Osteoporos Int (2011) 22:1915–1925
DOI 10.1007/s00198-010-1410-3

ORIGINAL ARTICLE

Changes in mandibular cortical width measurements with age in men and women

M. Roberts · J. Yuan · J. Graham · R. Jacobs · H. Devlin

Received: 8 June 2010 / Accepted: 18 August 2010 / Published online: 1 October 2010
© International Osteoporosis Foundation and National Osteoporosis Foundation 2010

Abstract
Summary Automated software was used to measure the mandibular cortical width in a large sample of dental radiographs. We determined that cortical thinning normally starts in women at age 42.5 years and accelerates thereafter. We can estimate population referral rates and thus enable cost benefit analyses for osteoporosis detection by dentists.
Introduction Previous studies have shown that the mandibular cortical width is significantly correlated with the bone mineral density at sites which may undergo osteoporotic fracture, e.g. hip. Mandibular cortical width can be determined automatically from dental panoramic radiographs that dentists frequently request, using appropriate software. We study the distribution of cortical width given age to predict those patients requiring further investigation for osteoporosis.
Methods The mandibular cortical width was measured in 4,949 dental panoramic tomograms, in patients aged 15–94 years. The inferior and superior cortical edges were detected automatically using a global active shape model image search, followed by an active appearance model search. Nonparametric statistical analysis and nonlinear piecewise linear/quadratic regression were used to analyse the data.
Results For females, the mean cortical width had a linear increase before the age of 17 years, a period of no change (estimate=3.25 mm, se=0.01) until the age of 42.5 years, followed by a quadratic decrease with age. For males, it had a linear increase before the age of 19 years, a constant value (estimate=0.37 mm, se=0.01) until the age of 36 years and then a slow linear decrease. The rate of decrease in mean cortical width goes from 0.049 to 0.105 standard deviations per year in the 60–80-year-old female age group, in line with published bone mineral density T-score reductions.
Conclusions The pattern of decrease in mandibular cortical width with age was similar to the known pattern of bone loss from the hip, accelerating in women after the age of 42.5 years.

Keywords Active appearance model · Cortical bone · Mandible

M. Roberts
Imaging Science and Biomedical Engineering, School of Cancer and Enabling Sciences, University of Manchester, Manchester, UK

J. Yuan
School of Mathematics, University of Manchester, Manchester, UK

J. Graham
Imaging Science and Biomedical Engineering, School of Cancer and Enabling Sciences, University of Manchester, Manchester, UK

R. Jacobs
Oral Imaging Center, Katholieke Universiteit, Leuven, Belgium

H. Devlin (✉)
School of Dentistry, University of Manchester, Manchester, UK
e-mail: Hugh.Devlin@Manchester.ac.uk

Introduction

Thin mandibular cortical width measurements in dental panoramic radiographs have been used as a predictive measure of systemic osteoporosis [1]. These radiographs are frequently used by dentists and may provide a useful opportunity to contribute to osteoporosis diagnosis, especially if the radiograph is taken for other reasons. In that

situation, there is no additional radiation dose due to the osteoporosis investigation. Early detection by dentists would allow the necessary preventive treatment to be instituted; however, what constitutes a normal value of mean cortical width or the expected degree of variation about this measurement in the population is unknown.

Ledgerton et al. [2] showed in a sample of British women that mandibular cortical width generally decreased from the age of 25 years, but this trend accelerated after the age of 60 years. The study was undertaken only in women, and the sample size of 500 subjects (with ten 5-year age groups) was limited by having to perform the measurement task manually. Manual measurement of the mandibular cortical width is usually undertaken in the region directly below the mental foramen, perpendicular to the lower border of the mandible. Detection of the mental foramen may be unreliable, as multiple foramina are sometimes present [3]. Allen et al. [4] showed that the cortical width could be measured automatically using computer image analysis without the requirement for identification of the mental foramen.

In cross-sectional studies, we have observed an association between mandibular cortical width and the bone mineral density (BMD) at important fracture prone sites such as the spine and hip [5].

Our aim in this study was to measure the mandibular cortical width in dental panoramic tomograms from a large sample of men and women and to determine how it relates to the patients' age and gender.

Materials and methods

All dental panoramic tomogram images were collected anonymously from one hospital database (Leuven, Belgium) over a 2-year period, 2006–2008. Radiographs from the same radiography machine at the same hospital site had contributed to a previous study [6], in which the absolute value of scaling between patient and image was calculated using a ball bearing of known size in the image. No calibration object was present in the radiographs for this study, but the same scaling factor has been assumed.

Data

A total of 6,096 Digital Imaging and Communications in Medicine (DICOM) images were received from the Oral Imaging Centre of the Department of Dentistry, Oral Pathology and Maxillofacial Surgery of the University Hospitals, Catholic University of Leuven. All images had been anonymized, and patients had given general informed consent to the use of their radiographic images for research. All images were inspected visually to exclude those that were unusable due to unacceptable quality or evidence of pathology, such as previous mandibular fractures, surgical procedures, cysts or tumours that could have influenced the mandibular cortex. The remaining set to be analysed contained 4,949 tomograms. Patient age and gender were extracted from the DICOM header files: age ranged from 15 to 94 years, and there were 2,386 males and 2,563 females.

Locating the edges of the mandibular cortex

The method of Allen et al. [4] for locating the inferior and superior edges of the mandibular cortex on dental panoramic radiographs used active shape models (ASMs) [7]. In this approach, the edges are found by optimisation—image search for the best fit to a model which had been built by expert annotation of a training set of images. Separate ASMs were used for the left and right halves of the mandible, modelling the region between the gonion and a point below the mental foramen. The ASM could be run either in a fully automatic mode, search being initiated from the mean shape derived from the training set, or by using a manual initialisation defined using four user-specified points (Fig. 1). However the former gave poorer diagnostic ability than the latter due to some search failures and some cases of lateral misalignment, as the edge-search used by the ASM has little information with which to position the shape laterally along the mandible.

Because of the large size of the dataset, we have developed the modelling of the mandible beyond that of Allen et al. [4] to increase the reliability of fully automatic search [8]. Briefly, the shape model was laterally extended by annotating the upper and lower edges of the mandible beyond the gonion to utilise more clues from the curvature of the mandible and therefore improve the lateral positioning. The search is conducted in two phases. In phase 1, a global ASM search is conducted for the combined (left and right) inferior mandible edge (Fig 1). This edge usually displays high contrast and can be located reliably to provide an initial configuration for phase 2: an active appearance model (AAM, [9]) search on the separate left and right halves of the mandible to further refine lateral position and locate the superior border. AAMs allow the search to make use of an image texture model, providing more information than is available to the ASMs. Technical details are given in Roberts et al. [8], in which it is demonstrated that this ASM/AAM hybrid method is more accurate that the ASM method previously described by Allen et al. [4] and gives superior diagnostic performance on the dataset of Devlin et al. [6].

Detection of search failures

This ASM/AAM hybrid method in fully automatic mode gave similar accuracy to the manually initialised ASM on

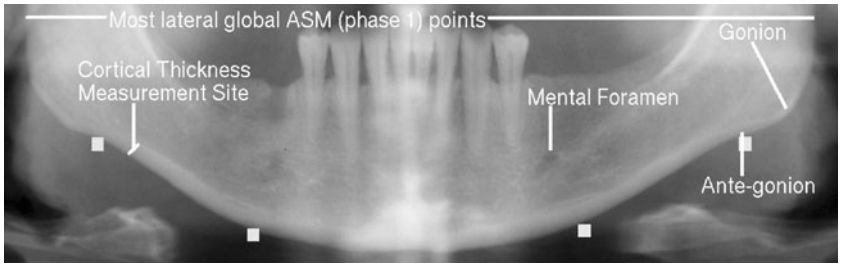


Fig. 1 Typical dental panoramic tomogram, with labelled anatomical points. The rectangles show the lateral positions of the 4 points used if employing a manual initialisation of the Active Shape Model (ASM). The full extent of the phase 1 global ASM beyond the Gonion is also indicated

the training set of Allen et al. [4]. However, when run on the more challenging dataset of Devlin et al. [6], there were partial search failures (convergence to incorrect locations) in about 10% of cases. Therefore, we devised an automatic diagnostic for incorrect convergence [8] based on:

1. The residual sum of squares (RSS) of the fit of the AAM texture model to the image (fits with low residuals are more reliable) and
2. A symmetry measure between the left and right halves (reliable fits should be similar on both halves of the mandible).

Thresholds on both measures were derived from calculating the median μ and a robust estimate of standard deviation σ from an independent dataset [6], but using solutions derived from the four-point manual initialisation. We examined images where either failure criterion exceeded $\mu+\alpha\sigma$. With α=5, the vast majority of searches classified as failures were indeed failures, whereas only 5% of detected failures between α=3 and α=4 were genuine. We therefore set a threshold at α=3, above which the search result was checked visually and if necessary subjected to a four-point manual initialisation to reach a reliable fit. This was applied to 1,273 images; in the remaining 3,676 cases, the superior and inferior edges were located using the fully automated ASM/AAM hybrid search.

Extraction of cortical width measurement

It was previously found that the correlation between mandibular cortical width and BMD at other skeletal sites varies depending on the position along the mandible at which the measurement is made, with a peak occurring at roughly three quarters of the distance from the sub-mental foramen point to the antegonion [4]. We found similarly that a position sited 67% of the distance from the sub-mental foramen point to the antegonion (Fig. 1) gave the best area under a receiver operating characteristic curve, for diagnosing osteoporosis from extracted width of the cortex. Defining this distance as L, the extracted width was smoothed over $\pm 0.1L$ from this point using a Gaussian kernel. Thus, the inevitable imprecision in the automatic lateral positioning of the extraction point is mitigated by the smoothing window. The final measurement is the mean of the width measured at this position in the left and right halves.

Statistical analysis

This is a cross-sectional observational study. The response variable is mandibular cortical width, and the predictor is the patients' age. We used nonparametric regression and parametric linear and nonlinear regression to estimate the mean cortical width as a function of age for males and females.

The basic model is $y=f(x)+\varepsilon$, where y is cortical width, x is age, and ε is an error term that follows a normal distribution with mean 0 and some unknown variance σ^2. Independence between observations is assumed, and a different mean cortical width function $f(x)$ is allowed for each gender group. This model is for nonparametric regression where no specific form of $f(x)$ is given a priori, it is only assumed to be a smooth function. Nonparametric regression can be used on its own to estimate $f(x)$ and can also help to derive a parametric form for $f(x)$.

We consider the age variable x as continuous even though the recorded age was integer valued, for example, someone who had had a 50th birthday, but not the 51st, was recorded as aged 50, but the exact age can be anywhere between 50 and 51. For this reason, we added 0.5 to the recorded age to give the centre of the interval. The large sample size (n=4,949) makes it less of an issue not to have exact age in the data. The sample mean cortical width for each age x available contains all the information about $f(x)$, and a confidence interval can be obtained. Plotting them together for all x provides a clearer indication to the shape of the function $f(x)$ than the original data.

Nonparametric analysis

The methodology of nonparametric estimation and hypothesis testing was entirely from Bowman and Azzalini [10], and their R package 'sm' was used. The mean cortical width function $f(x)$ was estimated for males and females using nonparametric regression by locally fitting linear functions of age to the data. There were as many linear regressions as values of x in the calculations. The data points(x_i,y_i) were

weighted using a Gaussian kernel according to how far x_i was from x relative to a smoothing parameter h—a larger (smaller) value of h means more (less) smoothing. There are various methods of selecting h, including cross validation which minimises the 'leave-one-out' mean square error and the default 'df' method in sm which tries to achieve a certain degree of freedom (default=6). After fitting a linear regression line like this to the data near x, it is used at x to estimate $f(x)$. This approach to the regression problem is nonparametric because a formula for $f(x)$ is not required.

The hypotheses of linearity and constancy were tested separately for males and females over various intervals of age, which were chosen based on inspections of the nonparametric regression plots. Here, linearity means the function is linear over a specific age interval and constancy means it does not change with age over this range.

Parametric analysis

We used piecewise linear regression which is like straight line linear regression, except that the line is allowed to bend at one or more places, resulting in a different equation for each section of the data. It can be fitted as a linear model with suitably constructed predictors. By estimating the equations simultaneously, the line segments will join perfectly. The bends can also be estimated using nonlinear regression together with the model coefficients, and we can allow a smooth quadratic transition at each bend. We did all these for the male and female data and were able to simplify the female model using a quadratic piece.

For parametric linear and nonlinear regression using the R functions 'lm' and 'nls', we refer to Venables and Ripley [11]. The programming was done in R, and we found the code provided by Lindstrom [12] useful when extending the hockey model [13].

Results

Of the 4,949 dental panoramic images, 2,563 were of women (51.8%). The mean age of the females was 44.25 years (sd=17.50) and of males was 43.09 years (sd=17.83). The mean mandibular cortical width was 3.21 mm (sd=0.46) overall, 3.14 mm (sd=0.46) for the females and 3.29 mm (sd=0.45) for the males.

Considering only those subjects below the age of 50 years, there was a statistically significant difference (0.1 mm) in mean mandibular cortical width between the males and females (3.34 mm for males and 3.24 mm for females, t=6.45, P<0.001). For subjects aged 50 or over, the difference of 0.2 mm between males (3.20 mm, sd=0.44) and females (2.98 mm, sd=0.52) was also statistically significant (P<0.001).

We calculated the sample mean cortical width for each age represented in the data separately for males and females. A crude estimate of the mean cortical width function $f(x)$ of age x can be obtained by simply joining the sample means, see the jagged red lines in Fig. 2a, b for females and males, respectively. The vertical lines in blue represent 95% confidence intervals for $f(x)$ for each x individually. There was obviously a need for smoothing to

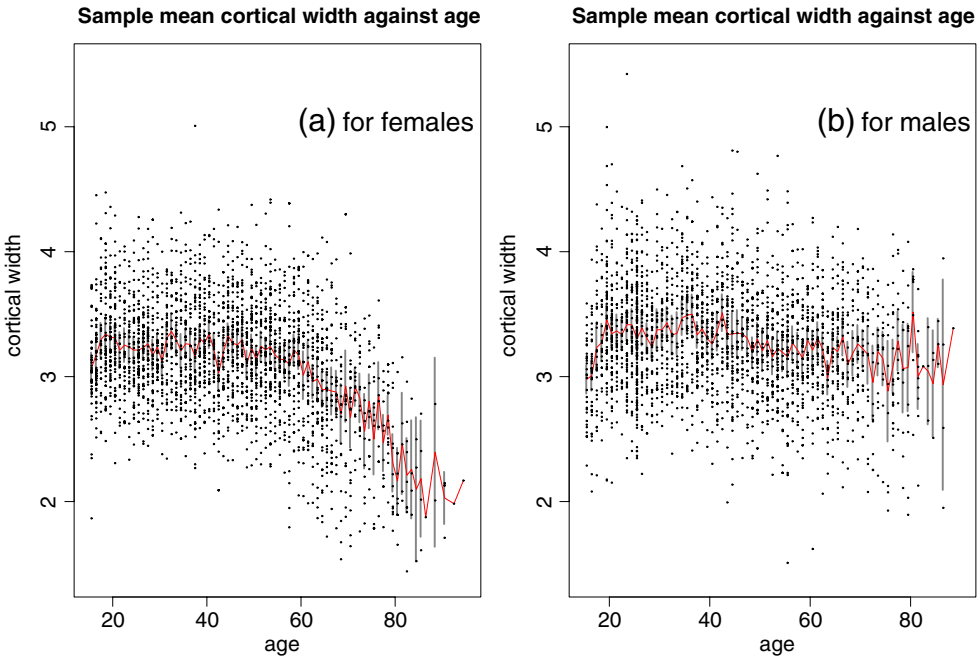


Fig. 2 Sample mean cortical width against age **a** for females and **b** for males

get better estimates of $f(x)$, as we only expect a gradual change from 1 year to another.

Nonparametric analysis

The curves in Fig. 3a, b represent smoothed estimates of $f(x)$ for females and males, respectively, obtained by fitting linear regressions locally as described earlier. The smoothing parameters $h=3.01$ for females and 2.58 for males were chosen by cross validation and used for data with age <25 to bring out more details, after which more smoothing was applied by using the values $h=6.99$ (females) and 6.42 (males), as determined by the degrees of freedom method ('df' in the R package 'sm'). The dotted lines in each graph provide a variation band that allows twice the standard errors of the estimates above and below. The circles represent sample means with varying sizes reflecting how many individuals were included at each age. The mean functions are different for males and females, and a nonparametric test of equality gave highly significant ($P<0.001$) evidence against it.

The sample means and smoothed values are given in Table 1 for females and Table 2 for males. From Fig. 3a and Table 1, the mean cortical width for females seems to increase with age before 19, remains more or less constant till about age 42 and then decreases in a nonlinear fashion afterwards but not by more than 0.1 mm before age 55. There is a slight hint of a bend at about age 75. The curve in Fig. 3b for males suggests a similar but almost linear pattern after age 40 with possible bends at 20, 40, 53 and 66. We conducted nonparametric tests and found no significant evidence ($P>0.10$) against linearity for males or females, over the 5 age intervals delineated by vertical lines in each of the two graphs in Fig. 3. There was significant evidence against linearity for females over 42 ($P<0.01$) but not significant evidence ($P=0.37$) against linearity with respect to $(\text{age}-42)^2$ over this interval. Further tests gave no significant evidence against constancy for females from age 19 to 42 ($P=0.39$) and males from 20 to 40 ($P=0.62$).

Parametric regression

Three parametric models of increasing complexity were used to fit the data.

1. A piecewise linear model with a single breakpoint at age 20 for males and age 55 for females
2. A piecewise linear model with breakpoints at 20 and 40 years for males and 19 and 55 years for females, allowing in each case a constant central segment
3. A combined linear and quadratic model for females, with a breakpoint between linear segments at age 19 and a quadratic segment of the form $a-b(x-42)^2$ starting at age 42

The breakpoints were initially specified by inspection of the nonparametric regression curve, but were later estimated along with other model parameters using nonlinear regression. A smooth quadratic transition of 12 months duration was allowed between adjacent linear sections to maintain a continuous first derivative. The fitted parametric curves for females and males are shown in Figs. 4 and 5, respectively.

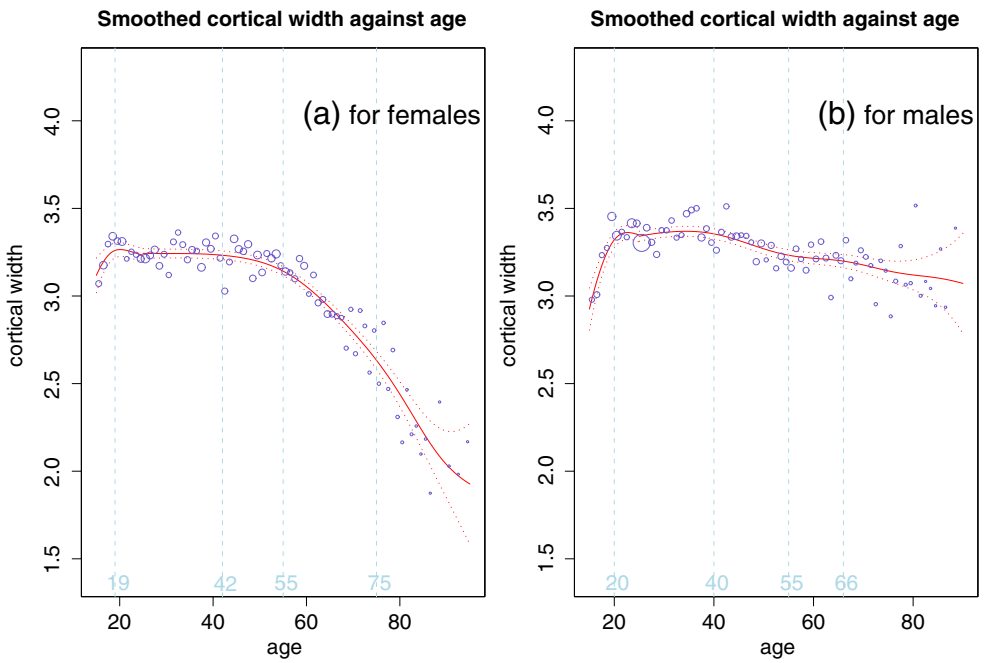


Fig. 3 Smoothed cortical width against age **a** for females and **b** for males

Table 1 Female age, sample means and smoothed cortical widths

Age (years)	15.5	16.5	17.5	18.5	19.5	20.5	21.5	22.5	23.5	24.5	25.5	26.5	27.5	28.5	29.5	30.5	31.5	32.5	33.5
Sample mean (mm)	3.07	3.18	3.30	3.34	3.31	3.31	3.21	3.25	3.24	3.21	3.21	3.23	3.26	3.17	3.24	3.12	3.31	3.36	3.29
Smoothed width (mm)	3.15	3.19	3.23	3.25	3.26	3.26	3.26	3.25	3.24	3.24	3.24	3.24	3.24	3.24	3.24	3.24	3.24	3.24	3.24
Age (years)	34.5	35.5	36.5	37.5	38.5	39.5	40.5	41.5	42.5	43.5	44.5	45.5	46.5	47.5	48.5	49.5	50.5	51.5	52.5
Sample mean (mm)	3.21	3.26	3.26	3.16	3.31	3.27	3.34	3.22	3.03	3.19	3.33	3.27	3.25	3.30	3.10	3.23	3.13	3.24	3.22
Smoothed width (mm)	3.24	3.24	3.24	3.24	3.24	3.24	3.24	3.23	3.23	3.23	3.22	3.22	3.22	3.21	3.21	3.20	3.19	3.18	3.17
Age (years)	53.5	54.5	55.5	56.5	57.5	58.5	59.5	60.5	61.5	62.5	63.5	64.5	65.5	66.5	67.5	68.5	69.5	70.5	71.5
Sample mean (mm)	3.24	3.17	3.14	3.13	3.10	3.21	3.17	3.01	3.12	2.96	2.98	2.90	2.90	2.88	2.88	2.70	2.92	2.67	2.92
Smoothed width (mm)	3.16	3.15	3.13	3.12	3.10	3.08	3.06	3.04	3.02	3.00	2.97	2.95	2.92	2.89	2.86	2.84	2.81	2.78	2.75
Age (years)	72.5	73.5	74.5	75.5	76.5	77.5	78.5	79.5	80.5	81.5	82.5	83.5	84.5	85.5	86.5	88.5	90.5	92.5	94.5
Sample mean (mm)	2.83	2.56	2.80	2.50	2.85	2.47	2.69	2.31	2.16	2.47	2.21	2.26	2.10	2.18	1.87	2.39	2.03	1.98	2.17
Smoothed width (mm)	2.72	2.68	2.65	2.62	2.58	2.54	2.50	2.46	2.42	2.37	2.33	2.29	2.24	2.20	2.16	2.09	2.02	1.97	1.94

Table 2 Male age, sample means and smoothed cortical widths

Age (years)	15.5	16.5	17.5	18.5	19.5	20.5	21.5	22.5	23.5	24.5	25.5	26.5	27.5	28.5	29.5	30.5	31.5	32.5	33.5
Sample mean (mm)	2.98	3.01	3.23	3.27	3.45	3.34	3.37	3.34	3.42	3.41	3.30	3.39	3.31	3.24	3.38	3.37	3.43	3.33	3.35
Smoothed width (mm)	2.98	3.09	3.17	3.25	3.30	3.34	3.36	3.36	3.36	3.35	3.34	3.35	3.35	3.36	3.36	3.36	3.37	3.37	3.37
Age (years)	34.5	35.5	36.5	37.5	38.5	39.5	40.5	41.5	42.5	43.5	44.5	45.5	46.5	47.5	48.5	49.5	50.5	51.5	52.5
Sample mean (mm)	3.47	3.49	3.50	3.33	3.38	3.30	3.26	3.36	3.51	3.34	3.34	3.35	3.34	3.31	3.20	3.30	3.21	3.29	3.16
Smoothed width (mm)	3.37	3.37	3.37	3.37	3.36	3.36	3.35	3.35	3.34	3.33	3.32	3.31	3.30	3.29	3.28	3.27	3.27	3.26	3.25
Age (years)	53.5	54.5	55.5	56.5	57.5	58.5	59.5	60.5	61.5	62.5	63.5	64.5	65.5	66.5	67.5	68.5	69.5	70.5	71.5
Sample mean (mm)	3.23	3.19	3.16	3.27	3.21	3.15	3.29	3.21	3.31	3.22	2.99	3.23	3.20	3.32	3.10	3.19	3.26	3.22	3.17
Smoothed width (mm)	3.24	3.24	3.23	3.23	3.22	3.22	3.22	3.21	3.21	3.21	3.21	3.20	3.20	3.20	3.19	3.19	3.18	3.17	3.17
Age (years)	72.5	73.5	74.5	75.5	76.5	77.5	78.5	79.5	80.5	81.5	82.5	83.5	84.5	85.5	86.5	88.5			
Sample mean (mm)	2.95	3.20	3.14	2.88	3.08	3.28	3.06	3.07	3.52	3.00	3.08	3.04	2.94	3.27	2.94	3.39			
Smoothed width (mm)	3.16	3.15	3.15	3.14	3.13	3.13	3.13	3.12	3.12	3.12	3.11	3.11	3.10	3.10	3.10	3.08			

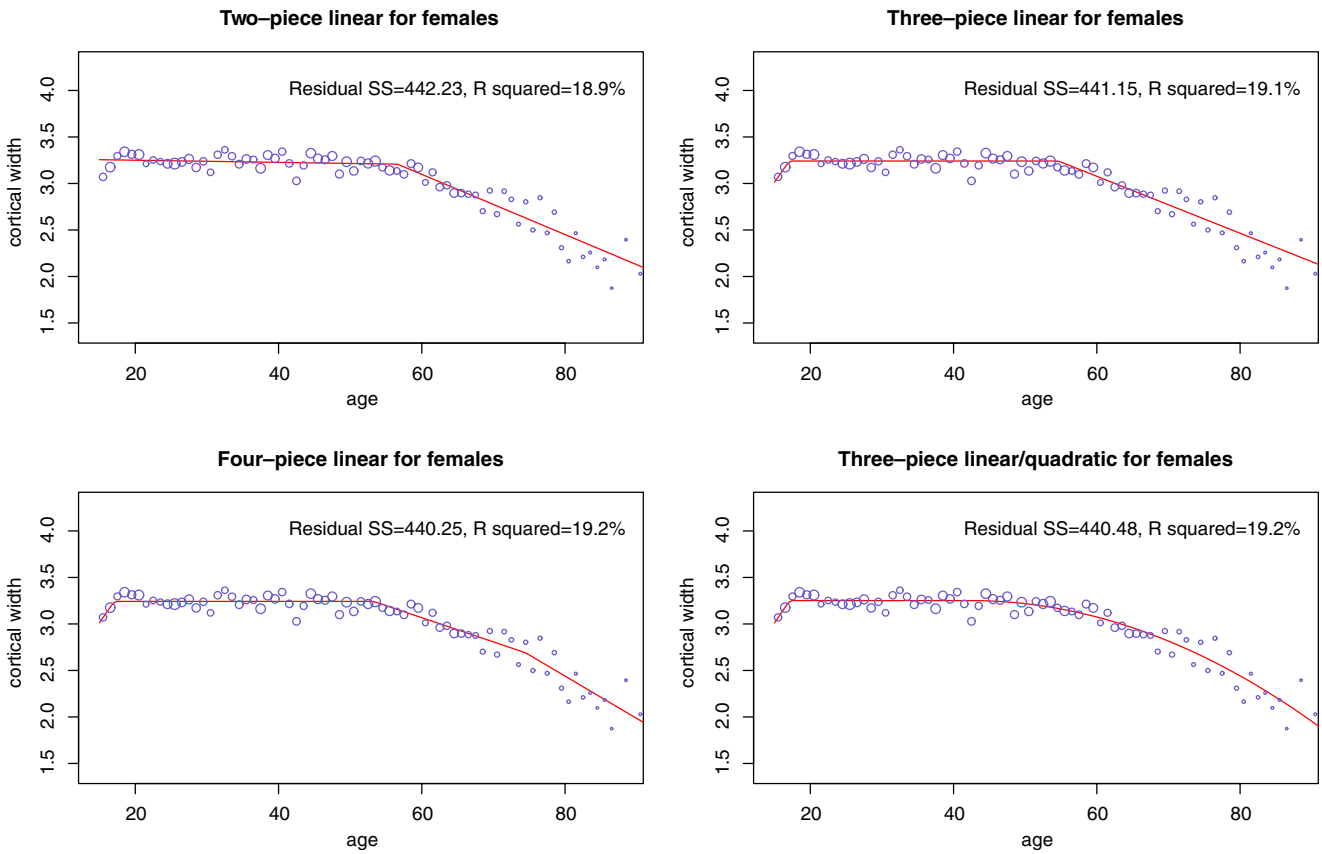


Fig. 4 Two-piece, three-piece and four-piece linear regression models and three-piece linear/quadratic regression model for females

There was a significant difference in RSS between the female two-piece and four-piece models (442.23 vs 440.25, P=0.01), and the improvement from the three-piece to the four-piece model was significant at 10% though not significant at 5% (441.15–440.25, P=0.07) (Fig. 4). The last model with a quadratic piece has a slightly larger RSS than the four-piece. However, it is the preferred model as it has a smaller Akaike's information criterion or AIC value (2,771.9 vs 2,774.5) taking into account its reduced number of parameters. It also looks more natural giving a smooth decay with an increasing rate with age. For the males, the three-piece linear model is preferred, as it provides a significant improvement in RSS from the two-piece (P= 0.01) but adding one or more bends does not do so (P= 0.13, 0.23) (Fig. 5).

The estimated bends in the preferred models are at age 17.12 (se=0.87) and 42.46 (se=1.85) for females and 19.08 (se=0.72) and 36.02 (se=4.40) for males. We set them to 17 and 42.5 for females and 19 and 36 for males and fitted the models for the final time. Details of the final models are given in Table 3 and plotted in Fig. 6 together with nonparametric estimates for comparison. The parametric and nonparametric estimates agree very well.

The model for females has a residual sum of squares (RSS) of 440.49, compared to a total sum of squares about the mean of 545.19. The proportion of the variance explained by the age dependence is measured by the pseudo-R^2 value:

$$1 - 440.49/545.19 = 19.20\%.$$

For males, RSS=456.57, but the total sum of squares of 477.51 was lower than for females, resulting in a reduced pseudo-R^2 value for males of:

$$1 - 456.57/477.51 = 4.38\%.$$

This reflects the fact that that the mean cortical width for males varies less with age than for females and changes little after about age 20.

The estimated mean mandibular cortical width of females decreased considerably with age after about 50 years (Fig. 4). The estimated mean cortical width (Table 4) decreased by 4.5% from 3.22 mm at age 50 to 3.07 mm at age 60 years and by a further 8.4% at age 70–2.82 mm. From age 40 to 70, the decrease was 13.4%. The male cortical width decreased by 1.8% from age 50 to 60 years and a further 1.9% from age 60 to 70 years. The reduction in the male cortical width from 3.35 mm at age 40 years to 3.16 mm at age 70 years was 5.4% (Table 4). Taking into account the effect of age, the estimated standard

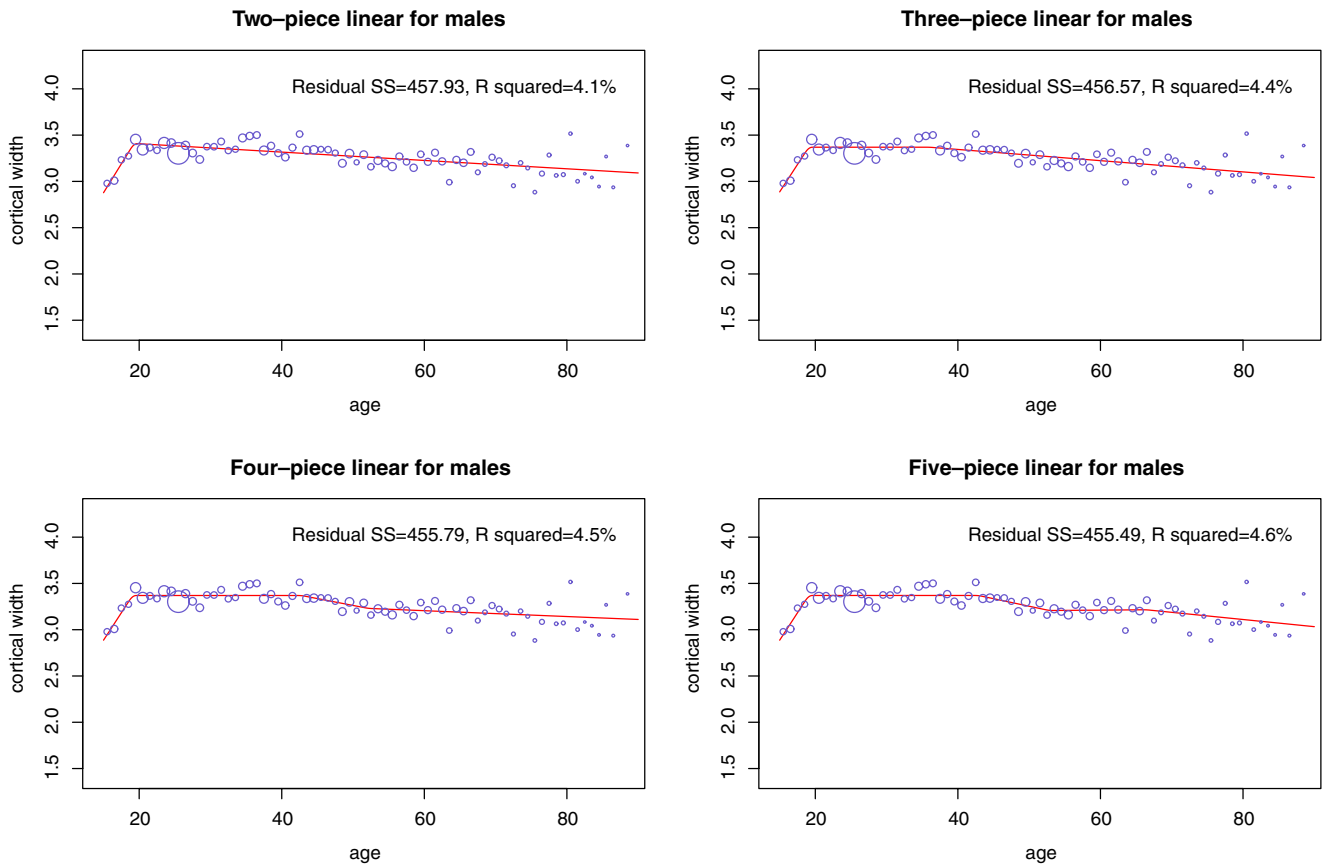


Fig. 5 Two-piece, three-piece, four-piece linear and five-piece linear regression models for males

deviation of cortical width for males was 0.44 and that for females was 0.41 (Table 3).

Discussion

Mandibular cortical width has been proposed as an opportunistic method of detecting osteoporosis in women despite some variation about the sample mean. In our sample, the estimated standard deviation in the cortical widths of females was 0.41 mm, which takes account of age. Potential sources of variability include errors in patient positioning in the radiography machine, inaccuracies in the accurate location of the mandibular measurement site, variable anisotropic scaling at different locations in the dental panoramic tomogram, measurement error in the software detection of the endosteal bone margins and inherent variation in the population. Our estimated standard deviation of 0.41 is less than the value of 1.2 obtained by Ledgerton et al. [2] using manual methods. To summarize, we have found a nonparametric estimate and a precise formula for the expected cortical width at any age after 15 years and a better estimate of the standard deviation. The improved estimate of the age-related variation in cortical

Table 3 Details of final fitted parametric models for females and males

	Females		Males	
$f(x)$	$a+c(x-17)$	$x<16.5$	$a+c(x-19)$	$x<18.5$
	$a-0.35c(x-17.5)^2$	$16.5\leq x7.5$	$a-0.5c(x-19.50)^2$	$18.5\leq x<19.5$
	a	$17.5\leq x<42.5$	a	$19.5\leq x<35.5$
	$a-\mathrm{b}(x-42.5)^2$	$x\geq 42.5$	$a-0.5\mathrm{b}(x-35.5)^2$	$35.5\leq x<36.5$
			$a-b(x-36)$,	$x\geq 36.5$
Estimates	$a=3.25(se=0.01)$	$b=0.000575$ (se=0.000023)	$a=3.37(se=0.01)$	$b=0.0061$ (se=0.0007)
	$c=0.12(se-0.01)$	$\sigma=0.41$	$c=0.12(se=0.02)$	$\sigma=0.44$

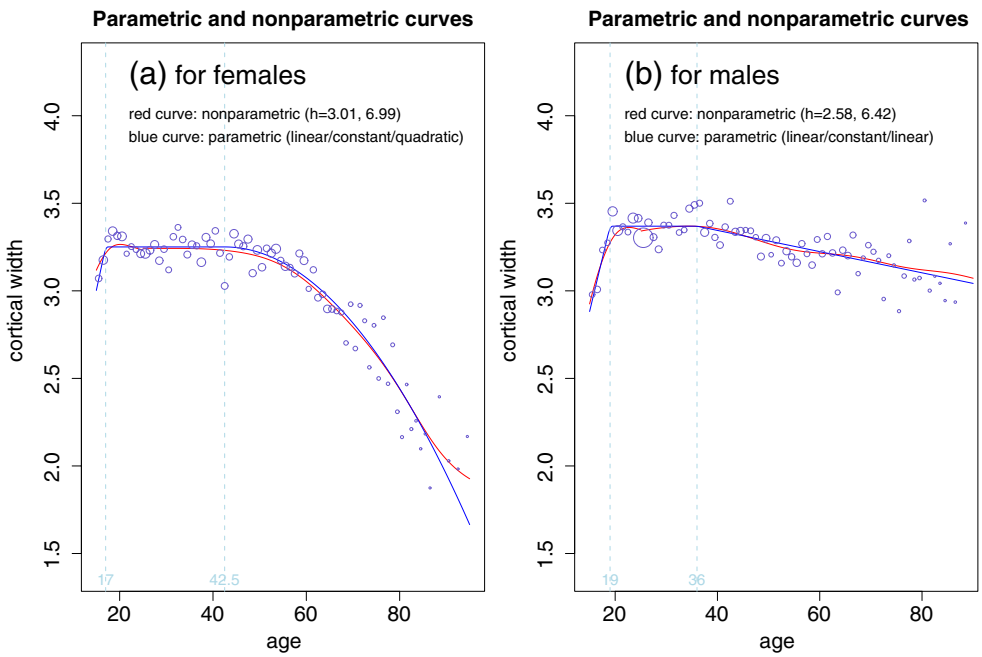


Fig. 6 Parametric and nonparametric curves **a** for females and **b** for males

width allows us to estimate referral rates implied by the application of various width thresholds, on which referral decisions might be made.

We observed that accelerated cortical thinning in women occurs after about 42.5 years of age. A number of other studies have produced results with which this observation may be compared. Morita et al. [14] measured the mandibular cortical width in 80-year-old men and women. They found that the prevalence of severe mandibular cortical erosions was ten times higher in the female sample than in the males (58.8% vs 5.9%). Figure 2 shows that this is where the differences in mean values of cortical thickness between sexes would be large. The accelerated decline in cortical width of women after the age of 50 years has been observed at various anatomical sites. Riggs et al. [15] showed that cortical thinning accelerated in the radius only after the age of 50 years in normal women, whereas men were little affected. Hyldstrup and Nielsen [16] used the "metacarpal index" (the ratio of the radiographic cortical thickness of the metacarpal bone to the total mid-metacarpal diameter) as a measure of osteoporosis. Using thickness measurements on digital radiographs, the index was maximal in the third decade and declined with age [16]. However, it needs to be taken into account that the metacarpal index is reduced by periosteal apposition, as well as endosteal cortical resorption [17].

Thinning of the femoral cortex with age is more rapid than the decline in areal BMD measured at the femoral neck, and it is the precipitous decline in cortical bone width at the hip which is present in hip fracture [18]. We have found a similar pattern of cortical bone loss in the mandible, which raises the intriguing question as to whether the mandibular measurements could be used to predict hip fracture.

We have previously proposed that mandibular cortical thickness measured from dental panoramic radiographs may be used as an opportunistic method for detecting osteoporosis [5]. Our earlier recommendation, based on manual measurements, was that a mandibular cortical width below the mental foramen of less than 3 mm merited referral of women for investigation of osteoporosis. The present study utilised a more lateral measurement site on the mandible (Fig. 1), which had previously been shown to have a better efficacy of detection of skeletal osteoporosis than measurement at the mental foramen [4]. Using this previously published dataset of females, the ASM/AAM method detected osteoporosis at the femoral neck with a sensitivity of 78% and specificity of 80% at a mandibular cortical width threshold value of 2.75 mm. In the analysis of the large dataset reported here, we can calculate where this threshold lies on the age-dependent width distributions. At age 40 years, it lies 1.2 standard deviations below the

Table 4 Estimated mean cortical widths at the age of 40, 50, 60 and 70 years together with standard errors for males and females

	Estimated mean cortical width	
Age (years)	Females (mm)	Males (mm)
40	3.25 (se=0.01)	3.35 (se=0.01)
50	3.22 (se=0.01)	3.28 (se=0.01)
60	3.07 (se=0.01)	3.22 (se=0.01)
70	2.82 (se=0.02)	3.16 (se=0.02)

mean female width, reducing at ages 50, 60 and 70 to 1.1, 0.7 and 0.1 standard deviations, respectively. As osteoporosis is rare before age 40, we estimate that using this threshold as a criterion for referral would result in a false-positive rate of referral of low bone mass of 12% in pre-menopausal women. Law et al. [19] found a similar false-positive rate of 15% when using bone density measurements at the femoral neck as a screening tool for predicting hip fractures, based on a cut off of 1 standard deviation below the mean density of the controls. Our threshold would refer 50% of women at age 71 and 34% at 65 years. The latter figure is around double the estimated National Health and Nutrition Examination Survey (NHANES) III prevalence rate of 15% at 65 [20], although the increase of 22% above our baseline false-positive rate of 12% for 40 years old is closer to the NHANES figure. The previously published prevalence figures for osteopenia in the femoral neck in women aged over 50 years is 50% [20], and we may be detecting a proportion of these. From Table 3, the rate of decrease in mean cortical width is $2b(x-42.5)$mm/year for females with age x in the 60–80 range, which increases from 0.049σ per year at $x=60$ to 0.105σ per year at $x=80$. This is in line with the published [21] reduction in BMD T-score in the UK, which varied between 0.25 and 1.3 per decade for femoral neck at different centres.

There has been no previous study on what mandibular cortical width threshold should be used for ageing males, and this study does not have BMD data to allow a definitive recommendation. However, given the female threshold, we make the provisional recommendation that males aged over 65 with a mandibular cortical width below the same threshold of 2.75 mm should consider DXA screening. This translates to 1 standard deviation below the male mean at age 65.

Measurement of the mandibular cortical width has been shown to be a poor predictor of fractures in an elderly study group, but this statistically nonsignificant result may be due to the small numbers of patients who developed fractures in some studies [22]. In our previous work, greater sensitivity in osteoporosis detection was introduced by using mandibular cortical width in combination with other clinical risk factors [1], but no other clinical risk factors were available for our sample in the present study.

It was a limitation of our study that definitive absolute scale measurements (e.g. including a ball bearing of known size in the dental panoramic tomogram) were not available relating image distance to anatomical distance. However, the previous study [6] included data using the same scanner, and in that study, absolute scale was measured using a ball bearing of known size that was placed intra-orally and incorporated into the radiographic image. This was used to calibrate for differences in magnification between images. As the images in the current study had been collected for other reasons, no calibration object was used. The standard error on the calculated mean scale (derived by bootstrap resampling of the data in [6]) was 1.2%. The limitations of our study also include a few unusually high or low averages in the cortical widths, e.g. at age 42.5 in both the male and female data.

Many osteopenic or osteoporotic female subjects, aged over 60 years, are likely to have cortical bone loss from both the hip and the mandible. The rate of loss of bone from the hip accelerates exponentially with age and follows a time course which is very similar to that observed for the mandibular bone [23]. The similarities with our results imply a phenomenon that is driving bone loss in older women, but which leaves men relatively unaffected. It is in this age group of women that the automated detection of cortical width will prove to be most useful in osteoporosis diagnosis and prevention of hip fractures.

Acknowledgements We would like to acknowledge the assistance of Herman Pauwels in the data collection for our sample. The study was supported financially by the Dunhill Medical Trust.

Conflicts of interest The authors declare that there are no conflicts of interest.

References

1. Devlin H, Allen PD, Graham J, Jacobs R, Karayianni K, Lindh C, van der Stelt PF, Marjanovic E, Adams JE, Pavitt S, Horner K (2008) The role of the dental surgeon in detecting osteoporosis: the OSTEODENT study. Br Dent J 204:E16. doi:10.1038/sj.bdj.2008.317
2. Ledgerton D, Horner K, Devlin H, Worthington H (1999) Radiomorphometric Indices of the mandible in a British female population. Dentomaxillofac Radiol 28:173–181
3. Yosue T, Brooks S, Arbor A (1989) The appearance of the mental foramina on panoramic and periapical radiographs. II. Evaluation of patients. Oral Surg Oral Med Oral Pathol 68:488–492
4. Allen PD, Graham J, Farnell DJJ, Harrison EJ, Jacobs R, Karayianni K, Lindh C, van der Stelt PF, Horner K, Devlin H (2007) Detecting reduced bone mineral density from dental radiographs using statistical shape models. IEEE Trans Inf Technol Biomed 11:601–610
5. Horner K, Devlin H (2002) Detecting patients with low skeletal bone mass. J Dent 30:171–175
6. Devlin H, Allen PD, Graham J, Jacobs R, Karayianni K, Lindh C, van der Stelt PF, Marjanovic E, Adams JE, Pavitt S, Horner K (2007) Automated osteoporosis risk assessment by dentists: a new pathway to diagnosis. Bone 40:835–842
7. Cootes TF, Taylor CJ, Cooper DH, Graham J (1995) Active shape models—their training and application. Comput Vis Image Underst 61(1):38–59
8. Roberts MG, Graham J, Devlin H (2010) Improving the detection of osteoporosis from dental radiographs using active appearance models. In: Niessen W, Meijering E (eds) Proceedings of the IEEE international symposium on biomedical imaging, Rotterdam, 14–17 April 2010, pp 440–443
9. Cootes TF, Edwards GJ, Taylor CJ (2001) Active appearance models. IEEE Trans Pattern Anal Mach Intell 61:38–59

10. Bowman AW, Azzalini A (1997) Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations. Oxford University Press, Oxford
11. Venables WN, Ripley BD (2002) Modern applied statistics with S., 4th edn. Springer, New York
12. Lindstrom M (2000). Contribution to online discussion thread [S] Piecewise Linear Regression, S-News Archive, Division of Biostatistics, School of Medicine, University of Washington in St. Louis. http://www.biostat.wustl.edu/archives/html/s-news/2000-04/msg00209.html. Accessed 24 Apr 2000
13. Bacon DW, Watts DG (1971) Estimating the transition between two intersecting straight lines. Biometrika 58(3):525–534
14. Morita I, Nakagaki H, Taguchi A, Kato K, Murakami T, Tsuboi S, Hayashizaki J, Inagaki K, Noguchi T (2009) Relationships between mandibular cortical bone measures and biochemical markers of bone turnover in elderly Japanese men and women. Oral Surg Oral Med Oral Pathol Oral Radiol Endod 108:777–783
15. Riggs BL, Wahner HW, Dunn WL, Mazess RB, Offord KP (1981) Differential changes in bone mineral density of the appendicular and axial skeleton with aging. J Clin Invest 67:328–335
16. Hyldstrup L, Nielsen SP (2001) Metacarpal index by digital X-ray radiogrammetry: normative reference values and comparison with dual X-ray absorptiometry. J Clin Densitom 4:299–306
17. Adams P, Davies GT, Sweetnam P (1970) Osteoporosis and the effects of ageing on bone mass in elderly men and women. Q J Med 156:601–615
18. Thomas CD, Mayhew PM, Power J, Poole KE, Loveridge N, Clement JG, Burgoyne CJ, Reeve J (2009) Femoral neck trabecular bone: loss with aging and role in preventing fracture. J Bone Miner Res 24:1808–1818
19. Law MR, Wald NJ, Meade TW (1991) Strategies for prevention of osteoporosis and hip fracture. BMJ 303(6800):453–459
20. Looker AC, Orwoll ES, CC JJR, Lindsay RL, Wahner HW, Dunn WL, Calvo MS, Harris TB, Heyse SP (1997) Prevalence of low femoral bone density in older U.S. Adults from NHANES III. J Bone Miner Res 12:1761–1768
21. Holt G, Khaw KT, Compston RDM, JE BA, Woolf AD, Crabtree NJ, Dalzell N, Warldey Smith B, Lunt M, Reeve J (2002) Prevalence of osteoporotic bone mineral density at the hip in Britain differs substantially from the US over 50 years of age: implications for clinical densitometry. Br J Radiol 75(897):736–742
22. Okabe S, Morimoto Y, Ansai T, Yoshioka I, Tanaka T, Taguchi A, Kito S, Wakasugi-Sato N, Oda M, Kuroiwa H, Ohba T, Awano S, Takata Y, Takehara T (2008) Assessment of the relationship between the mandibular cortex on panoramic radiographs and the risk of bone fracture and vascular disease in 80-year-olds. Oral Surg Oral Med Oral Pathol Oral Radiol Endod 106:433–442
23. Ensrud KE, Palermo L, Black DM, Cauley J, Jergas M, Orwoll ES, Nevitt MC, Fox KM, Cummings SR (1995) Hip and calcaneal bone loss increase with advancing age: longitudinal results from the study of osteoporotic fractures. J Bone Miner Res 10:1778–1787

43. **Image texture in dental panoramic radiographs as a potential biomarker of osteoporosis.** M.G. Roberts, J. Graham, H. Devlin, *IEEE Trans. Biomedical Engineering*, ***60(9)***, *2384 – 2392, 2013*. doi: 10.1109/TBME.2013.2256908

Image Texture in Dental Panoramic Radiographs as a Potential Biomarker of Osteoporosis

Martin G. Roberts, James Graham*, *Member, IEEE*, and Hugh Devlin

***Abstract*—Previous studies have shown an association between osteoporosis and automatic measurements of mandibular cortical width on dental panoramic radiographs (DPRs). In this study, we show that additional image texture features increase this association and propose the combined features as a potential biomarker for osteoporosis. We used an existing dataset of 663 DPRs of female patients with bone mineral density (BMD) measurements. The mandibular cortex was located using a previously described computer algorithm. Texture features, based on co-occurrence matrices and fractal dimension, were measured in the bone within the cortex and also in the superior basal bone above the cortex. These, augmented by cortical width measurements, were used by a random forest classifier to identify osteoporosis at femoral neck, total hip, and lumbar spine. Classification performance was assessed by ROC analysis. Area-under-curve (AUC) values for identifying osteoporosis at femoral neck were 0.830, 0.824, and 0.872 using, respectively, cortical width alone, cortical texture (co-occurrence matrix features) alone, and combined width and texture. At 80% sensitivity, these classifiers produced specificity values of 74.4%, 73.6%, and 80.0%, respectively. Fractal dimension was a less effective texture feature. Prediction of osteoporosis at the lumbar spine was poorer, but a combined width and superior basal bone texture classifier gave a significant improvement in AUC at $p < 0.05$ over the use of width alone.**

***Index Terms*—Co-occurrence matrices, dental panoramic radiograph (DPR), image texture, osteoporosis, random forest classifier.**

I. Introduction

OSTEPOROSIS is a progressive skeletal disease characterized by low bone mass and structural deterioration of bone tissue, leading to an increased susceptibility to fragility fracture. It is associated with increased morbidity and mortality: 27% of women who sustain a hip fracture die within 1 year [1]. Early detection of osteoporosis can allow therapeutic intervention, but the condition is often undiagnosed. There has been recent interest among dental researchers in identifying those at risk of reduced bone mineral density (BMD) from dental radiographs [2], [3]. Fig. 1 shows a dental panoramic radiograph (DPR), on which the inferior mandibular cortex is visible. It was reported in [3] that measuring the thickness of the cortical bone using active shape model (ASM) search [4] provides a good diagnostic of low BMD at other skeletal sites. Roberts *et al*. [5] showed that the reduction in the mean width of the mandibular cortex with age followed a similar curve to systemic BMD loss in post-menopausal females. As DPRs are often taken by dentists, they provide a useful opportunity for diagnosis of osteoporosis, without requiring additional radiation exposure to the patient. At-risk patients can be referred to their general medical practitioner for advice about further investigation.

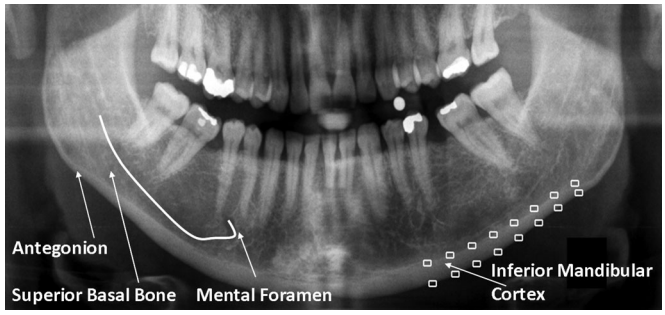


Fig. 1. Dental panoramic radiograph. The inferior mandibular cortex, antegonion, and mental foramen are indicated. The course of the mandibular canal is marked on the left side of the picture. The points marked around the cortex on the right side of the picture indicate the positions of the superior and inferior cortical border determined by image search. The cortical width is measured at a point close to the antegonion. Image texture is measured in the cortex between the antegonion and the mental foramen on both sides of the image and over the same extent in the superior basal bone in a region bounded by the cortex and the mandibular canal (indicated).

Other researchers [6]–[9] have examined the links between bone loss at multiple skeletal sites and various texture measures applied to DPRs (or intra-oral radiographs) in human cases and animal models [10], [11]. In this study, we examined a variety of texture measures applied to an existing dataset of DPRs from 663 females. Machine learning methods were applied to a large vector of image texture features measured within and around the mandibular cortex as predictors of osteoporosis. We also investigated prediction via a combined classifier using both cortical texture and width measurements.

II. Materials and Methods

A. Data

The dataset had been already collected during a previous study [2] and consisted of DPRs for 663 ambulant female

Manuscript received December 20, 2012; revised March 21, 2013; accepted March 26, 2013. Date of publication April 4, 2013; date of current version August 16, 2013. This work was supported by a research grant from the Dunhill Trust and used data previously acquired by a study funded by a technological development project grant from the European Commission (Osteodent, QLK6-2002-02243). *Asterisk indicates corresponding author.*

M. G. Roberts is with the Centre for Imaging Science, Institute of Population Health Sciences, Manchester Academic Health Sciences Centre, University of Manchester, Manchester, M13 9PT, U.K. (e-mail: zetamgr@gmail.com).

*J. Graham is with the Centre for Imaging Science, Institute of Population Health Sciences, Manchester Academic Health Sciences Centre, University of Manchester, Manchester, M13 9PT, U.K. (e-mail: jim.graham@manchester.ac.uk).

H. Devlin is with the School of Dentistry, Manchester Academic Health Sciences Centre, University of Manchester, Manchester, M15 6FH, U.K. (e-mail: Hugh.Devlin@manchester.ac.uk).

Digital Object Identifier 10.1109/TBME.2013.2256908

0018-9294 © 2013 IEEE

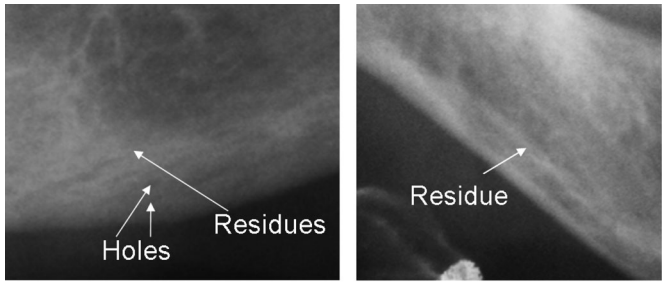


Fig. 2. Details of two DPRs showing the porous appearance of the cortex. (a) Holes and residues in the cortex of an osteoporotic patient. (b) Presence of an extensive residue creates uncertainty in the true position of the superior cortical border.

patients together with BMD values determined by dual energy X-ray absorptiometry (DXA). Robust radiographic protocols and regular calibration were applied to ensure consistency of radiographic and densitometric data. Patients were diagnosed osteoporotic according to the World Health Organization criteria, i.e., those with a BMD standardized T-score value below -2.5, evaluated at three skeletal sites (femoral neck, total hip, and lumbar spine). There were 140 patients who were osteoporotic at one or more sites.

B. Localization of the Mandibular Cortex

Devlin *et al*. [3] described location and measurement of the width of the mandibular cortex using an active shape model (ASM) [4]. The algorithm and independent training set are described in more detail in [12]. Roberts *et al*. [13] improved the search algorithm by extending the modeled region and by using an active appearance model (AAM) [14] search after an initial ASM, and this improved method was also used in this study. ASM and AAM use image search, which can proceed automatically to produce a delineation of the edges of the cortex. However, fewer search failures occur if the search is initialized from four landmark points defined interactively [12]. Here, we are interested in the effect of texture features, so the latter approach is taken to isolate errors due to texture measurement from those resulting from occasional search failures. The width measurements, and texture measurements reported here, were taken in the region of the cortex between the mental foramen and the antegonion (see Fig. 1).

C. Holes and Residues

One problem with the cortical width measurement is that the superior mandibular border (the endosteal margin) may become unclear, especially in osteoporotic cases in which the bone may be eroded more in some regions than in others. This can lead to "holes" within the cortex, where there has been significant bone loss; conversely, there may be bone islands or "residues" remaining above the line of the superior border (see Fig. 2). As a result, the upper edge of the cortical border becomes ill-defined, leading to errors in width measurement [see Fig. 2(b)]. The purpose of this study was to examine image texture features that might capture the appearance of these holes and residues, and use these in a statistical classifier to improve the diagnosis.

The mandibular cortical index (MCI) [15] is a visual assessment scale that has been developed to assess osteoporosis in the cortical area of the mandible using DPRs. In this technique, the inferior cortex is classified into three groups according to the following criteria:

1) MCl 1: The endosteal margin of the cortex is even and sharp on both sides of the mandible.
2) MCl 2: The endosteal margin has resorptive cavities with cortical residues one to three layers thick on one or both sides.
3) MCl 3: The endosteal margin consists of many cortical residues and is clearly porous.

Taguchi *et al*. [16], [17] have studied the diagnostic capability of the MCI, and found that it has a significant predictive value. For our study, an expert dental practitioner (HD) graded the data on the MCI 1-3 system. In initial investigations, we examined whether there were significant differences in sample means in putative texture features between patients in the MCI $= 1$ group, and a group of patients with an MCI of either 2 or 3. We also compared sample means for osteoporotic patients and nonosteoporotic individuals as a means of determining which features to include in a classification process.

D. Image Normalization

Image texture values are potentially sensitive to image exposure. While the data collection protocol sought to maintain image values within a consistent range, image variation is inevitable. For this reason, we applied a robust image normalization procedure prior to extraction of texture features.

The data are placed on (0, 1), with value 0.5 at the mean, according to

$$g' = \frac{1}{2}\left(1 + \frac{g - \bar{g}}{\sqrt{(g-\bar{g})^2 + 3\sigma^2}}\right). \tag{1}$$

This is essentially the square-rooted Geman–McClure kernel function [18] with maximum influence at σ. We use robust estimators for $\bar{g}$ and σ to allow for holes in the cortex, and also any overlaid bright artifacts, which are sometimes present on DPRs. For σ, we use the robust S_n estimate [19] of the standard deviation within the cortex, and $\bar{g}$ is a robust estimate of mean, derived by starting from the median and then iteratively updating the mean using a Geman–McClure weighting kernel given the deviation from the current estimated mean, and kernel scaling $\sqrt{3}\sigma$. The S_n statistic does not require an initial estimate of mean, as it uses only inter-point deviations.

The sigmoidal shape of this normalization ensures that extreme high and low image values become drawn closer to 1 and 0, respectively.

E. Fractal-Based Texture Measurement

Several authors have quantified bone texture in radiographs using fractal dimension. Of particular relevance here, Yasar *et al*. [9] showed that texture measured by a box-counting

fractal dimension gave significant differences between samples of MCI 1 and MCI 2 or 3. A similar method applied to radiographs of the hip [20] had shown that several fractal dimension measures correlated with patient age. These results suggest that image texture, in particular measured by fractal dimension, might be used to discriminate between osteoporotic and nonosteoporotic bone. The method used in [9] requires binary segmentation of the gray-scale image. The fractal dimension is calculated by counting the number of "1" pixels within boxes at a range of scales. We sought to avoid the use of an arbitrary threshold and adopted the method of Rose *et al*. [21] in analyzing the spatial heterogeneity of tumors. This box-counting method treats the image as a 3-D landscape.

In [22], Rose *et al*. used an alternative measure of fractal dimension based on Rényi entropy [23]. The Rényi entropy is a family of functions, parameterized by a unit-less scalar $q \geq 0$, and so, the Rényi dimensions are a family of fractal dimensions defined in

$$d_q = \lim_{s \to 0} \frac{\log \sum_i g_i^q}{(1-q)\log\left(\frac{1}{s}\right)} \tag{2}$$

where the gray levels are normalized so that $\sum_i g_i = 1$.

In effect, we are using the normalized gray level as a pseudo-probability density function for the distribution of bone within the cortex. The limit as $q \to 1$ (also known as the Rényi information dimension) uses the conventional Shannon entropy. We computed the Rényi dimensions for $q = 1, 2, 4$.

F. Co-Occurrence Matrices

There is no reason to believe that dental panoramic images display any fractal structure (such as self-similarity across scale), and the fractal dimension here and in the previous studies merely acts as a general "roughness" measure. As an alternative approach, we applied classical Haralick texture features based on gray-level co-occurrence matrices [24]. In this method, a 2-D histogram is constructed, representing the frequency of occurrence of pairs of gray values in pixels separated by a specified vector. The magnitude and direction of the vector are application-dependent parameters.

In this case, the normal to the superior cortical border is a natural local direction for this vector, as it is likely to respond to residues. The 2-D histograms of gray-level co-occurrence were computed for a range of pixel separations (from 2 to 8) along these local normals. The histograms were then normalized to probability distributions. For both co-occurrence features and fractal features, a band of pixels above the cortex was included in the texture calculation to allow for the size of the sampling window. In the case of the co-occurrence matrix features, the width of this sampling band was 10 and 16 pixels for fractal dimension calculation, to accommodate the increasing scales of box sizes used in box-counting. A large number of features can then be calculated which encode different characteristics about these distributions (see [24] for a full list of these). We evaluated the first 12 of the features defined in [24] setting the gray-scale quantization to $N_g = 32$, and then retained nine features which seemed to show significant differences between the manually scored MCI grades and also between osteoporotic and nonosteoporotic individuals (see Section III for details).

These features at the various scales were then input to a classifier. The range of scales (pixel separations) and width of the image band above the cortex are variable parameters of the method which were investigated in preliminary experiments. Although the classifier performance did not depend strongly on these, a roughly optimal performance was obtained by taking separations from 2 to 5 pixels.

G. Use of Superior Basal Bone Above the Cortex

We also computed the same set of fractal and co-occurrence matrix features for the basal bone situated above the cortex, which we refer to as the *superior basal bone*. The region used extended from 2 pixels above the notional cortical superior border to 2 pixels below the mandibular canal containing the alveolar nerve [see Fig. 1(a)]. The position of the inferior border of the mandibular canal was marked on the images using four interactively positioned points immediately below the mandibular canal on each side between the mental foramen and antegonion. Spline interpolation was then used to estimate the mandibular canal location, and the small border of a further 2 pixels was set to avoid sampling the brighter texture of the canal itself. In contrast to the detection of residues near the cortical border, there is no obvious sampling direction for the co-occurrence matrix in the superior basal bone. We therefore computed the co-occurrence features for four directions (normal to the cortical border, tangential to it, and the two diagonal directions). In view of the increased set of directions, and the fact that the separation between the mandibular canal and the cortical border can sometimes be quite small, we evaluated the co-occurrence features only at separation distances of 2 and 4 pixels.

The bone above the mandibular canal was not used to avoid sampling into the teeth, which would have a large effect on the texture results.

H. Classification Method

Because of the potentially large number of texture features, some of which may be weak and noisy, we used a random forest classifier [25]. This extends the ideas of a classification and regression tree (CART) [26], in which the dataset is recursively divided according to the decision variable threshold which best separates the training data into child nodes of different classes. Some measure of the mixing impurity is minimized, (e.g., entropy [26]) to select the branching criteria. In a random forest, many such trees are built by performing bootstrap aggregation (randomly selecting data points from the sample with replacement, also known as *bagging*) which helps to avoid overtraining. Furthermore, at each decision node, the best branch of a randomly selected subset of the possible decision variables is taken (we used a subset size of $\sqrt{n}$ for n decision variables). This increases the independence of the trees in the forest. Each single tree can then produce a (possibly weak) estimate of the probability of the object's classification, by taking the decision variable set through the tree's branching nodes. The output probability of that tree is then the ratio of class types at the final decision

TABLE I
CORTICAL BONE TEXTURE MEASURES FOR SEPARATING GROUPS

Dimension Measure	Mean MCI 1	Mean MCI≥2	t(MCI)	Mean Non-Ost	Mean OstF	t(OstF)	Mean OstL	t(OstL)
Fractal features								
Box Counting	2.7921	2.7801	6.0	2.7931	2.7723	5.5	2.7755	6.1
Rényi entropy q=1	2.0260	2.0223	3.1	2.0268	2.0161	5.2	2.0188	4.8
Rényi entropy q=2	2.0197	2.0131	4.3	2.0209	2.0043	6.4	2.0080	6.0
Rényi entropy q=4	2.0290	2.0187	5.5	2.0303	2.0093	6.4	2.0132	6.5
Co-occurrence features								
Angular 2nd Moment	0.0154	0.0168	1.78	0.0156	0.0180	2.66	0.0162	1.02
Contrast	5.338	6.030	3.97	5.286	6.578	6.60	6.255	6.85
Correlation	0.845	0.814	5.40	0.847	0.802	6.77	0.806	8.80
Sum entropy	3.344	3.313	5.27	3.343	3.302	5.80	3.314	6.50
Inverse Difference Moment	0.430	0.427	0.49	0.433	0.413	2.93	0.414	3.79
Difference Variance	0.380	0.447	5.37	0.376	0.482	7.35	0.460	8.32
Difference Entropy	1.710	1.759	2.72	1.703	1.812	5.68	1.790	6.15
Information correlation 1	-0.276	-0.256	3.61	-0.278	-0.242	6.19	-0.249	6.77
Information correlation 2	0.870	0.849	4.44	0.872	0.835	7.13	0.841	7.94

Mean texture values for fractal dimension and co-occurrence matrix features for separating MCI 1 and MCI 2 or 3 groups. Similarly shown are means of osteoporotic (OstF denotes at femoral neck, OstL denotes at lumbar spine) and nonosteoporotic (not osteoporotic anywhere) groups. t_{MCI} gives the number of standard errors between the MCI 1 and 2 & 3 sample means, and t_{OstF} and t_{OstL} likewise between the osteoporotic and nonosteoporotic groups as diagnosed at the femoral neck or lumbar spine, respectively. The co-occurrence measures shown are taken from the list defined in [24]. In the case of cortical width features, the nine texture measures shown gave significant separations. Four further features (variance, sum average, sum variance, and entropy) gave insignificant or only marginally significant separation at most scales, and results are omitted in the interests of space.

node. We used a minimum node size of five samples. So, for example, if a tree branch terminates with one normal and four osteoporotic cases in training, then the estimate of osteoporosis probability at that node is 0.8. These probabilities are then averaged over all the trees in the forest to produce a final estimate of the probability of a positive classification. Random forests are known to work well in combining large numbers of weak and noisy features, and are robust against the addition of noise variables [25].

The use of the bootstrap aggregation also means that unbiased estimates of population classification performance can be obtained without the additional complexity of multiple train/test cross-validation cycles. Instead, an *out-of-bag* (OOB) estimate [25] is obtained by classifying a training example using only those trees which did not use that example as part of their bootstrapped training sample. This will use around 37% ($1/e$) of the total forest. We used large forest sizes (10 000 trees) so that this reduction should be immaterial. Therefore, we did not separate the training set into train and test sets, but produced one large forest for all of the data, and then used OOB estimates for predicting the probability of osteoporosis of each patient. A varying detection threshold was applied to this probability to generate receiver-operating characteristic (ROC) curves [27].

Separate classifiers were trained for the cortical bone and the superior basal bone between the cortex and the mandibular canal.

I. Combination of Classifiers

An obvious approach to combining the width and texture features is to simply use all the texture features and the two cortical widths (left and right) together in one random forest. However, because the width is a much better single predictor than any single texture feature (see Section III), this is not necessarily the best combination due to the random feature subset selection at each tree node. Our previous experience in using unbalanced combinations of features indicated that a better approach was to use a cascade, first training a separate texture classifier for each half of the cortex, and then training a final random forest classifier using the two cortical widths and the predicted probability of osteoporosis from the two texture classifiers.

J. Comparison of Classifiers

We wished to determine whether the combined classifier was a statistically significant improvement over the classification using width alone. For this, we used two methods: area under the ROC curve (AUC) and comparison of sensitivity and specificity at selected operating points. While differences in AUC can quantify differences in classification performance, AUC is influenced by regions of the ROC curve that have little practical relevance (low sensitivity or specificity). It is also possible that two ROC curves with similar values of AUC may still be significantly different at important regions. For this reason, we also compared the specificities of a set of operating points in the 70–90% sensitivity region (at 5% intervals) using McNemar's test [27], [28].

We estimated the distribution of the difference in AUC by performing a smoothed bootstrap [29]. This involves randomly resampling with replacement from the sample, combined with kernel smoothing, and is explained in detail in the Appendix.

III. RESULTS

A. Texture Feature Evaluation

Table I shows the results of various texture features applied to cortical bone in separating the manually categorized MCI 1 from MCI 2&3 groups and in separating nonosteoporotic (at any of the three sites) and osteoporotic groups. For compactness, only the most significant results are shown; hence, only osteoporosis at the femoral neck (OstF: 66 cases) and osteoporosis at the lumbar spine (OstL: 120 cases) are included.

TABLE II
SUPERIOR BASAL BONE TEXTURE MEASURES FOR SEPARATING GROUPS

Dimension Measure	Mean Non-Ost	Mean OstF	t(OstF)
Tangential Direction			
Contrast	0.833	0.760	2.00
Correlation	0.943	0.937	1.55
Difference Variance	0.0646	0.0552	3.35
Information correlation 1	-0.512	-0.491	2.63
Information correlation 2	0.912	0.901	2.10
Normal Direction			
Contrast	3.11	2.11	5.88
Correlation	0.728	0.763	1.91
Difference Variance	0.297	0.185	6.06
Sum Average	8.01	7.30	1.68
Sum Variance	60.50	50.91	1.55

Mean texture values for co-occurrence matrix features for separating osteoporotic groups (at femoral neck—OstF) and nonosteoporotic (not osteoporotic anywhere) groups (cf., Table I). t_{OstF} gives the number of standard errors between osteoporotic and nonosteoporotic groups . In contrast to Table I, only the co-occurrence features shown (with a pixel spacing of four) resulted in significant separation of the groups.

In the case of the cortical bone features, there are significant differences between the population means for MCI value 1 and the 2 & 3 combined category for the fractal features: both the box counting dimension (after taking the mean value of the left and right halves of the cortex) and the Rényi entropy dimensions. In all these cases, there is a reduction in the fractal dimension measure with osteoporosis, indicating that less of the space is being filled. There is a similar reduction in fractal dimension for the osteoporotic cases in comparison to the nonosteoporotic group, with even more significant separation of population means, especially for patients who were osteoporotic at the femoral neck.

Table I also shows the nine co-occurrence matrix features that gave the largest separation between MCI groups and osteoporotic and nonosteoporotic groups at most scales, with a pixel separation of 4. (For details of these features, readers are directed to [24].)

Table II shows the corresponding results for the superior basal bone texture. In this case, no significant difference was found between MCI categories or osteoporotic and nonosteoporotic cases for any fractal dimension measure. Some significant differences were found for a small set of co-occurrence features. The table shows the mean and standard error separations of means, indicating some significant separation (or close to significant at some pixel distance) for directions normal to and tangential to the cortical border for an inter-pixel distance of 4. Results for the diagonal directions are similar to those along the cortical border and are omitted in the interests of space.

B. Classification of Osteoporotic Individuals

The classification results are summarized in Table III, with specific points highlighted in the following sections. The table shows the AUC value and the false positive rate at selected values of sensitivity.

1) Cortical Bone Features—Fractal Dimension: We trained a random forest classifier using both box counting dimensions and Rényi dimensions on the left and right sides of the image.

TABLE III
PREDICTION OF OSTEOPOROSIS

	AUC	FPR at 70% sensitivity	FPR at 80% sensitivity	FPR at 90% sensitivity
Osteoporosis at femoral neck				
Fractal Dim	0.720	33.0%	44.1%	54.2%
TextureC	0.824	22.8%	25.6%	42.8%
TextureB	0.662	40.5%	46.4%	82.5%
Width	0.830	20.0%	26.4%	43.5%
Combined Fractal/Width	0.844	16.8%	23.1%	38.3%
Combined TextureC/Width	0.872	17.4%	20.0%	28.2%
Combined TextureB/Width	0.850	18.3%	25.2%	46.2%
Osteporosis at lumbar spine				
Fractal Dim	0.638	48.2%	55.3%	77.3%
TextureC	0.730	36.3%	44.1%	64.8%
Width	0.802	23.1%	31.5%	60.7%
Combined Fractal Dimension/Width	0.802	22.7%	29.5%	62.4%
Combined TextureC/Width	0.800	21.1%	29.3%	62.6%
Osteoporosis at any of 3 sites				
Fractal Dim	0.617	55.1%	56.5%	94.2%
TextureC	0.730	37.2%	49.1%	64.5%
TextureB	0.629	49.8%	58.5%	83.2%
Width	0.800	27.3%	34.0%	52.7%
Combined Fractal Dimension/Width	0.799	24.1%	32.5%	55.3%
Combined TextureC/Width	0.804	23.1%	31.7%	48.1%
Combined TextureB/Width	0.820	21.1%	27.3%	50.2%

AUC and False Positive Rates (FPR) for prediction of osteoporosis at the femoral neck, lumbar spine, and any of the three sites for the various classifiers. TextureC refers to the texture of the cortex, and TextureB to the texture of the superior basal bone.

Despite the significant differences in population means, the fractal dimensions did not prove to be very useful for classifying individuals as osteoporotic. The AUC for predicting osteoporosis at the femoral neck was 0.720. The false positive rate at 75% sensitivity was 41.7%. Although this is better than random, the specificity is clearly poor for practical clinical use. A classifier trained on these fractal features and the two cortical widths gave an apparent slight improvement on AUC from one trained on width alone, increasing from 0.816 to 0.844. This was not significant at $p = 0.05$ (only at $p = 0.26$) on the bootstrap test. The false positive rate at 80% reduced from 26.4% to 23.1%.

The AUC of the fractal dimension classifier for predicting osteoporosis at the lumbar spine was substantially lower at 0.638, and the combined classifier was essentially indistinguishable from one trained on cortical width.

2) Cortical Bone Features—Co-Occurrence Matrix: We selected the features shown in Table I, which resulted in a significant difference at most separation scales, for inclusion into another random forest classifier.

Fig. 3 shows the ROC curves, produced by three random forest classifiers trained, respectively, on (1) the 72 dimensional texture feature vector (two sides, four scales, nine features), (2) left and right cortical widths, and (3) the combined classifier using the output of separate left and right texture classifiers and the widths. The ROC curves are for osteoporosis determined by

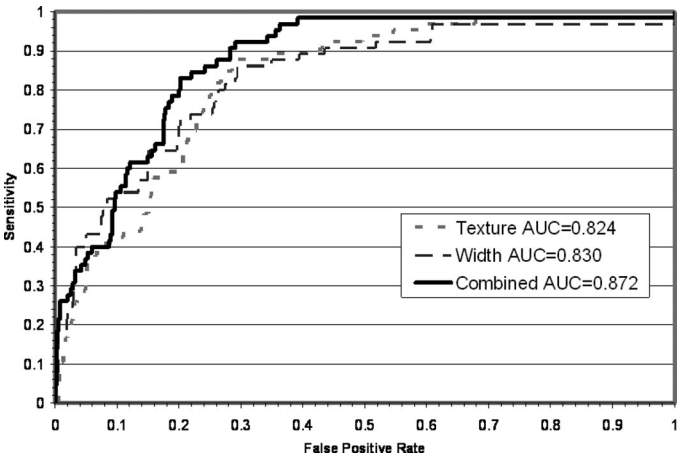


Fig. 3. ROC curves for detection of osteoporosis at the femoral neck. Three curves are shown for cortical texture classifier using co-occurrence features; classifier using left and right cortical widths; and the combined classifier using width and texture, as indicated in the key.

TABLE IV
MCNEMAR TEST STATISTICS FOR CLASSIFICATION

Sensitivity	70%	75%	80%	85%	90%
TextureC at femoral neck	9.77	17.75	23.52	91.39	65.7
TextureC at Lumbar spine	7.69	0.0	3.56	15.01	1.19
TextureC at any skeletal site	13.8	1.88	2.88	1.84	9,12
TextureB at any skeletal site	20.02	44.18	28.20	5.95	1.62

Comparing specificity differences between the combined texture/ width classifier and a width-only classifier. The table shows the McNemar statistic which should follow a χ_1^2 distribution under the null hypothesis. A significant result is reported if the statistic exceeds the 5% significance value for the χ_1^2 distribution (3.84). The first three comparisons (textureC) use the texture of the cortex and compare false positive rates when detecting osteoporosis at the femoral neck, lumbar spine, and any of the three sites, respectively. At the femoral neck, the difference between the combined classifier and the width-only classifier is significant at the 5% level for all sensitivity values. The significance of the difference reduces for the lumbar spine and any site. The last comparison (textureB) uses the superior basal texture and the comparison is for osteoporosis detected at any of the three skeletal sites. The 5% significance value is exceeded at all but 90% sensitivity.

BMD at the femoral neck (66 such cases). The corresponding AUCs are 0.824 (texture), 0.830 (width), and 0.872 (combined). These, together with the false positive rates at 70%, 80%, and 90% sensitivities, are shown in Table III. The difference in AUC between the combined classifier (cortical width plus cortical texture) and the width classifier is significant at $p = 0.05$, as the fifth percentile of the bootstrapped distribution of differences is positive (fifth percentile 0.011, bootstrapped median difference 0.039).

The additional benefit provided by the texture measures was confirmed at specific operating points by the results of the McNemar test on false positive rates for operating points in the 70–90% sensitivity range. The test statistics are given in Table IV, and are clearly all substantially larger than the 5% significance level of the χ_1^2 distribution (3.84). The improvement in specificity of the combined classifier appears to increase with increasing sensitivity so that at the 90% sensitivity point, the false positive rate reduces from 43.5% (width) to 28.2% (combined).

If osteoporosis is determined at the lumbar spine (120 cases) rather than femoral neck, then the prediction performance reduces (see Table III). The AUC for the texture classifier reduces substantially to 0.730 for osteoporosis at the lumbar spine, with a false positive rate of 44.1% at 80% sensitivity. There is no significant difference in overall AUC between the combined and width classifiers. At specific operating points (70% and 85%), the McNemar test indicates significant differences in specificity but the effect is marginal and much smaller than for the femoral neck (see Table IV).

When all three skeletal sites are used for a single diagnosis of osteoporosis (if any site is osteoporotic), then the results are similar to those for the lumbar spine, which in effect dominates (120 out of the overall 140 cases are osteoporotic at the lumbar spine). The combined classifier has a false positive rate at 80% sensitivity of 31.7% compared to 34.0% using only cortical width (see Table III). The bootstrap test indicates no significant difference in overall AUC, as zero difference is crossed at the 26th percentile. Similarly, the McNemar test indicates no significant difference in specificity between the combined classifier and a width classifier in the 75–85% sensitivity range, although the McNemar test does give a significant difference at both 70% and 90% sensitivity at $p = 0.05$. It appears that the state of the bone in the mandibular cortex is more strongly correlated with the femoral neck than other skeletal sites.

3) Superior Basal Bone Features: As no fractal dimension measure provided significant separation of groups (see Section III-A), we did not train a classifier with these features. A texture classifier was trained using the following co-occurrence matrix features: contrast, correlation, and difference variance (all orientations); information dimensions of correlation for the tangential orientation, and the sum average and sum variance for other orientations (see Table II).

The AUCs for superior basal texture classifiers are given in Table III. These were lower than for the cortical bone features and the classifiers on their own performed quite poorly. Nevertheless, the combined classifier for cortical width and superior basal texture performed better for predicting osteoporosis at any site (AUC 0.820). Although superior basal texture is poorer than cortical texture as a single predictor, it may be more independent of the cortical width measurement, and so provide a better overall predictor of bone status at predominantly trabecular sites such as the lumbar spine. Fig. 4 shows the ROC curves for predicting osteoporosis at any site for the superior basal texture, width, and combined classifiers.

The bootstrap test indicates a significant improvement in AUC for the combined classifier in predicting osteoporosis at any of the three sites ($p < 0.05$), and similarly significant differences are observed in the McNemar test comparing false positive rates at several sensitivities (see Table IV).

IV. DISCUSSION

Inoue and Ogawa [20] suggested the use of fractal dimension in analyzing trabecular patterns at the hip. Yasar and Akgunlu [9]

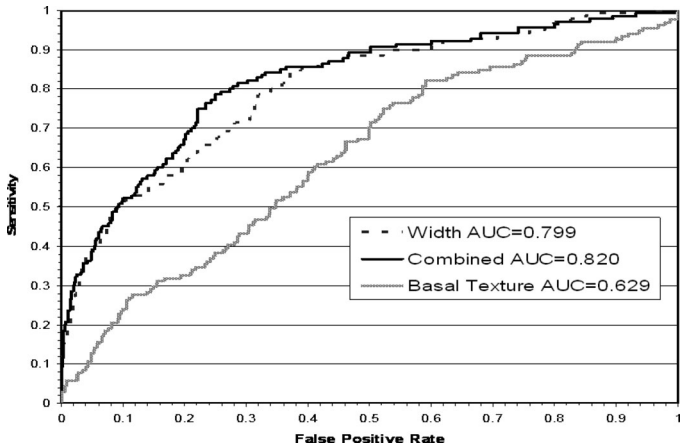


Fig. 4. ROC curves for detection of osteoporosis at any of the three skeletal sites; three curves are shown for superior basal texture classifier using co-occurrence features; classifier using left and right cortical widths; combined classifier using width and superior basal texture, as indicated in the key.

showed a significant difference in mean fractal dimension of the mandibular cortex for groups of patients with different MCI values. However, Southard *et al*. found no relation between the fractal dimension of the mandible and BMD at other skeletal sites for either human [30] or animal models [11]. Our results on a large dataset suggest that there is a relationship between various measures of fractal dimension and osteoporosis, particularly at the femoral neck. We have sought to improve on these rather formulaic measures of fractal dimension in a number of ways. First, we use a carefully implemented normalisation scheme based on robust statistics; second, we avoid arbitrary thresholds by regarding gray-level as an extruded third dimension; third, we use a family of different fractal dimensions in a multivariate classification framework. However, we conclude that although the relationship between the fractal dimension of the mandibular cortex and osteoporosis at other skeletal sites is clearly visible at the population level, it produces only poor specificity if used in predicting the osteoporotic status of individuals. This is due to the large variance of the fractal dimension compared to the small differences between the osteoporotic and nonosteoporotic groups. The standard deviations of box counting dimension in the nonosteoporotic and osteoporotic (femoral neck) groups are 0.0254 and 0.0275, respectively, whereas the separation in group means is 0.021. Similarly, the separation of group means for the Rényi correlation dimension ($q = 2$) is 0.66 times the standard deviation of the nonosteoporotic group. The substantial overlap between the two distributions results in poor classification performance. A similar problem appears implicit also in [9].

We investigated the use of co-occurrence matrix features, based on the intuition that they would efficiently capture the texture information about the clinically reported descriptions in terms of holes and residues. The magnitudes and directions of pixel separations were selected to correspond with this. This framework provides a larger set of features for this classification problem than the fractal measures. These features separately capture different aspects of the image brightness distribution across multiple scales resulting in a more sensitive classifier. There are, of course, many texture features that could be used [31] and other classifier designs that might have been employed. In selecting the random forest, we have used a classifier with well-attested performance, in which the bootstrap training regime allows us to make maximal use of the image dataset for training and evaluation. Our cortical texture classifier using co-occurrence features performs almost as well as cortical width as a predictor of osteoporosis at the femoral neck, and when combined with cortical width results in a statistically significant improvement in specificity in the most interesting region of the ROC curve. Lumbar spine osteoporosis is better predicted by combining cortical width with a texture classifier sampling the superior basal bone above the cortex, though this could be because of better response to cortical residues above the notional superior cortical border.

BMD at femoral neck and lumbar spine are highly correlated (Pearson's correlation $= 0.903$, s.e. $= 0.016$, $p = 0.001$). However, there appears to be a stronger relation between the texture of the mandibular cortex and BMD at the femoral neck, than at lumbar spine or other skeletal sites. This is consistent with the findings on cortical width in our earlier study [3] and might be expected since the mandibular cortex and femoral neck are both composed of primarily cortical bone, whereas the lumbar spine contains a greater proportion of trabecular bone. The cortical texture classifier has a false positive rate of 25.6% at 80% sensitivity in detecting low BMD at the femoral neck, while the corresponding false positive rate increases to 49.1% (specificity 50.9%) for osteoporosis at any of lumbar spine, total hip, and femoral neck. Nevertheless, this is a similar performance to expert human grading reported by Taguchi *et al*. [17], who found a sensitivity of 82.5% with specificity 46.2% for osteoporosis at either lumbar spine or femoral neck by diagnosing all MCI grade 2 or 3 patients as osteoporotic.

It is also possible to produce a cascaded classifier using both cortical and superior basal bone texture features, but this performs very similarly to the better of the single texture/width combination (i.e., width and superior basal texture for diagnosis at any site or width and cortical texture for diagnosis at the femoral neck).

Lerouxel *et al*. [10] used run-length moments as a texture measure of the alveolar bone in radiographs of the mandible of rats that had induced osteoporosis, and reported significant correlations between texture and densitometric parameters. Run-length moments are related to some co-occurrence features (e.g., longer run-lengths of the same pixel value will be observed with higher correlation in the co-occurrence distribution). However, we include a richer set of texture features which helps to boost the performance of the classifiers.

We suggest that our use of co-occurrence texture features and random forests may give a useful measure of bone quality at other skeletal sites. Given that modern machine learning methods such as random forests or boosting [32] can handle large feature vectors (including noisy and weak predictors), there is no need to restrict analysis to coarse single number summaries of texture distributions such as box-counting dimension. Chappard *et al*. [33] used a smaller set of three co-occurrence features at a single separation distance combined with geometrical features in linear regression models to

predict the failure load of excised femurs. It is also possible to use random forests for generalized regression as well as classification, and this may allow a larger feature set and nonlinear interactions to be explored without overtraining.

V. Conclusion

We have shown that image texture measured at the mandibular cortex have a strong association with osteoporosis diagnosed at the femoral neck, and a moderate association with osteoporosis at other skeletal sites, and is therefore a potential biomarker for osteoporosis. Texture classifiers based on co-occurrence statistics perform much better than those based on fractal dimensions that have been investigated previously, and can be more finely and objectively tuned than coarser grained human expert assignment of MCI grades. The cortical texture classifier performs similarly to cortical width as a biomarker for osteoporosis at the femoral neck, but there is a weaker link to osteoporosis in the lumbar spine, for which cortical width remains the best single predictor. The combined classifier using cortical texture and width results in a significantly stronger association with osteoporosis at the femoral neck than width-only methods, but at other skeletal sites there is little if any improvement. A smaller but significant improvement in AUC is obtained when diagnosing osteoporosis at any of the three sites by combining cortical width and similar texture features of the superior basal bone above the cortex. The cortical texture is easier to compute in a practical system, as it can be automatically calculated once the cortical borders have been determined, and the resulting improved association with femoral neck osteoporosis is particularly important, because hip fracture is one of the most serious consequences of osteoporosis.

Appendix
Bootstrapping the ROC Curve

A simple bootstrap method to compare two ROC curves would be to randomly select with replacement n_1 instances from the population of n_1 osteoporotic cases, and n_2 instances from the population of n_2 normal cases many times, use the same bootstrapped subsample for each classifier, and calculate the two AUCs, and then compute the difference. By repeating the bootstrapping many times (in this study, we have used 5000 bootstrap samples), we can estimate the distribution of the difference, and, for example, if the fifth percentile of the bootstrapped distribution exceeds zero, then the difference is significant at the 5% level. A better estimate is obtained by the smoothed bootstrap [29], in which the samples are drawn from kernel-smoothed estimates of the cumulative distributions [34], rather than sampling from the empirical sample. Equivalently, one can sample from the empirical distribution and then add small random errors to each data point drawn from the same kernel [29]. Kernel-smoothed estimators of the ROC curve have been shown to give better estimates than the simple sample ROC curve [34]. In this approach, each sample data point is in effect blurred out by a kernel function over some bandwidth interval. We follow Zhou and Harezlak [35] in using an Epanechnikov kernel [36] as the sampling distribution. Bowman *et al*. [37] showed that for an Epanechnikov kernel on [−h, h], the optimal kernel bandwidth is given by [2]

$$\hat{h} = \left(\frac{100\sqrt{\pi}}{7n}\right)^{1/3} \sigma. \tag{3}$$

This expression uses a Gaussian approximation of the decision variable distribution (see [37] for details), where n is the sample size, and σ is an estimate of standard deviation which we set according to

$$\sigma = \min\left\{\hat{\sigma}, (Q_{75} - Q_{25})/1.349\right\}. \tag{4}$$

Here, $\hat{\sigma}$ is the sample standard deviation, and $Q_{\cdot}$ represents the subscripted percentile points. We follow Zhou and Harezlak [35] in applying a log transformation prior to kernel smoothing, as the classifier probability outputs are right-skewed, and the log-transformed data tend to reduce oversmoothing caused by overestimation of constant kernel bandwidth due to skew (the Gaussian assumptions then cease to be valid). Note that the use of any monotonic transform (such as logarithm) has no effect on the empirical ROC curve, but can affect the kernel smoothing. Better estimates should be obtained if the transform aligns the distribution more closely to a Gaussian, for which the kernel bandwidth formula is optimal.

Given sample sizes n_1 for the osteoporotics and n_2 for the nonosteoporotics, we first perform bootstrap sampling with replacement picking subsamples of size n_1 and n_2 from the two populations. The probability of osteoporosis is taken from the full sample OOB estimates for both the cortical width and combined random forest classifiers. Then, Epanechnikov random noise is added using bandwidths computed as defined previously. The empirical AUC is calculated using this sample for each classifier method, and then the difference d_A between the two AUCs is taken, and its overall distribution is estimated by performing 5000 bootstrap repeats.

It could be argued that all of the random forest classifiers should be retrained on each bootstrap sample; however, the OOB estimates have already been derived through a bootstrap process, and the random forest methodology is very effective at removing overtraining, since the final estimate is a vote from thousands of trees each trained on a different subset of the sample. Also, the addition of the further noise (kernel smoothing) in effect already covers classifier training set random variability.

References

[1] C. Miller, "Survival and ambulation following hip fractures," *J. Bone Joint Surg*., vol. 60, pp. 930–934, 1978.

[2] K. Karayianni, K. Horner, A. Mitsea, L. Berkas, M. Mastoris, R. Jacobs, C. Lindh, P. F. van der Stelt, E. Harrison, J. E. Adams, S. Pavitt, and H. Devlin, "Accuracy in osteoporosis diagnosis of a combination of mandibular cortical width measurement on dental panoramic radiographs and a clinical risk index (OSIRIS): The OSTEODENT project," *Bone*, vol. 40, pp. 223–229, 2007.

[3] H. Devlin, P. D. Allen, J. Graham, R. Jacobs, K. Karayianni, C. Lindh, P. F. van der Stelt, E. Harrison, J. E. Adams, S. Pavitt, and K. Horner, "Automated osteoporosis risk assessment by dentists: A new pathway to diagnosis," *Bone*, vol. 40, pp. 835–842, 2007.

[4] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models—their training and application," *Comput. Vis. Imag. Understand*., vol. 61, pp. 38–59, 1995.

[5] M. G. Roberts, J. Yuan, J. Graham, R. Jacobs, and H. Devlin, "Changes in mandibular cortical width measurements with age in men and women," *Osteoporos. Int.*, vol. 22, pp. 1915–1925, 2011.
[6] S. Sakoda, R. Kawamata, T. Kaneda, and I. Kashima, "Application of the digital radiographic bone trabecular structure analysis to the mandible using morphological filter," *Oral Sci. Int.*, vol. 1, pp. 45–53, 2004.
[7] W. G. Geraets and P. F. van der Stelt, "Fractal properties of bone," *Dentomaxillofacial Radiol.*, vol. 29, pp. 144–153, 2000.
[8] G. Jonasson, L. Jonassan, and S. Kiliaridis, "Changes in the radiographic characteristics of the mandibular alveolar process in dentate women with varying bone mineral density: A 5-year prospective study," *Bone*, vol. 38, pp. 714–721, 2006.
[9] F. Yasar and F. Akgunlu, "Evaluating mandibular cortical index quantitatively," *Eur. J. Dentistry*, vol. 2, pp. 283–290, 2008.
[10] E. Lerouxel, H. Libouban, M. F. Moreau, M. F. Basle, M. Audran, and D. Chappard, "Mandibular bone loss in an animal model of male osteoporosis (orchidectomized rat): A radiographic and densitometric study," *Osteoporos. Int.*, vol. 15, pp. 814–819, 2004.
[11] T. E. Southard, K. A Southard, K. E. Krizan, S. L. Hills, J. W. Haller, J. Keller, and M. W. Vannier, "Mandibular bone density and fractal dimension in rabbits with induced osteoporosis," *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod.*, vol. 89, pp. 244–249, 2000.
[12] P. D. Allen, J. Graham, D. J. J. Farnell, E. J. Harrison, R. Jacobs, K. Nicopolou-Karayianni, C. Lindh, P. F. van der Stelt, K. Horner, and H. Devlin, "Detecting reduced bone mineral density from dental radiographs using statistical shape models," *IEEE Trans. Inf. Technol. Biomed.*, vol. 6, no. 6, pp. 601–610, Dec. 2007.
[13] M. G. Roberts, J. Graham, and H. Devlin, "Improving the detection of osteoporosis from dental radiographs using active appearance models," in *Proc. IEEE Int. Symp. Biomed. Imag.*, Rotterdam, The Netherlands, Apr. 14–17, 2010, pp. 440–443.
[14] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.
[15] E. Klemetti, S. Kolmakov, and H. Kroger, "Pantomography in assessment of the osteoporosis risk group," *Scand. J. Dent. Res.*, vol. 102, pp. 68–72, 1994.
[16] A. Taguchi, M. Tsuda, M. Ohtsuka, I. Kodama, M. Sanada, T. Nakamoto, K. Inagaki, T. Noguchi, Y. Kudo, Y. Suei, K. Tanimoto, and A.-M. Bollen, "Use of dental panoramic radiographs in identifying younger postmenopausal women with osteoporosis," *Osteoporos. Int.*, vol. 17, pp. 387–394, 2006.
[17] A. Taguchi, A. Asano, M. Ohtsuka, T. Nakamoto, Y. Suei, M. Tsuda, Y. Kudo, K. Inayaki, T. Noguchi, K. Tanimoto, R. Jacobs, E. Klemetti, S. C. White, and K. Horner, "Observer performance in diagnosing osteoporosis by dental panoramic radiographs: Results from the osteoporosis screening project in dentistry (OSPD)," *Bone*, vol. 43, pp. 209–213, 2008.
[18] S. Geman and D. McClure, "Statistical methods for tomographic image reconstruction," *Bull. Int. Statist. Inst.*, vol. LII, pp. 4–5, 1997.
[19] P. J. Rousseeuw and C. Croux, "Alternatives to the median absolute deviation," *J. Amer. Statist. Assoc.*, vol. 88, pp. 1273–1283, 1993.
[20] E. Inoue and K. Ogawa, "Analysis of trabecular patterns using fractal dimensions," in *Proc. Nucl. Sci. Symp. Med. Imag. Conf. Rec.*, 1995, vol. 3, pp. 1497–1500.
[21] C. J. Rose, S. J. Mills, J. P. O'Connor, G. A. Buonaccorsi, C. Roberts, Y. Watson, B. Whitcher, G. Jayson, A. Jackson, and G. J. Parker, "Quantifying heterogeneity in dynamic contrast-enhanced MRI parameter maps," in *Proc. Med. Image Comput. Comput.-Assist. Intervention Conf.*, Berlin, Germany, 2007, vol. 2, pp. 376–384.
[22] C. J. Rose, S. J. Mills, J. P. O'Connor, G. A. Buonaccorsi, C. Roberts, Y. Watson, S. Cheung, S. Zhao, B. Whitcher, A. Jackson, and G. J. Parker, "Quantifying spatial heterogeneity in dynamic contrast-enhanced MRI parameter maps," *Magn. Reson. Med.*, vol. 62, pp. 488–499, 2009.
[23] H.-O. Peitgen, H. Jurgens, and D. Saupe, *Chaos and Fractals*. Heidelberg, Germany: Springer, 2004.
[24] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man Cybern.*, vol. 3, no. 6, pp. 610–621, Nov. 1973.
[25] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.
[26] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. New York, NY, USA: Chapman & Hall, 1984.
[27] M. H. Zweig and G. Campbell, "Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine," *Clin. Chem.*, vol. 39, pp. 561–577, 1993.
[28] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, pp. 153–157, 1947.
[29] P. Bertail, S. Clemencon, and N. Vayatis, "On bootstrapping the ROC curve," in *Proc. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2008, pp. 137–144.
[30] T. E. Southard, K. A. Southard, and A. Lee, "Alveolar process, fractal dimension and postcranial bone density," *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod.*, vol. 91, pp. 486–491, 2001.
[31] M. Petrou and P. G. Sevilla, *Image Processing: Dealing With Texture*. Chichester, U.K.: Wiley, 2006.
[32] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proc. 13th Int. Conf. Mach. Learn.*, San Francisco, CA, USA, 1996, pp. 148–156.
[33] C. Chappard, V. Bousson, C. Bergot, D. Motton, A. Marchadier, T. Moser, C. L. Benhamou, and J-D. Laredo, "Prediction of femoral fracture load: Cross-sectional comparison of texture analysis and geometric measurements on plain radiographs versus bone mineral density," *Radiology*, vol. 255, pp. 536–543, 2010.
[34] C. J. Lloyd and Z. Yong, "Kernel estimators of the ROC curve are better than empirical," *Statist. Probab. Lett.*, vol. 44, pp. 221–228, 1999.
[35] X. Zhou and J. Harezlak, "Comparison of bandwidth selection methods for kernel smoothing of ROC curves," *Statist. Med.*, vol. 21, pp. 2045–2055, 2002.
[36] M. P. Wand and M. C. Jones, *Kernel Smoothing*. New York, NY, USA: Chapman & Hall, 1995.
[37] A. Bowman, P. Hall, and T. Trvan, "Bandwidth selection for the smoothing of distribution functions," *Biometrika*, vol. 85, pp. 799–808, 1998.

Authors' photographs and biographies not available at the time of publication.

44. **Multi-scale rigid registration to detect damage in micro-CT images of progressively loaded bones.** R. Green, J. Graham and H. Devlin, *Proceedings of the IEEE International Symposium on Biomedical Imaging Chicago IL. April 2011. pp 1231-1234.* doi: 10.1109/ISBI.2011.5872624

MULTI-SCALE RIGID REGISTRATION TO DETECT DAMAGE IN MICRO-CT IMAGES OF PROGRESSIVELY LOADED BONES

Richard Green[⋆†] *Jim Graham*[†] *Hugh Devlin*[⋆]

[†] Imaging Science and Biomedical Engineering, School of Cancer and Enabling Sciences
[⋆] School of Dentistry
Manchester Academic Health Science Centre, University of Manchester, UK

ABSTRACT

We present a method to detect damage in time-lapsed, micro-CT images of progressively loaded bone. The method we have developed splits the image into regions and performs registration on each region individually. The procedure is repeated with progressively smaller regions until either a minimum size or a maximum number of levels is reached. The regions are then classified as damaged or undamaged. This method has been successfully applied to three sets of images and tested with simulated damage. It will allow us to determine the characteristics of trabecular architecture that provide strength and to predict regions that are most likely to be damaged based on their structure.

Index Terms— Bone Strength, Micro-CT Imaging, Image Registration, Trabecular Structure

1. INTRODUCTION

Osteoporosis is a disease characterised by low bone mass and a deterioration in bone quality. Common clinical assessments are based on measures of the quantity of bone material (e.g. bone mineral density (BMD)). The architecture of trabecular bone, despite being an important component of bone strength, is not generally measured, partly due to the lack of availability of appropriate high resolution 3D imaging, but also the lack of a measure of bone quality to apply.

Appropriate imaging techniques such as dental cone beam computed tomography (CBCT) and high resolution peripheral quantitative computed tomography (HR-pQCT) are beginning to become available. Measures of bone quality could for example be used to improve the prediction of osteoporosis or the siting of dental implants.

The term 'bone quality' is not well defined. It covers the range of properties that provide strength and resistance to fracture including the quantity and density of bone, turnover and trabecular architecture. It is the last of these with which we are concerned.

Micro-CT imaging allows 3D trabecular structure to be imaged directly at high resolution. It has been widely adopted for both in-vivo and in-vitro assessment of trabecular architecture in animal models of osteoporosis. The most commonly used measures, such as trabecular thickness, structure model index (a measure of 'rod- or plate-likeness'), anisotropy and connectivity, show characteristic changes with osteoporosis and correlations to bone strength [1, 2]. These measures are mostly extensions of histomorphometric measures and not based on a knowledge of the way in which bone breaks.

We are interested in understanding the way in which bone fractures under loading in order to devise a measure of bone quality more clearly linked to strength and fracture resistance. To do this we are acquiring micro-CT images of progressively loaded rat vertebrae. From these images we will identify regions where damage occurs and look for features and structural characteristics that can predict them.

Nazarian and Müller were first to suggest the use of micro-CT based time-lapsed imaging to observe bone failure [3, 4]. They call the method 'image guided failure analysis' (IGFA). Their method involves in-situ, step-wise, compression with a strain held during scanning. Other authors have imaged damage, in these cases using synchrotron-based micro-CT, with a time lapsed method but again have not detected or measured damage from their images [5, 6].

Larrue et al. examined the automatic detection of micro-cracks from synchrotron-based micro-CT [7]. Their work differs from our own as they look specifically for micro-cracks (typically 1μm in width), which is only possible with synchrotron radiation, whereas we are seeking to characterise regions of weakness at a larger scale.

Perilli et. al. and Tassani et. al. both compressed cylindrical sections of human femoral head until a fracture plane formed through the entire specimen, producing two separate halves [8, 9]. In the former study a fracture region was defined through the entire cross-section of the specimen for the vertical extent of the fracture plane. Standard measures of trabecular architecture were shown to be statistically different in this region compared to that in the rest of the specimen. In the latter study the fracture region was determined automatically by registering each half of the the broken specimen to the undamaged original. This method proved to be as effective at delineating the damaged region as the manual approach.

In our study we seek to identify damage as it occurs at a scale intermediate between microcracks and complete breakage, prior to total failure of the bone. We require an automatic method with sufficient sensitivity to achieve this, in the context that such local deformation can be visually difficult to identify and distinguish from movement in the 3D volume. In particular we seek to differentiate between changes that are due to damage and those that are due to movement of undamaged bone. We register sections of loaded bone to the image prior to loading. To achieve detection at smaller scales we register increasingly small regions before classifying each region as damaged or undamaged. The method has been applied to images with artificially simulated damage to investigate the scale of detection that can be achieved, and to several sets of real data.

2. METHODS

2.1. Imaging and Loading

Individual lumbar vertebrae (L3, L4 and L5) were collected from 2 female rats, approximately 3 months old. Excess, external soft

Facilities for collecting in-situ loading data were kindly made available by Skyscan. This work was in part funded by the BBSRC.

tissue was removed. Micro-CT imaging is described below. Reconstruction was performed in the manufacturer's software, NRecon or InstaRecon (Skyscan, Kontich, Belgium). All images were resized by half. Where segmented images are used a global threshold was applied that was determined using Otsu's method.

Two different methods of preparation, loading and imaging have been used. The first involves separate compression and imaging steps where the bone is able to relax after the loading. The second method applies the load inside the scanner, the strain for each compression being maintained throughout the scans. The first method requires repositioning of the bone within the scanner between scans. Three sets of data have been analysed, the first two using the first method (1a, 1b) , the third using the second method (2a).

Specimens for the first method were prepared for loading by removing the parts of any processes that extended beyond the main body of the vertebra using a scalpel. Imaging was performed in a Skyscan 1072 scanner. Reconstructed images were $1024\times1024\times975$ voxels in size with a voxel size of 13.7 µm. Loading was performed in a Zwick universal testing machine. An increasing compressive strain was applied at a rate of 2 mm min^{-1} until a drop in force of 10% from the maximum occurred (typically indicating fracture). After the first scan of the intact bone the specimen was removed from the scanner for each compression step. Two vertebrae were loaded and imaged using this method. The first (1a) has 4 stages of loading, where the final loading was stopped at a strain of 20% as a drop of 10% did not occur. The second (1b) had 3 steps where the final loading was stopped at a force of 1000 N for similar reasons.

The specimen for the second method (2a) was prepared for loading by attaching PMMA caps to the each end of the vertebra. This has the advantage of providing flat, parallel surfaces for loading and the method has been devised to ensure repeatable, on-axis loading scenarios for future experiments. Imaging was performed using a Skyscan 1172 scanner and using Skyscan's in-situ loading stage for loading, minimising position differences between scans. A given strain was held during each scan with 2 loading steps, each increasing the strain by approximately 2.5%. Reconstructed images were $1500\times1500\times964$ voxels in size with a voxel size of 10.1 µm.

2.2. Simulated Damage

Artificial data was generated from two images of the same bone that were taken on separate occasions without any loading in between. The two images were globally registered and 7, 150^3 voxel sections were selected that contained a variety of structure types.

Damage to the bone was simulated by applying 3D deformation fields to one of the images. The deformation was described by three parameters: position and size of the damage region and the magnitude of the deformation. Each voxel in the damage region was moved by the deformation magnitude equally in each direction. Before being applied to the starting image the deformation fields were each smoothed (Gaussian smoothing with a SD of 6) to prevent discontinuities.

Two types of deformation were performed on each image pair. The first had a fixed magnitude of 15 voxels and the size was increased from 5^3 to 45^3 voxels in 5 steps. The second had a fixed size of 15^3 and the magnitude was increased from 5 to 45 voxels in 5 steps. A correctly identified damaged region was defined as one that coincided with 10% or more voxels above a threshold level in the deformation field. The threshold was 10% of the starting deformation magnitude for each region. The results for each level of damage have been averaged to provide a final result.

2.3. Registration

All damaged images were first registered onto the initial, undamaged, image using a global, rigid registration. This and all further registration was performed using FLIRT [10], part of FSL. This software was originally designed for use with neuro-MR images but has been found to perform well for our purposes. It uses a multi-scale, multi-start approach and applies de-weighting to edge pixels to decrease the effect of the changing overlap of two images being registered. In tests we found it reliably finds a global minimum for our images. All registrations are rigid and use the correlation-ratio cost function.

2.4. Location of Damage

We wish to identify regions that are damaged and distinguish them from those that have moved as a result of damage. Splitting the image into smaller regions and registering each individually will allow for heterogeneous movement and localised damage (see figure 1). It is hypothesised that damaged bone will not be able to be successfully registered using rigid registration and this can be used to classify regions. This method does not make any assumptions about the size of damage which could range from a single deformed or broken trabecula to diffuse damage throughout a larger region.

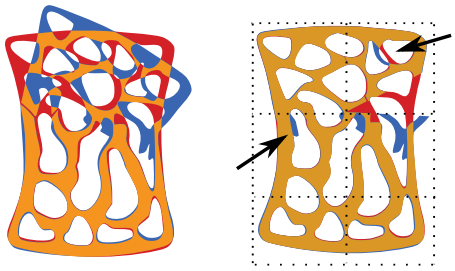


Fig. 1. Illustrating the detection of damaged regions by registration. Red indicates bone in the non-loaded image, blue in the loaded image and orange the overlap. In the left image the site of the damage is not obvious as global, rigid, registration can mask smaller, regional damage. Splitting the image up into smaller regions and registering individually in the right image reveals undamaged regions that have moved and small scale damage that was masked by the movement (arrows). At a larger scale the clear damage at the centre right will be identified.

The first stage involves registering at several scales. The loaded image is split into octants and each octant is registered individually onto the reference (non-loaded) image. This procedure is then repeated, splitting each region into octants and registering the subregions. With each level the number of regions increases exponentially.

Regions that do not have any structure in the reference image are discarded. The amount of structure is assessed by putting a threshold on the number of bone voxels in the segmented image. At any level registration is not performed if the loaded and reference images are sufficiently similar within the region. This decision is based on the difference between the registered images. This difference is quantified as the sum of non-coincident bone voxels, normalised by the number of edge voxels in the reference image to account for differing numbers and sizes of trabeculae in different regions. Edge voxels are found by performing binary erosion on the reference image. The

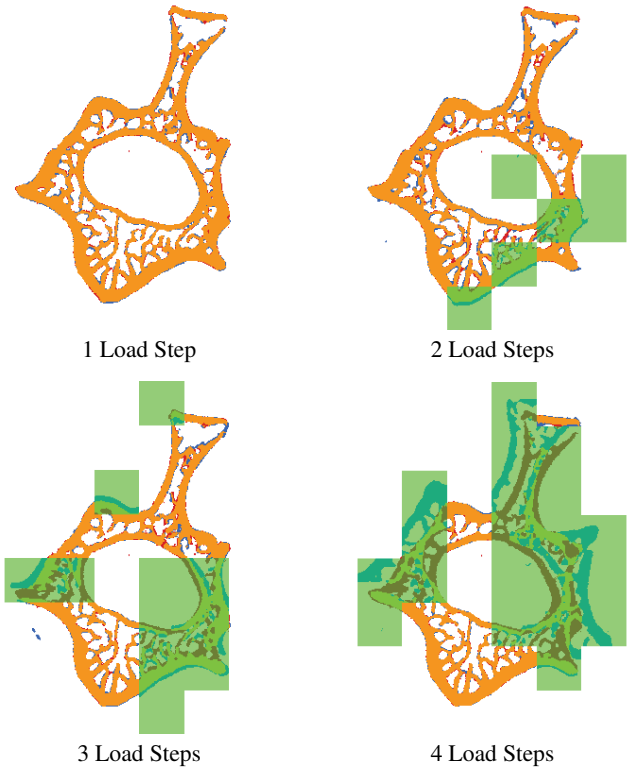


Fig. 2. A central slice perpendicular to the loading direction of non-loaded and loaded images of specimen 1a after multi-scale registration. The non-loaded image is shown in red, the loaded image in blue and the overlap in orange. The green overlay indicates regions that have been identified as damaged. As these are single slices from a volume, structure may be present in a region on another slice that is not visible.

process continues until either a minimum region size or maximum number of levels is reached.

Regions are then classified at the lowest splitting level on the basis of the remaining difference between the two images after registration. An undamaged region should show small differences; larger differences are indicative of damage. A region is thus classified as damaged if the difference between the reference and loaded image is above the threshold. If at any level the difference is below the threshold the region is classified as undamaged. Additionally a region may be classified as damaged if there is structure in the reference image but not in the loaded image.

Figure 2 shows overlays of the loaded and reference images. The loaded images are created as a mosaic of registered patches. Hence in undamaged regions the correspondence between the loaded and reference images is high, whereas there is low correspondence in the damaged regions.

The structure threshold was set at 200 voxels for the real data and 100 voxels for the artificial damage data. The splitting process was stopped after the third level (512 regions). The results are sensitive to the selection of the threshold on the difference ratio. It should be set at a level that prevents undamaged regions being classified as damaged due to variation in the segmentation arising from noise. Noise therefore limits the smallest levels of damage this method can detect. We have used a threshold value of 1 for the first set of real data, 6 for the second and a value of 0.6 for the artificial damage data, these values were determined empirically.

3. RESULTS

Table 1 shows the results of classification for the loaded images. The final region sizes were approximately 50^3 voxels. The number of regions classed as damaged increases with increasing damage for all specimens at all loading steps except for the final loading step of specimen 1b. This is due to the fact that catastrophic damage occurred quickly in this specimen. Consequently all the regions that were not discarded were classified as damaged in step 2. The increased damage at the next step therefore could not increase the number of damaged regions. Example images in figure 2 show the number of regions identified as damaged increasing with each loading step. Figure 3 shows the effect of regional registration on typical damaged and undamaged regions.

Specimen	Step	Undamaged	Damaged
1a	1	181 (35.4%)	48 (9.4%)
	2	130 (25.4%)	99 (19.3%)
	3	113 (22.1%)	116 (22.7%)
	4	58 (11.3%)	171 (33.4%)
1b	1	72 (14.1%)	124 (24.2%)
	2	0 (0%)	196 (38.3%)
	3	0 (0%)	196 (38.3%)
2a	1	281 (54.9%)	6 (1.2%)
	2	264 (51.6%)	23 (4.5%)

Table 1. Summary of classification results for the loaded images. As damage in the image increases regions that were previously classified as undamaged become damaged. Percentages are of the total number of regions, including those discarded.

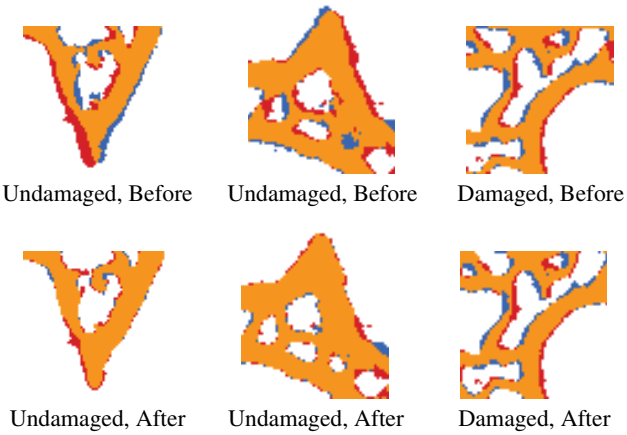


Fig. 3. Two regions classified as undamaged and one as damaged from the first loading stage of specimen 1b, before and after the fracture detection was applied. The non-loaded image is shown in red, the loaded image in blue and the overlap in orange. Local registration has improved the correspondence between the images for the undamaged regions resulting in a difference below threshold after regional registration. However, registration has been unable to reduce the difference in the damaged region (right).

Results for the detection of damaged regions in the artificially damaged images have been summarised in table 2 and example slices are shown in figure 4. Both types of damage show increasing sensitivity with increasing damage while specificity remains relatively constant.

Size /voxels	5^3	15^3	25^3	30^3	45^3
Sens. (%)	7.1	75.0	76.3	85.7	88.6
Spec. (%)	91.2	91.0	89.5	89.6	89.2
Magnitude /voxels	5	15	25	30	45
Sens. (%)	28.6	75.0	89.3	89.3	100
Spec. (%)	91.0	91.0	90.3	90.2	90.0

Table 2. Sensitivity and specificity of classification results for the artificially deformed images with increasing damage. Results are for an average of 7 regions. True positives are regions correctly identified as damaged, true negatives are those correctly identified as undamaged. The deformation field applied to create the simulated damage is used to provide the ground truth, see section 2.2.

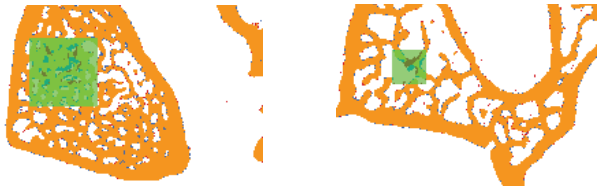


Fig. 4. Two example slices showing the artificial damage (level 2 on table 2) and the correct identification of a damaged region. The image before artificial damage has been applied is shown in red, the artificially damaged image is shown in blue and the overlap in orange. Regions classified as damaged are shown in green.

4. DISCUSSION

We have described a method based on multi-scale registration for identifying regions of damage in micro-CT images of loaded trabecular bone. The strength of this method is its ability to distinguish damage from movement where other techniques would not be able to do so. Non-rigid registration, for example, would have difficulty achieving this as it is necessary to identify large distortions at small scale while simultaneously allowing large scale rigid movements. It is not easy to see how a regularisation mechanism could deliver both of these outcomes.

Most similar studies use a cylinder of trabecular bone selected to have as uniform structure as possible. Our method allows whole vertebrae to be tested despite its heterogeneous structure.

In this study we have examples of progressive damage in two different loading protocols: repeated loading and progressive loading, extending to levels of damage greater than we require to detect. In all cases our method has identified regions of damage and distinguished these from regions of movement. Using synthesised ground-truth, we have been able to specify the minimum volume and extent of damage that can be detected using this approach.

The specific aim of this study is to propose methods for quantifying bone quality. We take as a working definition of bone quality the strength under loading. Regions of bone that suffer damage in the early stages of loading are, under this definition, weaker (of lower 'quality') than regions that remain undamaged. Our future work will seek to identify measures of local bone structure that allow us to discriminate the weaker regions from the others. Early damage is likely to occur in fairly small regions, where it may be difficult to detect visually. The methods described here are intended to identify regions of minimal damage and regions that resist damage. These detected regions will be confirmed visually before being used to train and evaluate local structure measures.

5. REFERENCES

[1] S. K. Boyd, P. Davison, R. Müller, and J. A. Gasser, "Monitoring individual morphological changes over time in ovariectomized rats by in vivo micro-computed tomography," *Bone*, vol. 39, no. 4, pp. 854–862, Oct. 2006.

[2] G. Campbell, H. Buie, and S. Boyd, "Signs of irreversible architectural changes occur early in the development of experimental osteoporosis as assessed by in vivo micro-ct," *Osteoporosis International*, vol. 19, no. 10, pp. 1409–1419, Oct. 2008.

[3] A. Nazarian and R. Müller, "Time-lapsed microstructural imaging of bone failure behavior," *Journal of Biomechanics*, vol. 37, no. 1, pp. 55–65, Jan. 2004.

[4] A. Nazarian, M. Stauber, and R. Müller, "Design and implementation of a novel mechanical testing system for cellular solids," *Journal of Biomedical Materials Research. Part B, Applied Biomaterials*, vol. 73, no. 2, pp. 400–411, May 2005, PMID: 15682380.

[5] R. Akhtar, M.R. Daymond, J.D. Almer, and P.M. Mummery, "Elastic strains in antler trabecular bone determined by synchrotron x-ray diffraction," *Acta Biomaterialia*, vol. 4, no. 6, pp. 1677–1687, Nov. 2008.

[6] P. J. Thurner, P. Wyss, R. Voide, M. Stauber, M. Stampanoni, U. Sennhauser, and R. Müller, "Time-lapsed investigation of three-dimensional failure and damage accumulation in trabecular bone using synchrotron light," *Bone*, vol. 39, no. 2, pp. 289–299, 2006.

[7] A. Larrue, A. Rattner, N. Laroche, L. Vico, and F. Peyrin, "Feasibility of Micro-Crack detection in human trabecular bone images from 3D synchrotron microtomography," in *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, 2007, pp. 3918–3921.

[8] E. Perilli, M. Baleani, C. Öhman, R. Fognani, F. Baruffaldi, and M. Viceconti, "Dependence of mechanical compressive strength on local variations in microarchitecture in cancellous bone of proximal human femur," *Journal of Biomechanics*, vol. 41, no. 2, pp. 438–446, 2008.

[9] S. Tassani, P. A. Asvestas, G. K. Matsopoulos, and F. Baruffaldi, "Automatic identification of trabecular bone fracture," in *XII Mediterranean Conference on Medical and Biological Engineering and Computing 2010*, vol. 29 of *IFMBE Proceedings*, pp. 296–299. Springer Berlin Heidelberg, 2010.

[10] M. Jenkinson, P. Bannister, M. Brady, and S. Smith, "Improved optimization for the robust and accurate linear registration and motion correction of brain images," *NeuroImage*, vol. 17, no. 2, pp. 825–841, Oct. 2002.

Applications of Image Analysis: Segmentation of the Prostate

45. **Differential segmentation of the prostate in MR images using tissue modelling and 3D Active Shape Models.** P.D. Allen, D Williamson, J. Graham, and C.E. Hutchinson. *Proceedings of the IEEE International Symposium on Biomedical Imaging Arlington VA. April 2006. J Kovacevic and E. Meijering, eds. IEEE. pp 410-413.* doi:10.1109/ISBI.2006.1624940

DIFFERENTIAL SEGMENTATION OF THE PROSTATE IN MR IMAGES USING COMBINED 3D SHAPE MODELLING AND VOXEL CLASSIFICATION

P. D. Allen, J. Graham, D. C. Williamson and C. E. Hutchinson

ISBE, University of Manchester, Oxford Road, M13 9PT, UK

ABSTRACT

Benign Prostatic Hyperplasia (BPH) is a non-cancerous expansion of the prostate, the progress of which can be quantified by measuring the relative volumes of the prostate's peripheral zone and central gland. Here we describe a method of automatic segmentation of both regions of the prostate from MR images using a combination of grey-level voxel classification and 3D statistical shape modelling.

1. INTRODUCTION

Benign Prostatic Hyperplasia (BPH) is a non-cancerous enlargement of the prostate which can cause constriction of the urethra and therefore obstruction of urinary flow. It affects 70% of men between the ages of 61 and 70, rising to 80% for men over 80 [1]. In 25% of men aged 80 symptoms are sufficiently severe to require surgical transurethral resection of the prostate (TURP), however this treatment has a high cost, morbidity (16%) and mortality (2.01%) [2] and so alternative treatments are sought.

Drugs such as finasteride can be used to treat BPH by shrinking the prostate, and evaluation of such candidate treatments requires a method of quantifying its effect. The current standard is Transrectal Ultrasound (TRUS) in which three orthogonal dimensions are measured and the volume is estimated using the formula for a prolate ellipsoid [2, 3]. Anatomically the prostate is divided into a number of zones: Peripheral (PZ), Central (CZ), Transitional (TZ), and fibromuscular. BPH primarily affects the TZ and so both the total Prostate (TP) and TZ volumes are measured using TRUS.

Tewari et al [2] have shown that the reduction in volume due to finasteride treatment over 12 months for the total TP and TZ are 8% and 27% respectively, however only the change in TZ volume correlates with improvement in urinary flow. For TRUS, intra-observer variability has been shown to be -18% to $+18\%$ for the TZ volume and -21% to $+30\%$ for the total volume [3].

Magnetic Resonance Imaging (MRI) is an attractive alternative to TRUS as it offers better definition of the prostate and is non invasive. MRI offers the possibility of accurately segmenting the prostate rather than assuming its volume from three orthogonal measurements. However, manual segmentation is time consuming, error prone and subjective, and so the goal of this project is to investigate the possibility of automatic segmentation of the appropriate regions of the prostate.

In MRI only two regions can be distinguished: the PZ, and what is referred to as the Central Gland (CG) comprising the remaining anatomical zones [4]. In cases of BPH the CG is mostly comprised of TZ due to the latter's expansion and so CG and TZ can be considered equivalent. Methods of segmentation [5] and registration [6] of just the outer prostate surface have been described for MR imaging, here we describe a method of both whole prostate and CG segmentation.

For this study we have used T2 weighted fat suppressed (T2FS) images as the CG/PZ contrast is enhanced in comparison with T2 or T1 weighting, and there is clearer separation of the prostate from surrounding tissue. The data were collected using a 1.5T Philips Gyroscan ACS MR scanner (software version NT5.3, Power Track 600, synergy body coil) from 22 patients with BPH. For each patient there are 50 axial slices with a thickness of 2mm and an in-plane resolution of 1.56mm.

Figure 1 shows a T2FS MR image of a prostate sliced in the axial, sagittal, and coronal planes. In T2 weighted images the PZ is generally brighter than the CG and in this case the two can be distinguished reasonably well.

Manual segmentation of the prostate is particularly difficult toward the superior portion where the seminal vesicles are very difficult to distinguish from the PZ, and toward the inferior portion where surrounding structures can become confused with the prostate and the prostate itself tends to bifurcate into two lobes. In the mid-section of the prostate blood vessels anterior to it can be confused with the PZ or CG depending on their relative intensity. The border between the PZ and CG can vary greatly from patient to patient depending on the severity of glandular enlargement. Figure 6 shows manual segmentation of the axial slice of the prostate, illustrating that the boundary between the regions is defined not only by the voxel values, but requires a model of expected shape.

2. AUTOMATIC SEGMENTATION

In seeking a method of automatic segmentation, our approach is to formalise the two level process behind manual segmen-

This project was funded by Pfizer.

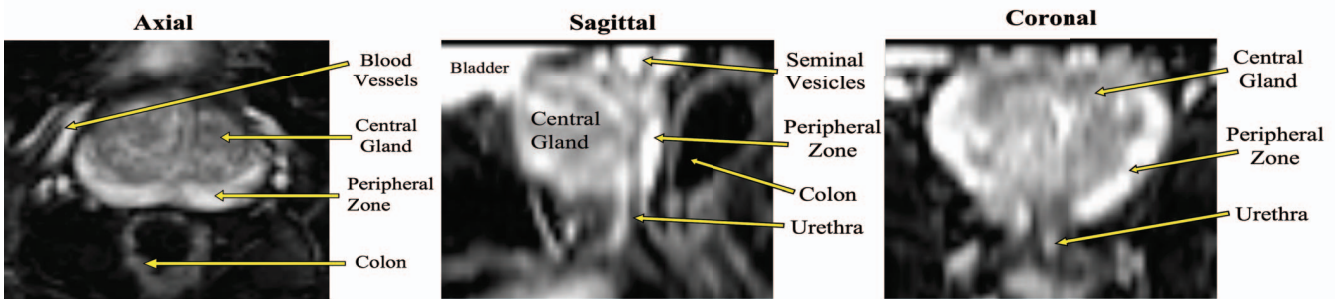


Fig. 1. *The appearance of the prostate in T2 weighted Fat suppressed MRI sliced in three orthogonal planes.*

tation using grey-level voxel classification to make the initial coarse segmentation, and to fit a 3D point distribution model (PDM) [7] to this classified data to form the smooth spatial constraint.

2.1. Voxel Classification

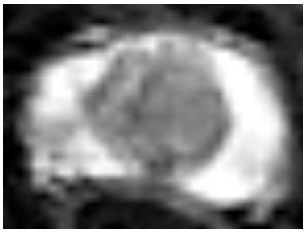


Fig. 2. *Axial T2FS MR image of a prostate.*

Fig. 3. *The result of applying grey-level tissue classification to figure 2.*

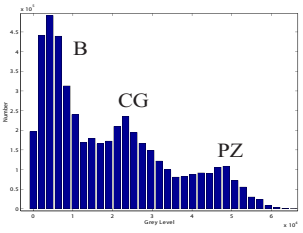


Fig. 4. *The grey-level histogram of figure 2.*

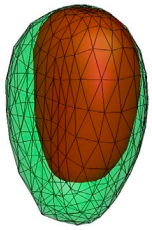


Fig. 5. *The mean shape of the double surface 3D PDM of the prostate*

Figure 2 shows an axial slice of a T2FS images cropped close to the prostate. The histogram of grey-levels in this image is shown in figure 4. There are three distinct peaks in this histogram corresponding to PZ, CG and what we can consider background (B), suggesting that we can assume the image as being composed of three tissues. In an MR image a pure tissue type would produce a distinct grey level intensity with Gaussian distributed noise, so if the three tissue assumption

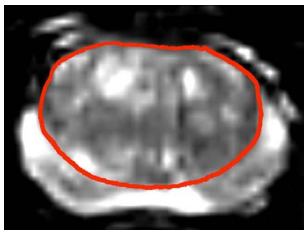


Fig. 6. *Manual segmentation of an axial slice into PZ and CG.*

Fig. 7. *The result of applying grey-level tissue classification to figure 6.*

holds we would expect the distribution of the histogram to be a sum of three Gaussians.

By fitting a three Gaussian-model to the histogram we can then calculate the class-conditional probability of each voxel in the image belonging to each of the three tissue types [8, 9]. Figure 3 shows the results of using these probabilities to classify the image into PZ, CG and B. The example in figure 2 is particularly well suited to voxel classification. However in a less homogenous case such as that shown in figure 7 bright BPH nodules will be wrongly classified as PZ and the dark compressed regions of PZ wrongly classified as CG - thus a further spatial constraint is required.

2.2. Shape Modelling

We are interested in fitting two surfaces: the Total Prostate (TP) and the Central Gland (CG) (see figure 5). The 22 images have been manually segmented to provide examples of each of these surfaces. To build a PDM from these surfaces requires a set of points on each surface which correspond across the data set and to achieve this we employ a method of automatic correspondence optimisation [10].

Using a leave-one-out evaluation of the ability of the PDM to represent the training data, it was found that using two separate PDMs for the TP and CG surfaces gave considerably better representation of the observed shape variability than using a single PDM of the two surfaces. This is because the 22 manually-segmented examples are not sufficient to ade-

quately describe the variation in the spatial relationships between the two surfaces.

2.3. Model Fitting

2.3.1. TP Surface:

From the tissue classification (section 2.1) each voxel has three values associated with it - P_{PZ}, P_{CG}, P_B representing the probabilities that it belongs to PZ, CG, or B. For an example surface we can sum these probabilities for the voxels enclosed by the surface giving the quantities which we can call PZ_{in}, CG_{in}, and B_{in}. For fitting the whole prostate surface a sensible objective function would then be:

$$PZ_{in} + CG_{in} - B_{in} \quad (1)$$

We initialise the search for fitting all of the shape parameters by first fitting the pose of the average shape. Using the objective function in equation 1, and pose paramters only, the search space is fairly smooth. We are able to use simplex to find an initial configuration for shape search. Optimisation of the surface shape however presents a far more complex search space with many local minima, and so here a genetic algorithm is used.

2.3.2. CG Surface:

To fit the CG surface it is the PZ/CG border that must be emphasised in the objective function and this can be done in the following way: Create a candidate CG surface $C1$, then dilate that surface by one voxel to form a second surface $C2$, which, for a correct surface $C1$, should be outside the CG. If we sum the probabilities *on* the surfaces rather than *in* them we can form the values PZ_{onC2} and PZ_{onC1}. The difference between these values should be a maximum when the CG surface is on the PZ/CG border. As the PZ does not always extend round to the anterior of the prostate (see figure 1), we also need to find the CG/B border in this region. We therefore also calculate B_{onC2} and B_{onC1}. From a search point of view this objective function is spiky as candidate solutions near but not at the correct position are no better than surfaces further away, and so counting the voxel probabilities inside the surface is also necessary. Thus the CG objective function becomes:

$$(PZ_{onC2} - PZ_{onC1}) + (B_{onC2} - B_{onC1}) + CG_{in} - B_{in} \quad (2)$$

In this case the sums of probabilities *in* the surface are normalised by surface volume to make them the same order of magnitude as the sums *on* the surface which are normalised by surface area. The term $' - PZ'_{in}$ is left out of the objective function because in some cases the there is considerable missclassification of CG voxels as PZ.

Table 1. Fit Results (see text).

Surface	**Point Diff (mm)** $\mu(\sigma)$	**Volume Diff (%)** $\mu(\sigma)$
TP	4.1 (1.1)	11.1 (9.5)
TP (mid)	2.8 (0.82)	6.5 (5.4)
CG	3.1 (2.5)	11.9 (8.9)
CG (mid)	2.0 (0.6)	6.8 (8.5)

Convergence of the CG fitting was not always successful starting from the mean shape. This difficulty was overcome by starting the search from a shape specified by a straightforward user interaction: the user marks four points on the middle slice roughly equally spaced around the CG, and selects the slices corresponding to the inferior and superior limits of the CG. This information is used to initialise the mean shape by stretching it in the X,Y, and Z axes and adjusting position in a simplex optimisation until the surface is as close to the user defined points as possible. The resulting surface is then used as a start point for a full GA optimisation of pose and shape.

2.4. Results

Prior to tissue classification and PDM fitting the image data was cropped manually around the prostate, as the full axial slices encompass the entire pelvic area in which the prostate is only a small region. The model fitting was tested in a series of leave-one-out experiments in which a surface model was built from the set of examples excluding the current example. GA optimisation was performed using MATLAB's genetic algorithm toolbox.

The results of fitting the TP and CG surfaces to each of the 22 patients are shown in table 1 as a mean point distance and percentage volume error. The effects of the anatomical ambiguity in the superior and inferior portions of the prostate (see section 1 can be reduced by only considering the mid-third of the cropped volume during shape and pose optimisation and the results of this are also included in table 1. Naturally the TP and CG volumes for the mid-third of the prostate are meaningless in themselves, however the CG/TP ratio calculated for this region from manual segmentation has a strong correlation (r=0.97) with the CG/TP ratio for the whole gland (figure 8). This suggests that to measure the CG/TP ratio, segmentation of the more clearly defined mid-section of the prostate may be sufficient.

2.4.1. Repeatability:

Since the GA includes a stochastic element the same fit given the same data is not guaranteed. The magnitude of this vari-

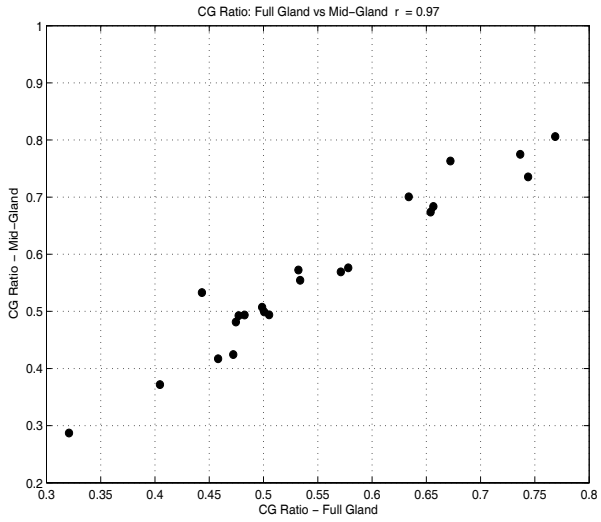


Fig. 8. *The CG/TP ratio for the mid-gland plotted against the CG/TP ratio for the whole prostate.*

ability can be estimated by repeating the fitting process 10 times and observing the variation in measured volume. On a subset of 10 of the patient group the results of this suggest standard deviations of 3% and 2% for the whole and mid-gland fits respectively.

3. DISCUSSION AND CONCLUSIONS

Table 1 demonstrates that in the majority of cases automatic segmentation results in Total Prostate and Central Gland surfaces that correspond accurately to the manual segmentation 'ground truth'. The key measure in this case is volume and even in cases where there is the greatest difference between automatic and manual segmentation these are comparable with the variation in volume estimates using TRUS. Automatic segmentation from MR images clearly has the potential to deliver precise estimates of volume change. Much of the discrepancy between manual and automatic segmentation arises in places where the boundary location is genuinely unclear, and the absolute nature of the ground truth is questionable. We intend to investigate the variability in manual measurement in these regions.

One of the more important measures to be derived is the ratio of volume of the Central Gland to Total Prostate. We have observed that this can be estimated reliably by only segmenting the more clearly defined central portion of the prostate. In a practical situation this may provide a solution to the more difficult cases.

4. REFERENCES

[1] A.M. Alam, K. Sugimura, H. Okizuka, J. Ishida, M. Igawa.: Comparison of MR Imaging and Urodynamic Findings in Benign Prostatic Hyperplasia. Radiation Medicine **18** (2000) 123–128

[2] A. Tewari, K.Shinohara, P. Narayan: Transitional Zone Volume and Transitional Zone Ratio: Predictor of Uroflow Response to Finasteride Therapy in Benign Prostatic Hyperplasia Patients. Urology **45(2)** (1995) 258–265

[3] A. R. Zlotta, B. Djavan, M. Damoun et al.: The Importance of Measuring the Prostatic Transition Zone: An Anatomical and Radiological Study. BJU International **84** (1999), 661–666

[4] A. Maio, M. D. Rifkin.: Magnetic Resonance Imaging of Prostate Cancer: Update. Topics in Magnetic Resonance Imaging **7(1)** (1995) 54–68

[5] Y. Zhu, S. Williams, R. Zwiggelaar.: Segmentation of Volumetric Prostate MRI Data Using Hybrid 2D+3D Shape Modelling. Medical Image Understanding and Analysis (2004) 61–64

[6] B. W. Fei, A. Wheaton, Z. H. Lee et al.: Automatic MR Volume Registration and its Evaluation for the Pelvis and Prostate. Phy. Med. Biol **47(5)** (2002) 832–838

[7] A. Hill, A. Thornham, C. J. Taylor.: Model-Based Interpretation of 3D Medical Images. British Machine Vision Conference (1993) 339–348

[8] D. C. Williamson, N. A. Thacker, S. R. Williams, M. Pokric.: Partial volume tissue segmentation using grey-level gradient. Medical Image Understanding and Analysis (2002) 17–20

[9] K. W. Fleischer, D. H. Laidlaw, A. H. Barr.: Partial-Volume Bayesian classification of material mixtures in MR volume data using voxel histograms. IEEE Transactions on Medical Imaging **17(1)** (1998), 74–86

[10] R. H. Davies, C. J. Twining, P. D. Allen et al.: Shape Discrimination in the Hippocampus Using an MDL Model. Information Processing in Medical Imaging (2003), 38–50

46. **Automatic differential segmentation of the prostate in 3-D MRI using random forest classification and graph-cuts optimisation.** E. Moschidis and J. Graham, *Proceedings of the IEEE International Symposium on Biomedical Imaging Barcelona, Spain. April 2012. pp 1727 – 1730.*
doi: 10.1109/ISBI.2012.6235913

AUTOMATIC DIFFERENTIAL SEGMENTATION OF THE PROSTATE IN 3-D MRI USING RANDOM FOREST CLASSIFICATION AND GRAPH-CUTS OPTIMIZATION

Emmanouil Moschidis, Jim Graham

Imaging Science and Biomedical Engineering, School of Cancer and Enabling Sciences, Manchester Academic Health Science Centre, The University of Manchester, Oxford Road, Manchester M13 9PT.
Emmanouil.Moschidis@postgrad.manchester.ac.uk, Jim.Graham@manchester.ac.uk

ABSTRACT

In this paper we address the problem of automated differential segmentation of the prostate in three dimensional (3-D) magnetic resonance images (MRI) of patients with benign prostatic hyperplasia (BPH). We suggest a framework that consists of two stages: in the first stage, a Random Forest classifier localizes the anatomy of interest. In the second stage, Graph-Cuts (GC) optimization is utilized for obtaining the final delineation. GC optimization regularizes the hypotheses produced by the classification scheme by imposing contextual constraints via a Markov Random Field model. Our method obtains comparable or better results in a fully automated fashion compared with a previous semi-automatic technique [6]. It also performs well, when small training sets are used. This is particularly useful in on-line interactive segmentation systems, where prior knowledge is limited, or in automated approaches that generate ground truth used for model-building.

Index Terms — Automatic Segmentation, Prostate Zones, MRI, Random Forests, Graph-Cuts.

1. INTRODUCTION

Biomedical image segmentation is a useful tool for clinical diagnosis and treatment planning. However, it is also a labor intensive task for radiologists, who segment a large number of images as part of their clinical routine. This is due to the fact that this task is often performed interactively, but also in some cases manually, thus involving a high cognitive load and time consumption for the human expert. Therefore, the automation of this process whenever this is possible is highly desired. In this paper we address the problem of automating the differential segmentation of the prostate in three dimensional (3-D) magnetic resonance images (MRI). The study was done in the context of identifying Benign Prostatic Hyperplasia - a non cancerous enlargement of the prostate that affects 50% of men over 60 years old [1].

The prostate is anatomically divided into the Peripheral (PZ), Central (CZ), Transitional (TZ) and Fibromuscular (FZ) zones. In BPH the prostatic enlargement is mainly due to the volumetric increase in the TZ. Therefore the estimation of the TZ volume and the TZ ratio (TZ volume/total prostate volume) is important for monitoring the progress of the disease and the effectiveness of drug treatments [2]. In MRI two regions are identified: the PZ and the Central Gland (CG), which includes the other three anatomical zones (fig. 1). However, in BPH the TZ is the predominant zone in the CG due to its expansion and therefore TZ and CG can be considered as equivalent [6]. Differential segmentation of the prostate is challenging. The appearance of the central and peripheral glands varies significantly among individuals. Furthermore surrounding tissue (seminal vesicles, blood vessels, the urethra and the bladder) present contrast challenges at different locations.

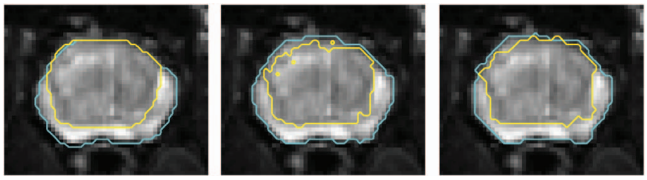


Fig. 1. An example axial mid-section of a prostate, depicting the ground truth (left), the RF hypothesis (middle) and the final GC regularized result (right). The yellow and cyan contours delineate the central gland and the total prostate respectively.

While some studies have addressed the problem of total prostate (ToP) segmentation in MRI [3, 4, 5], to the best of our knowledge it is only the study of Allen *et al.* [6] that addresses the problem of differential prostate segmentation. In their method they combine gray-level voxel classification and a double surface 3-D point distribution model. During classification, three Gaussian intensity models, which correspond to background, PZ and CG, are fit to the data. Subsequently, the 3-D point distribution model regularizes the classification output by providing smooth spatial constraints. The model fitting in this method is performed semi-automatically.

In this paper we present a fully automated method for the differential segmentation of the prostate. Our method operates in two stages: in the first stage a Random Forest (RF) classification scheme is used to localize the anatomy of interest [8]. In the second stage the final delineation is achieved via Graph-Cuts (GC) optimization. The optimization stage regularizes the hypotheses produced by the classification scheme by imposing contextual constraints via a Markov Random Field (MRF) model. In the context of comparing our method with the one of Allen *et al.* [6], we obtain comparable or better results in a fully automated fashion. Our method also performs well, when small training sets are utilized. This is particularly useful in on-line interactive segmentation systems, where prior knowledge is limited but may assist in reducing the user's cognitive load if it is utilized, or in automated approaches that generate ground truth for model-building algorithms.

2. METHODS

2.1. Dataset

We use the same dataset as in the study of Allen *et al.* [6], which consists of 22 3-D T2 fat suppressed MR images of the prostate from individuals with BPH. T2 fat suppressed MRI provides good contrast not only between the prostate and its surrounding tissue, but also between the prostatic anatomical zones. The images were acquired using a 1.5T Phillips Gyroscan ACS MR scanner. After their acquisition, all images were cropped close to the prostate as shown in fig. 1. The ground truth for each image is a binary volumetric mask produced after averaging the manual delineation of two radiologists on the cropped images. Prior to the experiments, the intensities of all images were normalized to lie in the bounded interval [0, 255]. Lastly, all images were resampled to allow for an iso-voxel resolution and volumes of equal sizes to be created.

2.2. Random Forests

RF classification constitutes the first stage of our framework, which aims for the localization of the anatomy of interest, given a set of segmented examples (training set). RF is a machine learning algorithm, which learns the properties of the data in the training set and makes predictions over unseen data using a linear combination of multiple base learners [8]. In classification, these learners are classifiers. In binary RF classification, the classifiers are *decision trees*, which are combined via a *majority voting* scheme. In such a scheme, the voxel in question is assigned the label of the class with more than 50% of the total votes. The winning votes can be seen as likelihood estimates.

Decision trees operate via test functions, similar to if-then rules, on observations of different image properties, termed *features*. In RF, each tree is trained on a *bootstrap* sample of the training set, using a random subset of the available features. A bootstrap sample is created from a given sample via random sampling with replacement. Combining the decision trees results in a classifier that yields outcomes with state-of-the-art accuracy. RF is fast during both training and prediction, it is inherently parallel and makes no assumption about the distribution of the data in the feature space. In this study we used the RF implementation available from [9].

In order to construct a feature space, which can encode the complex appearance of the anatomy of interest, we use voxel descriptors consisting of the gray level intensities, the x, y, z spatial coordinates, the Haralick [11] and the Laws [12] features. The last two sets of features are known to provide energy measures that are useful for the description of textured images. A similar approach has been adopted in [10], where textural measures are utilized to build the feature space of complex 2-D natural images, as part of an interactive segmentation system based on RF learning.

2.3. Graph-Cuts

GC optimization constitutes the second stage of our framework, which aims for the regularization of the RF hypothesis about the anatomy of interest and the provision of its final delineation (fig.1). This is achieved through the imposition of a MRF model, which incorporates contextual constraints in terms of neighboring voxel interactions. The maximum posterior probability (MAP) estimate of a MRF configuration of binary images can be computed by max-flow/min-cut algorithms [13]. The energy function that is usually minimized in segmentation problems is formulated as:

$$E(f)=\lambda\sum_{p\in P}D_p(f_p)+\sum_{\{p,q\}\in N}V_{pq}(f_p,f_q) \quad (1)$$

The first term in eq. 1 is called the *data* or *regional* term and incorporates the hypotheses that a voxel *p is* part of the foreground or the background set (having its label f_p assigned as foreground or background). In this study, these hypotheses are provided by the RF classification stage. The second term is the *smoothness* or *boundary* term, which encompasses the boundary constraints of the MRF model. It is also known as the interaction potential, since it handles the pairwise interactions between two voxels *p* and *q* of a neighborhood *N*. This term acts as the regularizer of the regional hypotheses and is written, following the notation of [14], as:

$$V_{pq}(f_p,f_q)=w_{pq}\cdot\delta(f_p\neq f_q) \quad (2)$$

$$\text{with } w_{pq}=\exp\frac{-(I_p-I_q)^2}{2\sigma^2} \quad (3)$$

where δ is a Kronecker delta function and I_p and I_q the intensities of voxels *p* and *q* respectively. The tunable parameters λ and σ, control the relative importance of the two terms in eq. 1 and the full width at half maximum of the Gaussian function's peak in eq. 3 respectively. In the context of this study we used our own implementation [7] of the GC algorithm described in [15].

2.4. Evaluation Metrics

We assess the performance of our method with a score of classification accuracy (*CA*), the Tanimoto coefficient (*Tc*), the volumetric difference (*VolDiff*) between the ground truth and the segmentation volumes and the maximum and mean point to surface distance between the segmentation and the ground truth surface (*MaxDist* and *MeanDist* respectively).

In order to calculate the *CA* metric, all the voxels are classified into true and false positives (*TP*, *FP*) and true and false negatives (*TN*, *FN*). The *CA* score is then defined as:

$$CA=100\times\frac{|TP|+|TN|}{|TP|+|TN|+|FP|+|FN|}\ \% \quad (4)$$

The Tanimoto coefficient *Tc* is computed as:

$$Tc=100\times\frac{|TP|}{|TP|+|FP|+|FN|}\ \% \quad (5)$$

The *VolDiff* metric is defined as:

$$VolDiff=100\times\frac{||FN|-|FP||}{|TP|+|FN|}\ \% \quad (6)$$

The *MaxDist* and *MeanDist* metrics are calculated via a 3-D distance transform. The distance is given in voxels, but since we use images with iso-tropic voxels, this result can be reported in millimeters as well.

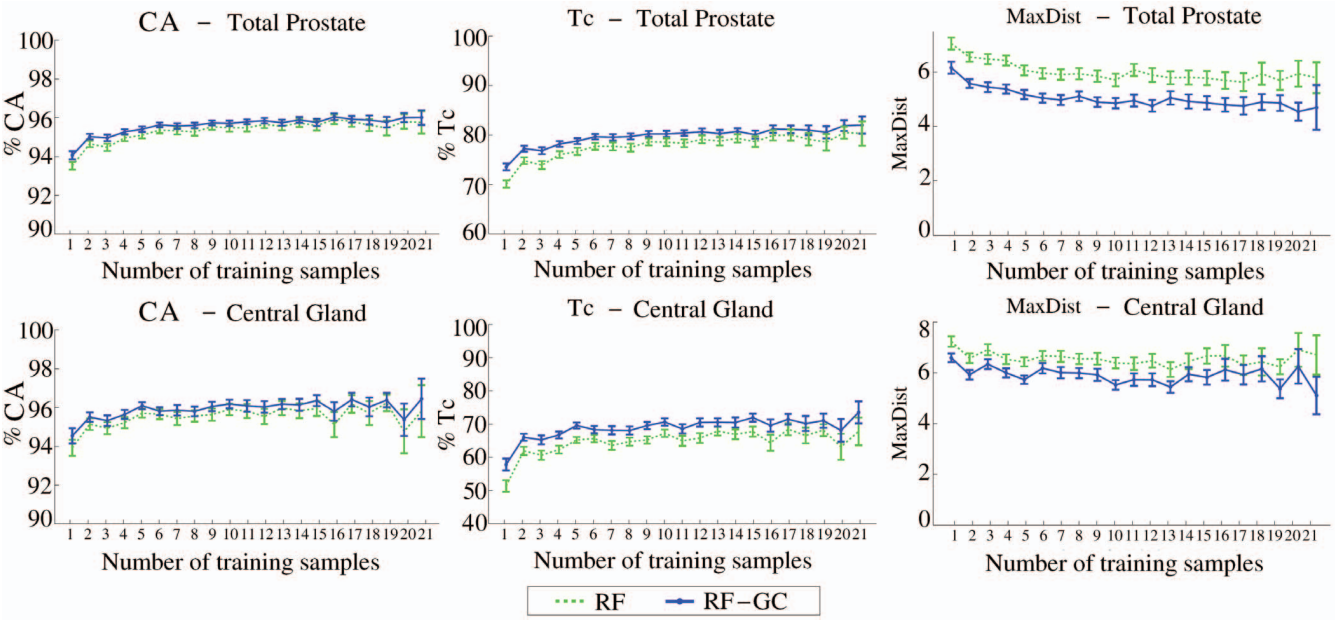


Fig. 2. Graphs depicting the performance of our framework before (RF) and after (RF-GC) regularization with respect to the *CA* (left), *Tc* (middle) and *MaxDist* (right) accuracy metrics for the Total Prostate (top row) and the Central Gland (bottom row) segmentation tasks. The error bars represent the ± 1.96 × standard error of the mean.

The *VolDiff* and *MeanDist* metrics are used to allow for the direct comparison between our results and those reported in [6]. All metrics quantify different aspects of the accuracy of segmentation.

3. EXPERIMENTS AND RESULTS

In order to observe the dependency of our method's accuracy on the number of the training samples, we created 21 training sets with sizes that ranged from 1 to 21. The images in each set were selected randomly. Those that did not take part in the training set, were used to form the test set. The process was repeated 10 times to allow for the production of training sets with considerable variability, leading to a total of 210 training sets, 10 from each size.

Having training sets with 3-D images leads to a large number of voxels that need to be stored in memory during the training phase of RF. This issue intensifies when a large number of features is also utilized. In order to improve the computational efficiency of the method, we performed the following actions: firstly, we sampled a fixed number of voxels from each training set, instead of using all the available voxels. The same number of voxels was sampled from each image of the training set, using a proportionate stratified sampling strategy. Secondly, we performed feature selection before training the final RF classifier, which led to the reduction of the number of features used from 753 to 16.

During RF training about a third of the training data are separated from the original training set and they are used by the algorithm as a validation test. In this stage an error estimate and the importance of the features in the prediction accuracy are provided. We used the *Gini impurity criterion* [8] for selecting the most important features, which assesses how well separated (pure) are the data, once a test function (split) is performed on a tree node. It is noteworthy that during the feature reduction phase, we used a sparse sampling strategy, sampling 5×10^3 voxels from the training set. Once the feature selection was performed, we trained the final RF classifier using denser sampling (5×10^4 sampling voxels per training set). The final RF model was then used for the localization of the ToP and the CG on the unseen images.

Once the RF hypotheses were obtained, we performed GC regularization. The σ and λ parameters were set for all the regularization experiments as follows: $\sigma=10$ and $\lambda=0.1$.

Fig. 2 and table 1 summarize the segmentation results obtained from the first (RF) and second (RF-GC) stage of our framework. Overall, GC regularization improves the RF-based segmentation outcome. The main effect of the regularization is the removal of holes inside the foreground and islands outside it, which exist in the RF hypotheses (fig. 1). Also, in both ToP and CG segmentation, all metrics demonstrate improved segmentation accuracy as the size of the training set increases. The recorded mean values for the *CA*, *Tc* and *MaxDist* score for ToP segmentation improve from 94.1%, 73.6% and 6.2 mm for training sets of size 1 to 96.0%, 82.0%, and 4.7 mm respectively for training sets of size 21. The recorded mean values for the same scores for CG segmentation improve from 94.5%, 57.8% and 6.6 mm for training sets of size 1 to 96.5%, 73.6%, and 5.1 mm for training sets of size 21. It is noteworthy that the method provides accurate results even with small sized training sets (≈10 images).

Table 1 presents the *VolDiff* and *MeanDist* scores, for training sets of size 21, which can be considered equivalent to the leave-one-out experiments used in the study of Allen *et al.* [6]. Our method obtained better *VoldDiff* score for the ToP segmentation task but slightly worse for the CG segmentation task. The *MeanDist* score was better in both segmentation tasks. The CG is a small anatomical region. In such cases, few misclassified voxels

lead to a large decline of the accuracy metrics. We believe that our results are promising, given the fact that our method is fully automated and that no shape priors are imposed to the final delineation.

Table 1: *VolDiff* and *MeanDist* results for the Total Prostate and the Central Gland segmentation tasks.

	Metrics	
Method (ROI)	***VolDiff (%)*** ***μ (σ)***	***MeanDist (mm)*** ***μ (σ)***
RF (ToP)	8.5 (7.2)	1.1 (0.2)
RF-GC (ToP)	**7.3 (4.2)**	**1.0 (0.1)**
Allen *et al.* [6] (ToP)	11.1 (9.5)	4.1 (1.1)
RF (CG)	21.3 (13.1)	1.5 (0.4)
RF-GC (CG)	13.3 (10.9)	**1.2 (0.3)**
Allen *et al.* [6] (CG)	**11.9 (8.9)**	3.1 (2.5)

4. CONCLUDING REMARKS

In this study we presented a fully automated approach for the differential segmentation of the prostate. Our method consists of two stages: in the first stage, a RF classification produces hypotheses about the anatomy of interest, thus tackling the localization problem. In the second stage, GC optimization is used to regularize the previous hypotheses via the MAP-MRF framework. Our empirical performance evaluation shows that the framework can provide accurate segmentation outcomes for a challenging task. Our method demonstrates comparable or better results in a fully automated fashion compared with the semi-automatic method of Allen *et al.* [6].

Furthermore, fig. 2 demonstrates that the method can be used to produce fairly accurate results using small training sets. This is particularly useful for on-line interactive segmentation. Model-based approaches, such as that of [6] require an accurately annotated training set, which is generally obtained by manual segmentation. Fig. 2 shows that automatic classification starts improving after the first training example and classification close to final performance can be achieved after a small number of training examples. This can lead to a dramatic improvement in cognitive load in the training phase of statistical model building.

In our study the ground truth was produced by delineation of the anatomy of interest on a slice by slice basis, thus producing a result that is unlikely to be in complete agreement with the segmentation outcome of a 3-D segmentation method.

It is also noteworthy, that no shape priors were incorporated in the suggested approach. Future work could investigate if such priors could offer any further improvement to the segmentation accuracy.

5. ACKNOWLEDGEMENTS

The project has been partially funded by the United Kingdom Biotechnology and Biological Sciences Research Council (BBSRC).

6. REFERENCES

[1] A. Thorpe and D. Neal, "Benign prostatic hyperplasia," *Lancet*, vol. 361, no. 9366, pp. 1359–67, 2003.

[2] A. Tewari, K. Shinohara and P. Narayan, "Transitional Zone Volume and Transitional Zone Ratio: Predictor of Uroflow Response to Finasteride Therapy in Benign Prostatic Hyperplasia Patients," *Urology*, vol. 45, no. 2, pp. 258–265, 1995.

[3] N. Makni *et al.*, "Automatic 3D segmentation of prostate in MRI combining a priori knowledge, Markov fields and Bayesian framework," in *Proc. EMBS,* 2008, pp. 2992-2995.

[4] S. Klein *et al.*, "Automatic segmentation of the prostate in 3-D MR images by atlas matching using localized mutual information," *Medical Physics*, vol. 35, no. 4, pp. 1407–1417, 2008.

[5] E. Moschidis and J. Graham, "Propagating interactive segmentation of a single 3D example on similar images: an evaluation study using MR images of the prostate," in *Proc. ISBI*, 2011, pp. 1472– 1475.

[6] P. D. Allen *et al.,* "Differential segmentation of the prostate in MR images using combined 3D shape modelling and voxel classification," in *Proc. ISBI*, 2006, pp. 410–413.

[7] E. Moschidis and J. Graham, "A systematic performance evaluation of interactive image segmentation methods based on Simulated User Interaction," in *Proc. ISBI*, 2010, pp. 928– 931.

[8] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[9] A. Jaiantilal, "randomforest-matlab," November 2011. [Online]. Available: http://code.google.com/p/randomforest-matlab/

[10] J. Santner *et al.*, "Interactive Texture Segmentation using Random Forests and Total Variation," in *Proc. BMVC*, 2009, pp. 1-12.

[11] R. M. Haralick, "Statistical and structural approaches to texture," *Proc. IEEE*, vol. 67, no. 5, pp. 786–804, 1979.

[12] M. T. Suzuki, Y. Yaginuma and H. Kodama, "A texture energy measurement technique for 3D volumetric data," in *Proc. SMC*, 2009, pp. 3779-3785.

[13] D. Greig, B. Porteous, and A. Seheult, "Exact maximum a posteriori estimation for binary images," *Journal of the Royal Statistical Society, Series B*, vol. 51, no. 2, pp. 271-279, 1989.

[14] Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images," in *Proc. ICCV,* 2001, vol. 1, pp.105–112.

[15] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE T-PAMI*, vol. 26, no. 9, pp. 1124–1137, 2004.

47. **The accuracy of prostate volume measurement from ultrasound images: A quasi-Monte Carlo simulation study using magnetic resonance imaging**. D.-O. Azulay, P. Murphy and J. Graham. *Computerised Medical Imaging and Graphics, **37(7)**, 628-636, 2013.*
doi: 10.1016/j.compmedimag.2013.09.001.

Computerized Medical Imaging and Graphics 37 (2013) 628–635

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

The accuracy of prostate volume measurement from ultrasound images: A quasi-Monte Carlo simulation study using magnetic resonance imaging

David-Olivier D. Azulay [a,*], Philip Murphy [a], Jim Graham [b]

[a] Precision Medicine, Pfizer, Sandwich CT13 9NJ, UK
[b] Centre for Imaging Science, Manchester Academic Health Sciences Centre, University of Manchester, M13 9PT, UK

ARTICLE INFO

Article history:
Received 27 June 2012
Received in revised form 3 July 2013
Accepted 3 September 2013

Keywords:
quasi-Monte Carlo simulation
Ultrasound
Magnetic resonance imaging
Prostate
Volume
Estimation

ABSTRACT

Prostate volume is an important parameter to guide management of patients with benign prostatic hyperplasia (BPH) and to deliver clinical trial endpoints. Generally, simple 2D ultrasound (US) approaches are favoured despite the potential for greater accuracy afforded by magnetic resonance imaging (MRI) or complex US procedures. In this study, different approaches to estimate prostate size are evaluated with a simulation to select multiple organ cross-sections and diameters from 22 MRI-defined prostate shapes. A quasi-Monte Carlo (qMC) approach is used to simulate multiple probe positions and angles within prescribed limits resulting in a range of dimensions. The basic ellipsoid calculation which uses two scanning planes compares well to the MRI volume across the range of prostate shapes and sizes ($R = 0.992$). However, using an appropriate linear regression model, accurate volume estimates can be made using prostate diameters calculated from a single scanning plane.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

Estimation of a prostate volume is an integral component in the evaluation of patients with BPH. While measurement performance is important for routine clinical assessment, precision, reproducibility and practicality of measurement are crucial to enable derivation of clinical trial endpoints [6,15].

Volume estimation of the prostate is normally carried out using trans-rectal US images. The measurement can be made by a planimetric method where a stack of 2D slices is constructed by step movements of the ultrasound probe. Clinically it is more convenient to make NPUS measurements of prostate diameters in the latero-lateral (LL), anterio-posterior (AP) and cranio-caudal (CC) directions from 2D images using the assumption of an ellipsoidal shape. By displacing the probe the operator determines the orientation for the best estimate of diameters.

Bazinet et al. [3] have noted that volume measured in this way is inaccurate when compared with US (PUS) or magnetic resonance (MR) and conducted a reproducibility trial, finding that differences in volume estimates of up to 25% could be obtained in successive NPUS examinations of the same patient.

Direct 3D measurements of a volume by PUS or MR have the potential to deliver more accurate and reproducible results but are considerably more expensive, acquisition is time consuming and, without automated interpretation, analysis is highly labour-intensive [13,26,1]. US remains widely available to the urologist and therefore remains the most practical method of estimating prostate volume [22]. In this study we sought to investigate which estimation method can best be used with 2D NPUS to predict prostate volume.

Allen et al. [2] have investigated the use of active shape model (ASM) search for measuring the volumes of the complete gland and central gland using MR images. An ASM is a statistical model of shape, built from a training set of images which have been segmented (the important surfaces defined) manually [7,9]. For their study [2] collected twenty two fat-suppressed MR images of prostates from patients diagnosed with BPH attending the urology clinic at Salford Royal Hospital, UK.

The objective in this study is to simulate volume measurements by NPUS using the manual segmentation obtained from the ASM study in order to quantify the limitations on the accuracy and reproducibility of taking this approach. Rahmouni et al. [21] and Lee and Chung [13] have reported that measurements of volume using MR accurately represent the volumes of real specimens after prostatectomy. We therefore take the manually delineated borders of the prostate used by Allen et al. [2] to be realistic ground truth representing the prostate boundary.

* Corresponding author. Tel.: +44 1304 620 314.
E-mail addresses: david-olivier.azulay@pfizer.com, david-olivier.azulay@wanadoo.fr (D.-O.D. Azulay).

0895-6111/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compmedimag.2013.09.001

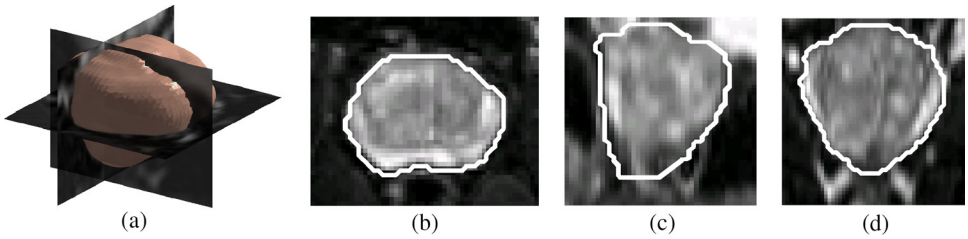


Fig. 1. Example of a 3D rendered prostate: the approximately elliptical boundaries are taken to represent the projections that would be observed in NPUS images.

Fig. 1a shows a 3D rendered prostate volume from the manually annotated data. Axial, sagittal and coronal sections are indicated and the corresponding slices through the volume are shown in Fig. 1b–d, with the prostate contour superimposed.

While Allen et al. [2] addressed the segmentation of both the whole gland and the central gland, only the whole gland is considered here, as this is what may be measured using NPUS. Furthermore, internal structures are not necessarily comparable between MR and US images. We make the assumption that the US-defined borders are close approximations to the borders defined in MR images. Making realistic hypotheses about the range of possible orientations of the NPUS probe, we have implemented a qMC method to assess the distribution of the error in volume estimates produced by the different direct ellipsoidal formulae and the linear models extrapolated from these formulae. The errors have been successively quantified with standard linear regression goodness-of-fit parameters, correlation coefficients and Bland–Altman (BA) plots.

2. Methods

The most common clinical approach to estimating volume using NPUS is to capture US images from two approximately orthogonal planes passing through the centre of the prostate. Rather than collecting US and MR images from volunteers, we simulate the US measurements using a qMC procedure that allows us to investigate the error distributions arising from varying the choices of the US planes used and the analytical methods applied to produce the volume estimates. The reference volume for the simulation procedure is a triangulated closed surface defining the 3D volume of the prostate derived from manual segmentation of MR images (see Fig. 1a). An evenly distributed selection of pairs of orthogonal cuts through this surface is simulated to generate perimeters that would have been obtained from the corresponding NPUS images. There is evidently an infinite number of planes cutting a 3D volume but the generation of a representative subset is required. Triplets of points distributed on the surface of a sphere are successively picked to define planes that belong to this subset.

2.1. Devising an evenly distributed set of orthogonal cuts

The general procedure is illustrated with the 2D example sketched in Fig. 2.

The blue line represents the outer boundary of the 2D slice of an artificial prostate, the dotted rectangle its bounding box. Cuts are simulated across this boundary by selecting random pairs of points on the circumference of a circle. The figure shows six points uniformly distributed on the circumference of a small circle near the centre of the prostate. A small circle is preferred, rather than a large one circumscribing the bounding box because the cuts should pass close to the centre of the prostate. There are several cuts that are geometrically equivalent in that they divide the circle into segments whose areas are in the same ratio (e.g. AB, BC, CD, DE, EF and FA, or AC, AE, BD, BF, CE and DF). However, only some of those are suitably oriented. To start with the angle between the probe and the anterior-posterior axis should be small to reflect probe positioning. The selections are therefore limited by requiring that this angle should be less than 22.5°. While somewhat arbitrary, this tolerance probably represents an upper limit on a human observer's ability to estimate the orientation of the LL diameter. On Fig. 2, AD is an acceptable choice because the angle α with the AP axis is small, whereas CE is not acceptable because the angle β is too large. Then it is assumed that an NPUS operator should judge fairly accurately the position of the centre of the prostate which would be close to the middle zone of the bounding box. In determining next the LL diameter, it is also assumed that the two diameters should intersect close to the true centre.

The extension to 3D is less straightforward because there is no easy parameterisation that allows fine control over the likely position and orientation of the plane that is necessary to simulate the clinical situation. The obvious extension to Fig. 2 is to distribute points on the surface of a sphere. However, distributing points evenly on a sphere is a complex problem [23]. The solution we have adopted is due to Li et al. [14], which generates a "low discrepancy" sequence of points by using the terms of the sequence to approximate the sizes of surface on the sphere defined by neighbouring groups of points. In other words, each point on the surface is equidistant to its nearest neighbours. This guarantees the existence of a bounded error (low discrepancy with the uniform distribution).

A plane is fully determined from the selection of three distinct points. Because these points are drawn from an evenly spaced sequence (and not from a random sequence), the simulation is based on a qMC (and not a Monte Carlo) method [18]. Sloane et al. [24] have calculated a convenient set of distributions, which were downloaded and used for this study (see Fig. 3). Planes generated

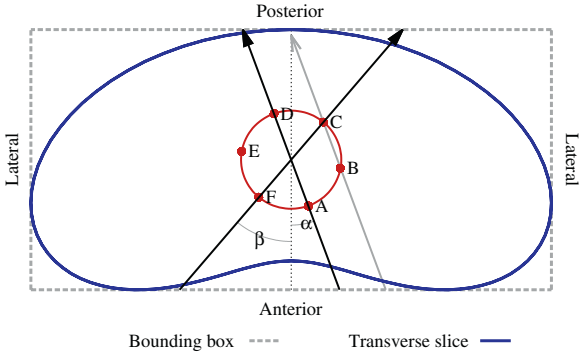


Fig. 2. Principle of the cut selection procedure on the schematic of a 2D slice: the α cut is accepted but not the β cut because it is too far from the AP axis. The BC cut would be equally accepted as its angle is α as well, even though it does not pass through the center of the circle.

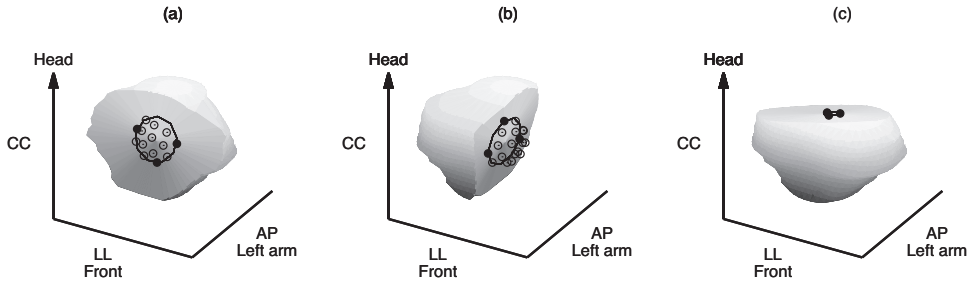


Fig. 3. Example of three possible cuts in 3D. Three points from the uniform distribution of 50 points on a sphere are chosen to define each cutting plane; (a) cut defining CC and LL on the coronal plane, (b) defining AP and CC on the sagittal plane, (c) defining AP and LL along the transverse plane. (c) is an example of an extreme cut that passes just through the central sphere trimming the top of the prostate horizontally.

Table 1
Linear regressions on the top 5% and 10% estimates. Every one of the 10 listed formulae generates a different set of 22 estimates that was used to predict the 22 fixed MR volumes. All the SE_b are statistically significant ($p^{\ddagger} < 0.001$) demonstrating the appropriateness of the linear fits. The likelihoods of the χ^2 observations were calculated as well to confirm the validity of the linear regressions ($Q'' > 0.1$, $Q' > 0.01$ and $Q^{\star} > 0.001$). The CI of the average $\overline{V_d}$ is printed in the last column. The first three linear regressions on the top 5% estimates are illustrated in Fig. 7.

	Top 5%					Top 10%				
V_d (cm^3)	MR = $a + bV_d$	R	SE_b	$\overline{V_d}$	CI($\overline{V_d}$)	MR = $a + bV_d$	R	SE_b	$\overline{V_d}$	CI($\overline{V_d}$)
AP.CC.LL	$2.43+0.962V_d''$	0.992	$0.028^{\ddagger}$	44.2	1.14	$2.80+0.974V_d''$	0.990	$0.030^{\ddagger}$	43.3	1.23
AP . LL^2	$0.08+0.894V_d''$	0.983	$0.037^{\ddagger}$	50.2	1.63	$0.65+0.920V_d''$	0.987	$0.033^{\ddagger}$	48.2	1.40
CC . LL^2	$-5.16+0.818V_d''$	0.983	$0.034^{\ddagger}$	61.3	1.64	$-3.88+0.822V_d''$	0.981	$0.036^{\ddagger}$	59.4	1.71
LL . CC^2	$-6.38+0.819V_d''$	0.976	$0.041^{\ddagger}$	62.6	1.94	$-5.73+0.824V_d''$	0.975	$0.042^{\ddagger}$	61.5	1.96
LL . AP^2	$8.13+0.995V_{d'}$	0.969	$0.057^{\ddagger}$	37.0	2.21	$8.52+1.024V_{d'}$	0.968	$0.059^{\ddagger}$	35.6	2.23
CC . AP^2	$5.87+1.031V_{d'}$	0.965	$0.062^{\ddagger}$	37.9	2.32	$6.62+1.047V_{d'}$	0.964	$0.065^{\ddagger}$	36.6	2.38
AP . CC^2	$-1.37+0.900V_d^{\star}$	0.961	$0.058^{\ddagger}$	51.5	2.46	$-1.04+0.918V_{d'}$	0.965	$0.056^{\ddagger}$	50.1	2.35
AP^3	$11.36+1.090V_d\star$	0.948	$0.082^{\ddagger}$	30.8	2.83	$12.30+1.114V_d\star$	0.941	$0.090^{\ddagger}$	29.3	3.01
LL^3	$-7.34+0.708V_d\star$	0.941	$0.057^{\ddagger}$	73.9	3.00	$-6.90+0.740V_d\star$	0.956	$0.051^{\ddagger}$	70.1	2.61
CC^3	$-6.18+0.673V_d\star$	0.887	$0.078^{\ddagger}$	75.9	4.11	$-5.83+0.686V_d\star$	0.894	$0.077^{\ddagger}$	74.0	3.98

by close neighhbours would miss the center of the sphere and therefore the approximate center of the prostate. These extreme cases contribute to the observed variability in volume estimates.

There are three formulae for calculating volumes based on different ellipsoidal approximations, requiring up to three diameters (the $\pi/6$ term will be omitted when referring to a specific formula to simplify the notations):

- spherical $\pi/6 \times d_1^3$,
- spheroidal $\pi/6 \times d_1^2 \times d_2$ which falls into either the egg-shaped prolate ($d_1 < d_2$) or disk-shaped oblate ($d_1 > d_2$) categories,
- ellipsoidal $\pi/6 \times d_1 \times d_2 \times d_3$.

Each one of the three diameters d_1, d_2 and d_3 can successively be AP, CC and LL leading to ten possibilities (see Table 1).

The LL, AP and CC diameters in the transverse and sagittal planes should all intersect at their middle point and be perpendicular to each other, which means that the sagittal plane needs to be perpendicular to the transverse one while containg the AP diameter. Strict perpendicularity between planes, not diameters, is maintained during the qMC simulation to prevent the overestimation of the CC diameter. That is due to the latent inclination of the sagittal plane which artificially increases the CC diameter when manually measured [8].

2.2. Simulation algorithm

One potential approach could have been the exhaustive enumeration of perfectly orthogonal cutting diameters across the 3D volume. However, in a practical situation, the positions and angles of the diameters and planes are estimated visually and therefore of limited precision. Rahmouni et al. [21] and Kim and Kim [12] have reported that the choice of the transducer orientation relative to the prostate is operator dependent. The transverse slice is determined first since it is less error prone when delimited by the human eye [17,11]. As a consequence, some flexibility has been introduced in the automated procedure: for instance the longest diameter is drawn at the expense of true orthogonality. We set the tolerances on perpendicular angles between diameters to 5° and on middle point localisations to a 5% precision of the lengths of respective diameters. The purpose is to compensate for both the visual imprecision of the operator and the potential shape deformation of the prostate due to the probe.

The steps implemented to achieve the qMC simulation are summarised in the following algorithm:

Algorithm 2.1.

1: define points of a sphere such that each point is equidistant to its nearest neighbours
2: **repeat**
3: choose a triplet of points from this uniform distribution to define a transverse plane
4: compute the 2D transverse slice
5: search the longest AP diameter within the 22.5° tolerance angle of the real AP axis
6: search the longest LL diameter which is orthogonal to the AP diameter within the 5° tolerance angle and intersects at their middle points within the 5% tolerance distance
7: compute the 2D sagittal slice which is perpendicular to the transverse one and contains the AP segment
8: search the longest CC diameter which is orthogonal to the AP diameter within the 5° tolerance angle and intersects at their middle points within the 5% tolerance distance
9: **until** all triplets of points have been exhausted

A distribution of $n = 50$ points on the sphere was chosen, describing $C_3^{50} = 19,600$ distinct planes, out of which $m = 4808$ are properly oriented. Each accepted pair of planes provides estimates of the LL,

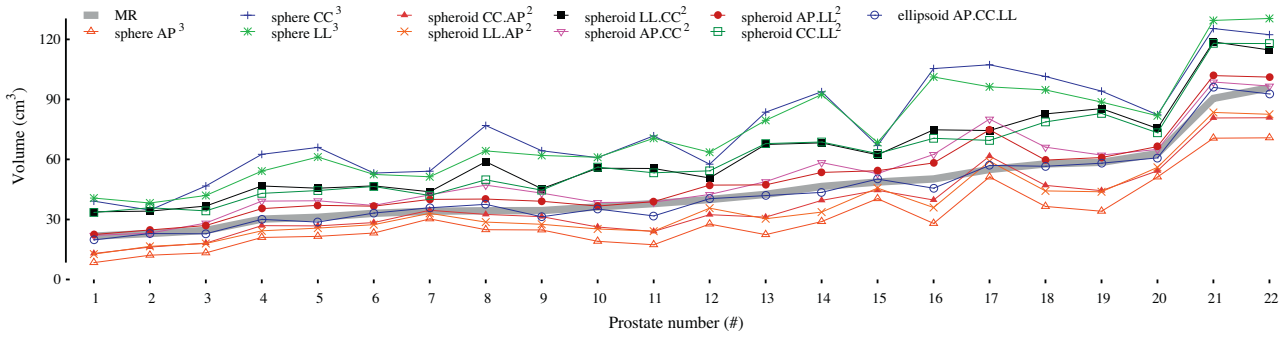


Fig. 4. Averages of the top 5% estimates for the ten formulae and each prostate, sorted by MR volume.

CC and AP diameters from which ten estimated volumes are computed: one with each of the formulae and every combination of parameters. These are the estimates to be compared with the true volume obtained from the manual segmentation of the MR images. In NPUS imaging an operator would normally seek to maximise the diameters. Kim and Kim [12] have shown that experienced operators produce larger estimates than beginner and trained ones. So all the main statistics in Section 3 are derived from the top 5% (240 largest) volume estimates to reflect how a user would be likely to decide on the maximum measured prostate size.

The figure of 5% is selected to provide sufficient statistics to construct a linear model, constrained to produce the largest estimates. The threshold is rather arbitrary and we include figures in Table 1 representing models calculated using the largest 10% for comparison.

2.3. Error prediction for the linear models

Each formula for calculating the volume from measured diameters corresponding to each pair of planes provides average and variance values for the volume of each of the 22 images. This forms a training set that is used to define a linear model of the form $y = a + bx$. The predictor, x, corresponds to one of the spherical, spheroidal or ellipsoidal estimates and the outcome, y, is the true MR volume [5].

Leave-one-out cross-validation analyses were completed in order to quantify the sensitivity of these linear models [20]. For every formula, 22 distinct training sets are created including 21 averages, by leaving one average out from the original set and a new linear regression calculated for each of them.

2.4. Assessing the agreement of the estimates with the MR volumes

Bland and Altman [4] described a statistical procedure that here can be used to evaluate the relative concordance between assumed ground truth from MR and simulated NPUS imaging as a function of prostate size. The BA plot shows how the difference between measurements of the same quantity is related to the mean of the measurement. If the differences follow a normal distribution with average ave and standard deviation SD then 95% of their values should be within the limits ave $\pm 1.96 \times SD$, also called limits of agreement (LoA).

3. Results

Fig. 4 indicates that application of different ellipsoidal approximations from the top 5% estimates tend to systematically overestimate or underestimate the true prostate volume. This error can be corrected by training a regression model that allows measured values to predict the true volume. Here we use a linear model of the form $y = a + bx$ for simplicity. A slope lower or higher than 1 will correct for overestimates and underestimates respectively [27,11].

Table 1 shows the linear models derived from the training data corresponding to each of the ellipsoidal calculations along with their correlation coefficients R, standard error of the slope (SE_b), mean value of the estimated volume and 95% confidence interval around the mean.

The statistical significance of the linear regression parameters can be estimated by the probability (Q) of the chi-square distribution [19]. Values of Q greater than 0.1 represent believable fits. The linear model is adequate to extrapolate the true volume of the prostate, especially when the $AP.CC.LL$, $AP.LL^2$, $CC.LL^2$ or $LL.CC^2$ formula is used.

The size of the 95% confidence interval (CI) of the average as a proportion of the average itself is about 3% for the first four ranked estimates. The arbitrary top 5% criterion reflects the experience of the operator searching for the maximum diameters. A less experienced operator is more likely to miss these maxima; this is quantified by the extension to 10%, that is to say the addition of 5% lower volumes to the previous calculations. Using the top 10% clearly results in slightly lower volume estimates (around 2 cm^3), however the model parameters and standard errors are not affected greatly by the acceptance threshold.

Fig. 5 shows the variation in the correlation coefficient that arises when predicting the true (MR) volume for each of the training set using a model constructed in a leave-one-out manner from the remaining data. Values for the three-diameter ellipsoidal calculation and the two estimates using only two diameters with the highest correlation are shown. Fig. 6 shows the Bland–Altman plots for the same three volume estimation formulae. The grey bands indicate the 95% confidence limits around the limits of agreement values.

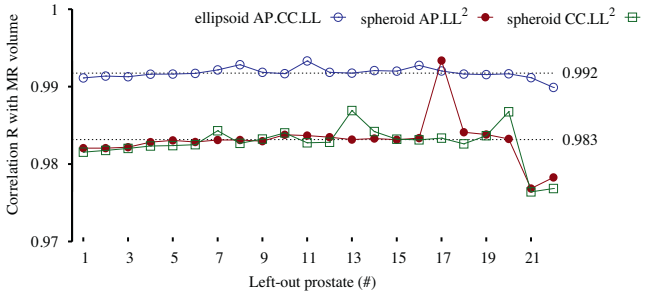


Fig. 5. Leave-one-out correlation coefficient R from the linear regressions on the top 5% estimates. The horizontal dotted lines display the corresponding values of the complete training set shown in Table 1.

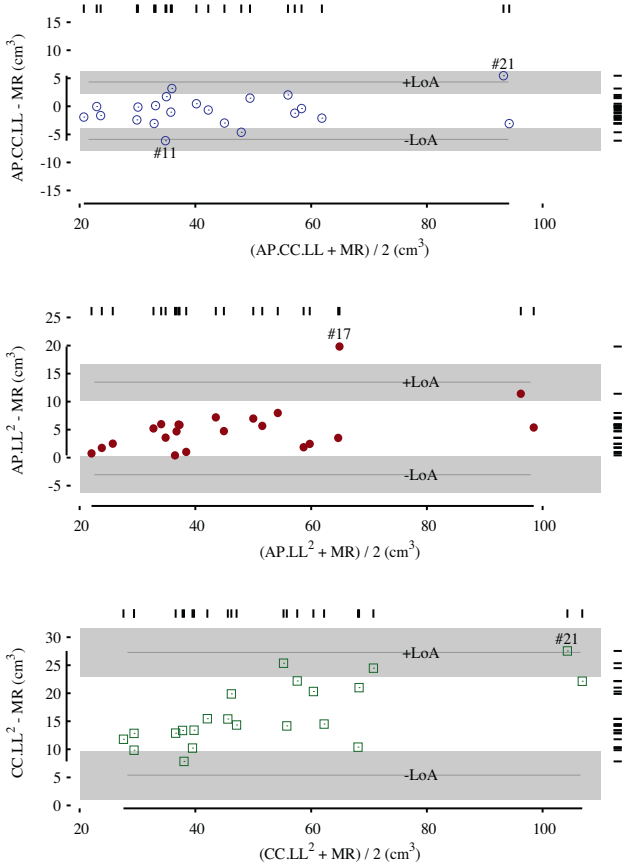


Fig. 6. BA plots for the three direct formulae. Estimates are obtained from the top 5% largest volumes and scales are preserved across the plots to facilitate the visual comparisons.

The means μ and standard deviations σ of three of the formulae and their respective diameters are listed in Table 2 along with the measured ground-truth volume (MR).

Table 2 also shows the distribution of the 4808 simulated ellipsoidal volumes for each prostate as a histogram. The vertical bar in the histogram marks the reference volume, that is to say the 0% deviation from the MR volume.

4. Discussion

A good linear model depends on the reproducibility of the deviation, rather than its amplitude. For instance, in Table 2, the $CC.LL^2$ direct estimate systematically overestimates the MR by 39% on average; however its derived linear model is acceptable as an estimate according to the goodness-of-fit parameters R and SE_b in Table 1.

[4] explain why correlation coefficients can be misleading to quantify the quality of the agreement between two measurements and suggest more appropriate calculations to perform better assessments. They complement the standard linear regression plot presented in Fig. 7 by the examination of the estimate errors, *i.e.* the differences between the real and approximated values (Fig. 6). The CI of the LoA are delimited by the gray bands whose widths are both equal to $1.96 \times CI(\pm LoA)$.

The $CC.LL^2$ formula exhibits wide gray bands delimiting the 95% LoA confidence intervals. Those are inadequate for an approximation method even if the correlation coefficient R for the $CC.LL^2$ model is appealingly high in Table 1. Conversely, the spheroidal $AP.LL^2$ estimate is satisfactory according to both its correlation coefficient and its BA plot notwithstanding the fact that its performance is undermined by the extreme outlier #17 (see also Fig. 5). The high correlation of the $AP.LL^2$ model approximation indicates that most of the prostates in the sample can be reasonably approximated as oblate spheroids with circular cross-section in the coronal plane (the LL and CC diameters are often of similar length in Table 2). Examination of the shape of prostate #17 indicates that its shape approximates a prolate spheroid with a circular cross-section in the transverse plane. Fig. 5 reveals that the correlation coefficient of the $AP.LL^2$ spheroidal model is better than those of the ellipsoidal model when this prostate does not contribute to the model. The correlation coefficient is slightly higher than the ellipsoidal model in that case.

Fig. 4 demonstrates that the direct ellipsoidal calculation is consistently close to the truth when restricted to the top 5% estimates. Nevertheless there are significant variations as indicated by the horizontal range of all the histograms displayed in Table 2: estimates can vary from −50% to +20% which is in agreement with the observation of a high variability in reproducibility studies by [3]. The histograms also reveal that the majority of the qMC ellipsoidal volumes underestimate the true value. Because the prostate is not a perfectly ellipsoidal object [21], the probability of coming across an orientation which captures the three largest diameters in a given pair of orthogonal planes at the same time is quite low. This explains why underestimation is expected and justifies the use of the largest volumes only in the analysis [16,10,22].

The volume based on the ellipsoidal choice benefits from narrower confidence intervals, but requires the acquisition of both the transverse and sagittal planes. The $AP.LL^2$ linear regression appears to be of good accuracy but is obtained from the transverse plane only; [10], [25] and [28] raised the same practical consideration in terms of clinical feasibility. As determined by Fig. 4, the spherical formula provides a lower limit on the volume estimate if the

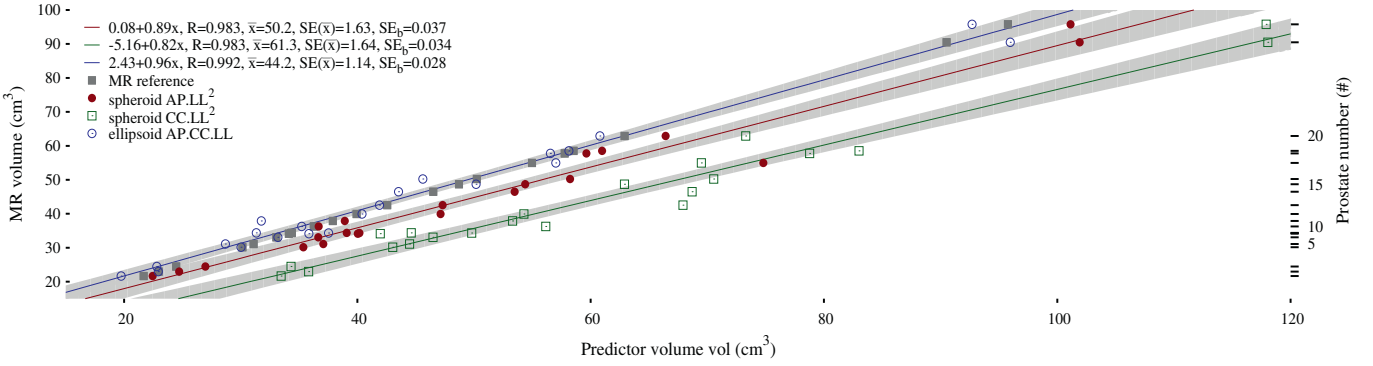


Fig. 7. Linear regressions on the averages of the top 5% volumes for three formulae (ellipsoidal and two spheroidal models). An overestimate is found on the right side of the MR reference value, an underestimate on the left side. The CI are delimited by the grey bands centered around their line equation.

Table 2
Means (μ) and standard deviations (σ) for the volume estimates for three of the formulae: ellipsoid ($AP.CC.LL$) and spheroids ($AP.LL^2$ and $CC.LL^2$) together with the values of the diameters giving rise to these volumes for each image of the 22 subjects. These are derived from the 5% highest volumes obtained in the qMC simulation. The histograms show the distribution of all the ellipsoidal volumes (grey) as a percentage deviation from the MR volume (indicated by the vertical dotted line), while the bins shown in red represent the 5% largest estimates used to derive the statistics shown. The average deviation of each estimate from the true volume is denoted Δ. Overall estimates of μ and σ appear at the foot of the table.

#		MR	AP	CC	LL		AP.CC.LL	Δ	AP	LL	$AP.LL^2$	Δ	CC	LL	$CC.LL^2$	Δ
1	μ	21.7	2.4	4.1	3.9		19.8	−8.9%	2.4	4.3	22.4	+3.6%	3.7	4.2	33.5	+54.4%
	σ		0.1	0.2	0.2		0.6	2.8%	0.0	0.1	0.5	2.3%	0.3	0.1	1.0	4.6%
2	μ	23.0	2.7	3.9	4.2		22.9	−0.2%	2.7	4.2	24.7	+7.6%	3.9	4.2	35.8	+56.0%
	σ		0.1	0.1	0.1		0.4	1.7%	0.1	0.1	0.8	3.3%	0.1	0.1	1.1	4.8%
3	μ	24.5	2.8	4.0	3.9		22.8	−6.8%	2.8	4.3	27.0	+10.2%	3.6	4.3	34.3	+40.2%
	σ		0.1	0.3	0.3		0.3	1.3%	0.1	0.1	1.4	5.5%	0.2	0.1	1.1	4.4%
4	μ	30.1	3.2	4.6	3.9		30.0	−0.4%	3.1	4.7	35.4	+17.3%	3.7	4.7	43.0	+42.7%
	σ		0.1	0.4	0.4		0.5	1.6%	0.1	0.1	1.9	6.4%	0.1	0.1	2.2	7.3%
5	μ	31.1	3.3	4.4	3.9		28.7	−7.7%	3.1	4.8	37.1	+19.2%	3.6	4.9	44.5	+43.0%
	σ		0.1	0.6	0.6		0.8	2.5%	0.2	0.2	2.9	9.5%	0.1	0.2	2.4	7.7%
6	μ	33.1	3.4	4.3	4.3		33.2	+0.4%	3.3	4.6	36.6	+10.8%	4.2	4.6	46.5	+40.6%
	σ		0.1	0.2	0.3		0.4	1.3%	0.1	0.1	0.9	2.7%	0.1	0.1	1.5	4.6%
7	μ	34.1	3.8	4.4	4.2		35.8	+5.0%	3.7	4.6	40.0	+17.3%	4.0	4.5	42.0	+23.0%
	σ		0.1	0.2	0.2		0.4	1.1%	0.2	0.1	1.5	4.3%	0.3	0.2	1.3	3.7%
8	μ	34.3	3.5	5.0	4.1		37.5	+9.3 %	3.2	4.9	40.2	+17.0 %	4.2	4.8	49.8	+45.0 %
	σ		0.2	0.3	0.3		1.3	3.7%	0.2	0.2	3.2	9.4%	0.5	0.4	4.0	11.7%
9	μ	34.4	3.6	4.5	3.8		31.3	−8.9%	3.2	4.8	39.1	+13.7%	3.6	4.9	44.6	+29.7%
	σ		0.2	0.5	0.5		1.3	3.7%	0.2	0.2	2.0	5.9%	0.1	0.1	1.6	4.6%
10	μ	36.3	3.2	4.7	4.5		35.2	−2.9%	3.0	4.8	36.7	+1.1 %	4.5	4.9	56.1	+54.8 %
	σ		0.1	0.2	0.2		0.7	2.0%	0.2	0.1	1.4	3.8%	0.1	0.1	1.5	4.2%
11	μ	37.9	3.1	4.6	4.4		31.8	−16.2 %	2.9	5.1	38.9	+2.7 %	4.0	5.0	53.3	+40.7 %
	σ		0.2	0.5	0.4		1.1	2.8%	0.2	0.2	1.7	4.5%	0.2	0.2	2.1	5.6%
12	μ	39.9	3.7	4.3	4.9		40.4	+1.2%	3.7	4.9	47.1	+18.0%	4.3	4.9	54.3	+35.9%
	σ		0.1	0.1	0.1		0.5	1.4%	0.1	0.1	1.5	3.8%	0.1	0.1	0.7	1.8%
13	μ	42.6	3.4	5.0	4.8		41.9	−1.6%	3.2	5.3	47.3	+11.2 %	4.7	5.3	67.9	+59.6 %
	σ		0.1	0.3	0.3		0.9	2.1%	0.2	0.1	2.2	5.1%	0.2	0.1	1.3	3.2%
14	μ	46.5	3.7	5.3	4.2		43.5	−6.4 %	3.3	5.6	53.5	+15.0 %	4.2	5.6	68.7	+47.7 %
	σ		0.2	0.4	0.3		1.1	2.4%	0.1	0.1	2.2	4.7%	0.1	0.1	2.3	5.0%
15	μ	48.7	4.1	4.8	4.9		50.2	+3.0 %	4.1	5.0	54.4	+11.7 %	4.7	5.1	62.9	+29.1 %
	σ		0.1	0.2	0.2		0.5	1.0%	0.1	0.1	1.5	3.1%	0.1	0.1	1.5	3.1%
16	μ	50.2	3.6	5.4	4.5		45.6	−9.2 %	3.4	5.7	58.2	+15.9 %	4.1	5.8	70.5	+40.4 %
	σ		0.2	0.7	0.6		1.8	3.6%	0.2	0.2	3.8	7.5%	0.1	0.1	3.0	6.0%

Table 2 (*Continued*)

#		MR	AP	CC	LL		AP.CC.LL	Δ	AP	LL	AP . LL^2	Δ	CC	LL	CC . LL^2	Δ
17	μ	55.0	4.5	5.3	4.6		57.0	+3.7 %	4.4	5.7	74.8	+36.1 %	4.1	5.7	69.5	+26.4 %
	σ		0.1	0.7	0.7		0.9	1.6%	0.1	0.2	5.5	9.9%	0.1	0.2	4.7	8.5%
18	μ	57.8	4.0	5.4	5.0		56.6	−2.1 %	3.7	5.5	59.6	+3.2 %	4.8	5.6	78.8	+36.3 %
	σ		0.2	0.4	0.3		2.8	4.9%	0.3	0.2	2.4	4.2%	0.2	0.2	3.2	5.6%
19	μ	58.5	3.9	5.4	5.3		58.1	−0.7 %	3.8	5.5	61.0	+4.2 %	5.4	5.4	83.0	+41.8 %
	σ		0.1	0.1	0.2		0.9	1.5%	0.1	0.1	1.4	2.3%	0.2	0.1	1.5	2.6%
20	μ	62.9	4.4	5.2	5.2		60.8	−3.4 %	4.4	5.4	66.4	+5.6 %	5.1	5.2	73.3	+16.5 %
	σ		0.2	0.2	0.1		1.8	2.9%	0.2	0.1	2.5	3.9%	0.3	0.2	2.3	3.7%
21	μ	90.5	5.0	6.0	6.1		96.0	+6.0 %	5.0	6.2	101.9	+12.6 %	6.0	6.1	118.0	+30.4 %
	σ		0.1	0.1	0.1		1.3	1.5%	0.1	0.1	1.0	1.1%	0.1	0.0	0.7	0.8%
22	μ	95.8	5.0	5.8	6.1		92.7	−3.2 %	5.0	6.2	101.1	+5.6 %	5.8	6.3	117.9	+23.1 %
	σ		0.2	0.2	0.2		1.5	1.6%	0.2	0.1	3.0	3.1%	0.2	0.1	2.5	2.7%
	$\overline{\mu}$	44.9	3.6	4.8	4.6		44.2	−2.3 %	3.5	5.1	50.2	+11.8 %	4.4	5.1	61.3	+39.0 %
	$\overline{\sigma}$		0.1	0.3	0.3		1.0	2.2 %	0.1	0.1	2.0	4.8 %	0.2	0.1	2.0	4.8%

Table 3
Mapping between measured values of first three non-planimetric approximations of Table 1 and the estimates of true (MR) volume V_e using the regression models of Fig. 7. The left column provides some example values of the measured direct volume V_d. The remaining columns show the corresponding estimates V_e of the true (MR) volume derived from each model, together with their CI.

V_d	$V_d = \text{AP} . \text{CC} . \text{LL}$ $V_e = 2.43 + 0.962 V_d$ $V_e \pm \text{CI}(V_e)\text{cm}^3$	$V_d = \text{AP} . \text{LL}^2$ $V_e = 0.08 + 0.894 V_d$ $V_e \pm \text{CI}(V_e)\text{cm}^3$	$V_d = \text{CC} . \text{LL}^2$ $V_e = -5.16 + 0.818 V_d$ $V_e \pm \text{CI}(V_e)\text{cm}^3$
20	21.68 ± 1.81	17.97 ± 2.85	11.19 ± 3.39
30	31.31 ± 1.41	26.91 ± 2.26	19.37 ± 2.79
40	40.93 ± 1.17	35.86 ± 1.81	27.54 ± 2.25
50	50.56 ± 1.19	44.80 ± 1.63	35.72 ± 1.83
60	60.18 ± 1.46	53.75 ± 1.79	43.90 ± 1.65
70	69.81 ± 1.88	62.69 ± 2.24	52.07 ± 1.76
80	79.43 ± 2.37	71.64 ± 2.83	60.25 ± 2.12
90	89.06 ± 2.90	80.58 ± 3.49	68.43 ± 2.64
100	98.68 ± 3.44	89.53 ± 4.20	76.61 ± 3.23

AP diameter is used and an upper limit if the *LL* diameter is used. Generally speaking the AP diameter is the shortest and the LL diameter is the longest, with the result that the $AP . LL^2$ oblate spheroid provides a suitable estimate of volume.

The regression lines specified in Fig. 7 allow us to estimate the true volume of the prostate given specific measurements of the ellipsoidal ($AP . CC . LL$), oblate ($AP . LL^2$) and prolate ($CC . LL^2$) approximations. Table 3 lists the estimated values V_e corresponding to examples of each of these direct measurements V_d together with associated confidence limits.

For instance, a direct volume of 50 cm^3 using the $AP . LL^2$ approximation will result in a 95% chance that the true volume is 44.8 cm^3.

Using this table as a guide, measured values of each of these three approximations can be translated into an estimated true volume.

The results of the leave-one-out exercise in Fig. 5 shows how sensitive the linear models are to large prostates. [16], [13] and [10] already noticed that correlation coefficients were different if small or large prostates were categorised. The ellipsoidal model estimate is clearly more stable across all the leave-one-out tests.

5. Conclusion

Using ground truth from manually segmented MR images of the prostates the precision of direct volume estimates given by several approximation formulae was evaluated. This has been achieved with simulations of NPUS measurements covering a range of reasonable combinations of capture planes and estimates of the prostate dimension in each of these.

Despite the relatively small number of subjects, our observations are consistent with previously published results, such as the high variability of replicates [3], the sensitivity of the direct volume estimations to large prostates [16,13,22] and the good accuracy and practicality of the $AP . LL^2$ estimates [10].

We have measured the parameters of a regression formula that generates an estimate of prostate volume, using an ellipsoidal approximation, that is more accurate than simply using the direct ellipsoidal calculation. This estimate requires the measurement of prostate diameters in two orthogonal planes. Using a single plane, a volume estimate of good accuracy can be also obtained from the regression model for the spheroidal model based on $AP . LL^2$. The slope of this model is less than unity, indicating that

an estimate based only on the direct spheroidal formula would tend to overestimate the volumes of large prostates.

We have demonstrated the use of a simulation study for interrogating sources of variability in estimating prostate volumes. This variability arises from factors such as image planes not intersecting at the centre or not being accurately perpendicular, operator proficiency in finding the largest diameters, oddly shaped prostates, etc.. This has resulted in an understanding of the sources of variability, and hence has enabled us to identify the most appropriate measurement strategy.

The approach could also be extended to optimise volume estimation in other organs such as the liver by studying the 3D shape.

References

[1] Aarnink R, De La Rosette J, Debruyne F, Wijkstra H. Reproducibility of prostate volume measurements from transrectal ultrasonography by an automated and a manual technique. Br J Urol 1996;78:219–23.

[2] Allen PD, Williamson D, Graham J, Hutchinson C. Differential segmentation of the prostate in MR images using combined 3D shape modelling and voxel classification. In: Kovacevic J, Meijering E, editors. IEEE international symposium on biomedical imaging. Arlington; 2006 April. p. 410–3.

[3] Bazinet M, Karakiewicz PI, Aprikian AG, Trudel C, Péloquin F, Dessureault J, et al. Reassessment of nonplanimetric transrectal ultrasound prostate volume estimates. Urology 1996 June;47:857–62.

[4] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986:307–10.

[5] Bland M. An introduction to medical statistics. Oxford: Oxford University Press; 2000.

[6] Chen ME, Troncoso P, Johnston D, Tang K, Babaian RJ. Prostate cancer detection: relationship to prostate size. Urology 1999;53:764–8.

[7] Cootes T, Taylor C, Cooper D, Graham J. Active shape models-their training and application. Comput Vis Image Understanding 1995;61:28–59.

[8] Dähnert WF. Determination of prostate volume with transrectal US for cancer screening. Radiology 1992:625–6.

[9] Davies R, Twining C, Cootes T, Waterton J, Taylor C. A minimum description length approach to statistical shape modeling. IEEE Trans Med Imaging 2002;21:525–37.

[10] Eri L, Thomassen H, Brennhovd B, Håheim L. Accuracy and repeatability of prostate volume measurements by transrectal ultrasound. Prostate Cancer Prostatic Dis 2002;5:273–8.

[11] Jeong CW, Park HK, Hong SK, Byun S-S, Lee HJ, Lee SE. Comparison of prostate volume measured by transrectal ultrasonography and MRI with the actual prostate volume measured after radical prostatectomy. Urol Int 2008;81:179–85.

[12] Kim SH, Kim SH. Correlation between the various methods of estimating prostate volume: transabdominal, transrectal, and three-dimensional US. Kor J Radiol 2008;9:134–9.

[13] Lee JS, Chung BH. Transrectal ultrasound versus magnetic resonance imaging in the estimation of prostate volume as compared with radical prostatectomy specimens. Urol Int 2007;78:323–7.

[14] Li X, Wang W, Martin RR, Bowyer A. Using low-discrepancy sequences and the Crofton formula to compute surface areas of geometric models. Computer-aided Des 2003;35:771–82.

[15] Loeb S, Han M, Roehl KA, Antenor JAV, Catalona WJ. Accuracy of prostate weight estimation by digital rectal examinantion versus transrectal ultrasonography. J Urol 2005;173:63–5.

[16] Matthews GJ, Motta J, Fracchia JA. The accuracy of transrectal ultrasound prostate volume estimation: clinical correlations. J Clin Ultrasound 1996;24:501–5.

[17] Nathan M, Seenivasagam K, Mei Q, Wickham J, Miller R. Transrectal ultrasonography: why are estimates of prostate volume and dimension so innacurate? Br J Urol 1996;77:401–7.

[18] Niederreiter H. Quasi-Monte Carlo methods and pseudo-random numbers. Bull Am Math Soc 1978;84:957–1041.

[19] Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical recipes: the art of scientific computing. 3rd ed. Cambridge Press; 2007. p. 780–5. Ch. Modeling of data.

[20] Quenouille MH. Note on bias in estimation. Biometrika 1956;43:353–60.

[21] Rahmouni A, Yang A, Tempany CMC, Frenkel T, Epstein J, Walsh, et al. Accuracy of in-vivo assessment of prostatic volume by MRI and transrectal ultrasonography. J Comput Assist Tomogr 1992;16(6):935–40.

[22] Rodriguez Jr E, Skarecky D, Narula N, Ahlering TE. Prostate volume estimation using the ellipsoid formula consistently underestimates actual gland size. J Urol 2008;179:501–3.

[23] Saff EB, Kuijlaars ABJ. Distributing many points on a sphere. Math Intell 1997;19:5–11.

[24] Sloane N, Hardin R, Smith W. Tables of spherical codes. Tech. rep. AT&T, Shannon Lab.; 1994, www.research.att.com/njas/.

[25] Terris MK, Stamey TA. Determination of prostate volume by transrectal ultrasound. J Urology 1991;145:984–7.

[26] Tong S, Cardinal HN, McLoughlin RF, Downey DB, Fenster A. Intra- and inter-observer variability and reliability of prostate volume measurement via two-dimensional and three-dimensional ultrasound imaging. Ultrasound Med Biol 1998;24(5):673–81.

[27] Weiss BE, Wein AJ, Malkowicz SB, Guzzo TJ. Comparison of prostate volume measured by transrectal ultrasound and magnetic resonance imaging: is transrectal ultrasound suitable to determine which patients should undergo active surveillance? Urol Oncol Semin Orig Invest 2012;(1078–1439), http://dx.doi.org/10.1016/j.urolonc.2012.03.002 http://www.sciencedirect.com/science/article/pii/S107814391200083X

[28] Wolff JM, Boeckmann W, Mattelaer P, Handt S, Adam G, Jakse G. Determination of prostate gland volume by transrectal ultrasound: correclation with radical prostatectomy specimens. Eur Urol 1995;28:10–2.

48. **A systematic performance evaluation of interactive image segmentation methods based on simulated user interaction.** E. Moschidis and J. Graham, *Proceedings of the IEEE International Symposium on Biomedical Imaging Rotterdam, The Netherlands. April 2010. W. Niessen and E. Meijering, eds. IEEE. pp 928-931.* doi: 10.1109/ISBI.2010.5490139

A SYSTEMATIC PERFORMANCE EVALUATION OF INTERACTIVE IMAGE SEGMENTATION METHODS BASED ON SIMULATED USER INTERACTION

Emmanouil Moschidis, Jim Graham

Imaging Science and Biomedical Engineering, School of Cancer and Enabling Sciences, Stopford Building, The University of Manchester, Oxford Road, Manchester M13 9PT.
Emmanouil.Moschidis@postgrad.manchester.ac.uk, Jim.Graham@manchester.ac.uk

ABSTRACT

In this paper we report on the results of a systematic performance evaluation of three efficient image segmentation algorithms, namely Graph-Cuts, Random-Walker and Grow-Cut. The evaluation focuses on their function as the computational part of an interactive segmentation system. The implications caused by the human involvement in the overall process are avoided by simulating two different patterns of user interaction. The methods are evaluated with respect to accuracy, precision, efficiency and parameter sensitivity on three dimensional medical images. The results provide useful insight regarding the algorithmic performance of the selected techniques and the effect of the identified patterns of user interaction on the segmentation outcome.

Index Terms— Performance Evaluation, Interactive Image Segmentation, Simulated User Interaction, 3-D Medical Images, Graph-Cuts, Random-Walker, Grow-Cut.

1. INTRODUCTION

Human experts, such as radiologists, often segment large three dimensional (3-D) images as part of their clinical routine. While manual segmentation is tedious, subjective and inefficient, automatic methods rarely work as well as is required. Interactive image segmentation methods may be employed as a useful alternative. The recent advent of different efficient computational methods offers a large variety of candidates for the segmentation task.

Interactive segmentation techniques generally consist of three components: the Graphical User Interface (GUI), the interactive part and the computational part [1]. In this study we focus on the computational part, assessing three popular algorithms, namely Graph-Cuts [7], Random-Walker [8] and Grow-Cut [9]. Graph-Cuts is a graph-based method that considers an image as a flow network represented by a graph. The segmentation outcome is obtained when the edges of the graph carry the maximum possible flow. This also provides the cut on the graph, with the minimum cost among all candidates, which separates the seeds provided by the user. Random-Walker is a graph-based method that calculates the probability that a random walk, which initiates from any voxel of an image, will reach a seed specified by the user, given the bias that it cannot cross high image gradients. Finally, Grow-Cut uses iteratively heuristic rules based on cellular automata that define the labels in a voxel neighborhood. When the algorithm converges the final labeling provides the segmentation outcome.

These interactive algorithms share a widely used pattern of interaction, the bush-strokes. In the context of such an interaction, the expert selects certain groups of pixels that belong to a specific class by drawing scribbles with the mouse or a pointing device. In case of binary segmentation the classes are two, foreground and background. In this study we evaluate the three methods with respect to their ability to perform binary segmentation on 3-D medical images, given a controlled user input.

The brush as an interactive tool enables the user to mark large groups of pixels, which provides the algorithms with sufficient input to create plausible segmentations. However, for the purposes of an evaluation study this input should be minimal and controlled. This allows for a better observation of the algorithmic response of the assessed methods by revealing potential failure modes. Therefore, instead of brush-strokes we provide the algorithms with a small number of seeds (individual voxels). Furthermore, in order to focus the evaluation on the computational part of the methods, we provide this input via simulated user interaction. This allows for the exclusion of inconsistent human interaction from the evaluation framework.

The rest of the paper is organized as follows: In section 2 we present the evaluation framework of our study and the metrics associated with it. In section 3 we present our experiments and results, followed by a discussion in section 4.

2. THE EVALUATION FRAMEWORK

The evaluation framework of our study is thoroughly described in [6]. In this section we summarize its main aspects, which are the simulation of the user interaction and the metrics used for assessing the accuracy, the precision and the efficiency of the algorithms.

In our simulation we identify two main patterns of interaction as illustrated in Fig. 1. In the first pattern, the user is selecting seeds that are spread throughout the volume of the anatomy of interest and the background, simulating a general specification of foreground and background; we refer to these as Volume Seeds. In the second, the user selects seeds one voxel away on either side of the ground truth boundary of the anatomy of interest. This is intended to simulate careful local specification of the boundary, and is referred to here as Surface Seeds.

In the context of our study we use two sets of real medical images and two synthetic ones. Specifically these are: five 3-D (83×80×104) brain MR images, in which the task is to segment the brain ventricles, five 3-D (30×40×29) prostate MR images, in which the task is to segment the whole prostate and two simulated 3-D (227×172×11) brain MR images, one T1 and one Inversion

Recovery Turbo Spin Echo (IRTSE), in which the task is to separate the gray matter from the white matter. The ventricle and prostate datasets were chosen to represent a range of difficulty in segmentation. The ventricle images are relatively straightforward to segment, while the prostate images are challenging as they show considerable internal variation in gray level. The synthetic data provide known ground truth for a segmentation task (gray/white matter separation), in which manual delineation is likely to be unreliable, while they demonstrate realistic tissue properties. The surrogate of truth of the real images is provided by manual delineation of the anatomy of interest by one expert. The ground truth of the synthetic data was provided by the forward model of the segmentation algorithm described in [5].

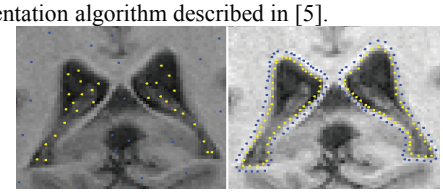


Fig. 1: An illustration of the Volume Seeds (left) and the Surface Seeds selection pattern (right) on a coronal slice of a brain MR image. The foreground and background seeds are colored yellow and blue respectively.

The accuracy of segmentation, with respect to the surrogate ground truth, is evaluated as follows: all the voxels are classified into true and false positives (TP, FP) and true and false negatives (TN, FN). The metric for segmentation accuracy (A) is defined as:

$$A=100\times \frac{|TP|+|TN|}{|TP|+|TN|+|FP|+|FN|}\% \quad (1)$$

The precision of the methods is assessed by measuring the effect of perturbation of the input seeds (see 3.2) on the segmentation result. For every pair of segmentations V_{S1}, V_{S2} the Tanimoto coefficient (Tc)[4] is calculated as:

$$Tc=\frac{V_{S1}\cap V_{S2}}{V_{S1}\cup V_{S2}}=\frac{|TP|}{|TP|+|FP|+|FN|} \quad (2)$$

The efficiency of the interactive segmentation methods is related to the speed and the cognitive load of the overall process. The amount of interaction required for a plausible segmentation, the extent of the precision required during the input provision by the user and the computational speed are the three aspects of a method's efficiency. The first one, is calculated in terms of clicks or input seeds. The fewer input seeds (low cognitive load) that are required for the segmentation task, the higher is the efficiency of the algorithm. With respect to the second aspect, an efficient algorithm does not demand precise pictorial input from the user (low cognitive load). Therefore, a repeatable method is also efficient. Lastly, the computational efficiency is usually reported in seconds.

In section 3.4 we investigate the sensitivity of these methods to operating parameters. The comparisons in sections 3.1-3.3 use optimal parameter settings.

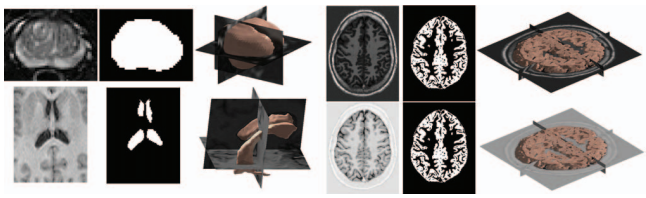


Fig. 2: An axial slice, its ground truth and a 3-D segmentation example of a prostate MR image (top left), a brain MR image (bottom left), a T1 (top right) and an IRTSE (bottom right) synthetic brain MR image. The anatomy to be segmented is the prostate, the brain ventricles and the gray matter, respectively.

3. EXPERIMENTS AND RESULTS

3.1 Accuracy

In order to assess the ability of the algorithms to provide accurate segmentation outcomes, we varied the number of input seeds, both foreground and background, from 1 to 30. For each number of seeds 30 different random initializations were used. The results were averaged over the ventricles and prostate image sets. The first two rows of fig. 3 depict the results of the experiments for the identified patterns of interaction. Overall the graphs show that the accuracy of the segmentation outcome improves as the number of seeds increases. Also, Volume Seeds selection provided better results than Surface Seeds selection. In addition, apart from the prostate experiment in which Random-Walker demonstrated slightly higher accuracy, Graph-Cuts produced the most accurate results, generally reaching its best performance with a smaller number of input seeds than required for the other methods. Visual inspection of the segmentation volumes also revealed that Grow-Cut is prone to leakages when seeds are placed close to partial volumes, which leads to low accuracy due to increased number of false positive. Random-Walker on the other hand demonstrates a resistance to leakages, which generally is a desirable property. However, when seeds are placed in partial volumes, the algorithm tends to segment the partial volumes as a separate tissue, which reduces its accuracy due to increased number of false negatives.

3.2 Precision

In this part of the study we performed a variable seed displacement of 2^i of one initialization per number of seeds, where $i\in[0,8]$; when $i=0$ the seed is only displaced to its immediate neighbor, whereas when $i=8$ it is displaced by 256 positions. All the possible values of i correspond to 9 perturbations and equal number of segmentation outcomes. All the possible pairs (36 in total) of segmentations, were compared with respect to their overlap (eq. 2). For each seed initialization the results were averaged. The last two rows of Fig. 3 depict the results of the experiments. Volume Seeds selection provided better results than Surface Seeds selection in this experiment as well. Graph-Cuts demonstrated high robustness to imprecise input on most datasets. Random-Walker performed slightly better on the prostate dataset and overall was quite robust to imprecise input, although the input required for reaching results comparable to Graph-Cuts was higher. Grow-Cut demonstrated limited robustness to alterations of the seed placement.

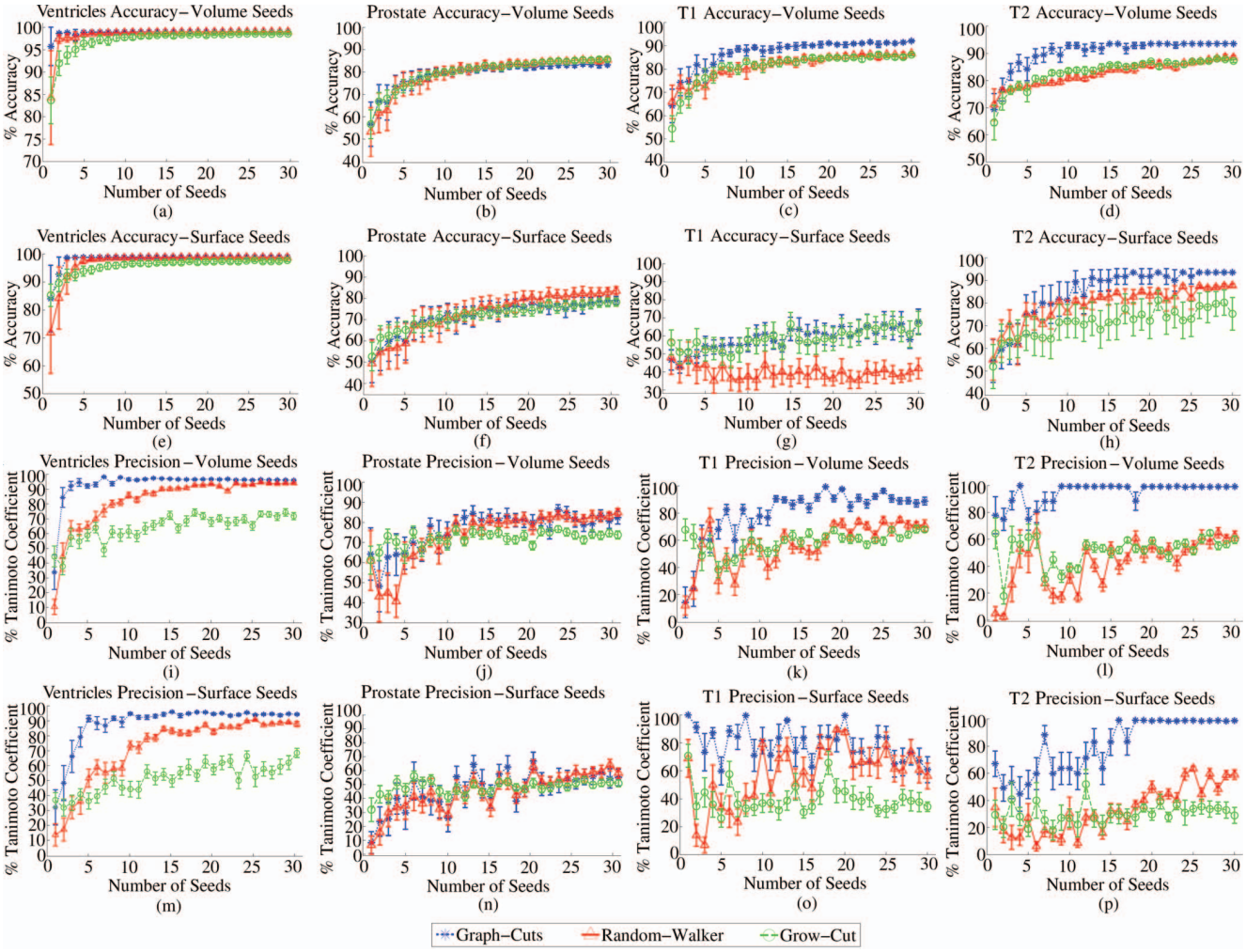


Fig. 3: Accuracy (a-h) and Precision (i-p) for Graph-Cuts, Random-Walker and Grow-Cut as a function of the number of input seeds on Ventricles (1st column), Prostate (2nd column), T1 (3rd column) and T2 (4th column) simulated images for Volume Seeds (1st and 3rd row) and Surface Seeds (2nd and 4th row). The error bars represent the ± 1.96 × standard error of the mean.

3.3 Efficiency

In terms of computational speed, among the implementations that we possess, Graph-Cuts is the most computationally efficient method, whereas Random-Walker is the most computationally expensive. Grow-Cut is slower than the former, but much faster than the latter. Random Walker's and Grow-Cut's shortest execution times were around 10 times and 1.5 times longer than Graph-Cuts' execution time respectively.

In terms of cognitive load, all algorithms can equally provide plausible results even with a few input seeds. Also, Grow-Cut's limited robustness to imprecise input signifies that some seed initializations are better than others. Consequently, the user should spend more time and effort, in order to provide the algorithm with "good" input seeds or to correct inaccurate segmentation suggestions. Therefore, its cognitive load is higher than the other two methods, which demonstrate high robustness to imprecise input.

3.4 Sensitivity to parameters

Since Grow-Cut is a non parametric method, this section concerns only Graph-Cuts and Random-Walker. Graph-Cuts was used without a regional term. Therefore, the only free parameter of both methods is the one that controls their Gaussian weighting function. This function assigns certain weights in the weighted graph based on the intensity differences of neighboring voxels. The weighted graph provides an enhanced data representation; therefore its construction can be seen as an implicit preprocessing step.

In order to assess how σ and β affect the segmentation outcome, we varied them for 30 different random initializations of 30 input seeds. A step of 5 was selected for the alteration of σ, whereas the equivalent step for β was chosen as 50 ($2\sigma^2$). The results (fig. 4) show that Graph-Cuts demonstrates slightly higher sensitivity than Random-Walker to its parameter. Small changes of σ can cause great changes to the segmentation accuracy for Graph-Cuts, whereas Random-Walker responds more smoothly to alterations of

β. Values for σ and β that provide the best possible accuracy are found within the intervals [5,50] and [50,500] respectively. It was also seen that changes of the value of the free parameter cause alteration to the time required for the completion of the segmentation task. However, only in Random-Walker was a strong correlation between β and the time required for the segmentation task observed. Generally in our experiments higher β values caused longer execution times.

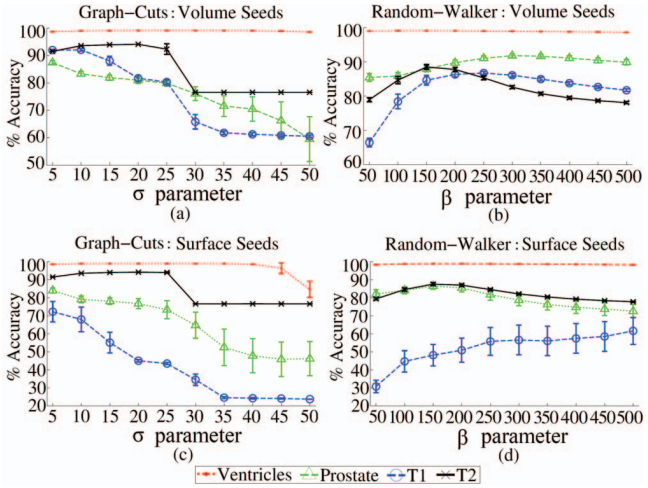


Fig. 4: Accuracy of Graph-Cuts (a,c) and Random-Walker (b,d) segmentations for different values of their free parameter for Volume Seeds (a,b) and Surface Seeds (c,d) on the datasets used in our study.

In Random-Walker's 3-D implementation, we used the preconditioned conjugate gradients method. Since iterative solvers give a solution within a desired limit of accuracy, we observed the effect of the solver's accuracy on the segmentation outcome. For values 10^{-8}-10^{-4} the segmentation outcome remains constant, whereas for values less than 10^{-3} it degrades heavily.

4. DISCUSSION

The experiments performed in our study led to a couple of interesting observations; a major observation is the fact that generally Volume Seeds selection enables the algorithms to provide more accurate and more repeatable segmentations than the Surface Seeds selection. We believe that this happens due to the partial volume effects which are prominent in medical images. If the user selects only voxels close to the boundaries of the anatomy of interest, s/he may select partial volume voxels, which may be recognized as a third separate tissue. Therefore, information from the internal of the organ to be segmented is important for the segmentation task.

In the context of our evaluation, Graph-Cuts proved to be more efficient than its competitors in terms of computational speed and cognitive load. Random-Walker demonstrated robustness against leakages towards partial volumes, which should not be overlooked, since this feature can be valuable in cases where strong edges are not present, a situation which would suggest a failure mode for Graph-Cuts that detects the lowest cost surface in a 3-D fashion. Grow-Cut demonstrated higher sensitivity to imprecise input than the other methods, which leads to higher user cognitive load.

Lastly, the implicit preprocessing step of the graph-based methods, which incorporates the construction of the weighted graph is important for both Graph-Cuts and Random-Walker. Random-Walker is however more robust to the alteration of its parameter.

5. ACKNOWLEDGMENTS

We would like to thank Leo Grady for source code and useful insight regarding the implementation of Random-Walker in 3-D and Neil Thacker for the synthetic images used in our study. The project is funded by the Biotechnology and Biological Sciences Research Council (BBSRC).

6. REFERENCES

[1] S.D. Olabarriaga and A.W.M. Smeulders "Interaction in the segmentation of medical images: A survey", *Medical Image Analysis*, vol.5, no.2, pp. 127-142, 2001.

[2] J.K. Udupa *et al.* "A framework for evaluating image segmentation algorithms", *Computerized Medical Imaging and Graphics*, vol.30, no.2, pp.75-87, 2006.

[3] G. Gerig, M. Jomier and M. Chakos "Valmet: A New Validation Tool for Assessing and Improving 3D Object Segmentation", *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 516-523, 2001.

[4] D.W. Shattuck *et al.* "Online resource for validation of brain segmentation methods", *NeuroImage*, vol.45, no.2, pp.431-439, 2009.

[5] P.A. Bromiley and N.A.Thacker "Multi-dimensional Medical Image Segmentation with Partial Volume and Gradient Modelling", *Annals of the BMVA 2008,* no. 2, pp. 1-22, 2008.

[6] E. Moschidis and J. Graham "Simulation of User Interaction for Performance Evaluation of Interactive Image Segmentation Methods", In *Proceedings of Medical Image Analysis and Understanding 2009 (MIUA 2009)*, pp. 209-213, 2009.

[7] Y.Y.Boykov and M.P. Jolly "Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images", In *Proceedings of the Eighth IEEE International Conference on Computer Vision (ICCV)*, vol. I, pp.105-112, 2001.

[8] L. Grady "Random walks for image segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 28, no. 11, pp. 1768-1783, 2006.

[9] V. Vezhnevets and V. Konouchine ""Grow-Cut" - Interactive Multi-Label N-D Image Segmentation by Cellular Automata", In *Proceedings of the Fifteenth International Conference on Computer Graphics and Applications (Graphicon-2005)*, pp.150-156 , 2005.

49. **Propagating segmentation of a single example to similar images: Differential segmentation of the prostate in 3D MRI**. E. Moschidis and J. Graham, in *'Abdomen and Thoracic Imaging: An Engineering & Clinical Perspective', A.S. El BAz,, L. Saba, J. Suri (eds.), Springer Science +Business Media, New York, 2013, Chapter 25, pp 657-684. ISBN 978-1-4614-8498-1.* doi 10.1007/978-1-4614-8498-1_25

Propagating Segmentation of a Single Example to Similar Images: Differential Segmentation of the Prostate in 3-D MRI

Emmanouil Moschidis and James Graham

Abstract In this chapter, we address the online and real-time segmentation propagation from one example onto similar images. We consider segmentation as a process consisting of two stages: the localization of the anatomy of interest and its boundary delineation. For each stage, we identify and evaluate different potential candidate methods. All methods are assessed regarding their ability to tackle the differential segmentation of the prostate on a dataset of 22 three-dimensional magnetic resonance images of individuals with benign prostatic hyperplasia (BPH). The estimation of the volume of different anatomical zones of the prostate is important for monitoring the progress of the disease. Differential segmentation of the prostate is challenging due to contrast challenges at different locations from surrounding tissues. Also, the high variation of appearance of the prostate across individuals affects the repeatability of frameworks that leverage prior knowledge from one image example. Our observation is that the repeatability is improved, when a two-stage methodology is employed, based on DROP (deformable registration using discrete optimization) registration followed by graph cuts-based segmentation. Our methodology achieves automatically results close to the ground truth, which can serve as an advanced starting point of an interactive process with reduced human operator workload.

Introduction

The main objective of this study is to offer assistance in the context of an interactive framework for building models of three-dimensional (3-D) medical images. In such a framework, a human operator obtains by interaction the surfaces of the anatomy

E. Moschidis (✉) • J. Graham
Centre for Imaging Sciences, Institute of Population Health, The University of Manchester, Stopford Building, Oxford Road, Manchester M13 9PT, UK
e-mail: emmanouil.moschidis@manchester.ac.uk

A.S. El-Baz et al. (eds.), *Abdomen and Thoracic Imaging: An Engineering & Clinical Perspective*, DOI 10.1007/978-1-4614-8498-1_25,
© Springer Science+Business Media New York 2014

of interest, which will be subsequently modeled. These surfaces are extracted from images, which are similar to each other in the sense that they depict the same anatomy of interest and they are acquired by the same imaging modality using the same protocol.

While the extraction of the organ surfaces can be tackled interactively on each image separately, this approach results in an inefficient pipeline with increased workload. As segmentation advances, the number of processed images and therefore the knowledge about the anatomy of interest increases. If this insight is exploited, there is the potential for the design of a framework, which can reduce the amount of user intervention by predicting the segmentation of unseen images. The accuracy of the prediction with respect to the ground truth is associated with the reduction of the amount of intervention that is further required until the desired segmentation is obtained; an accurate prediction requires fewer interactive maneuvers than an inaccurate one. Consequently, the provision of an accurate segmentation prediction reduces the operator's workload.

In this chapter, we address the problem of minimizing the user interaction when similar images are processed, given a *single* previously segmented image as an example of the desired outcome. We tackle this by propagating the segmentation example onto the subsequently processed images. This way the user is freed from the entire process and may intervene only if necessary at a later refinement stage or in case the framework fails to provide satisfactory outcomes. Our main assumption is that the processed images do not exist as a dataset, but they rather appear one at a time (online), as often happens in real life. Therefore, groupwise approaches are not included in the evaluation. Also, since one single image cannot capture the variation of a population, we exclude model-based methods and we restrict our study to data-driven ones. We illustrate the approach using the example of differential segmentation of the prostate in fat-suppressed T2-weighted magnetic resonance (MR) images. The prostate is anatomically divided into several zones, but in MR images, two regions can be identified: the central gland and the peripheral zone. Figure 1 shows a schematic diagram of the relationship between these regions and examples of their appearance in MR images.

We consider segmentation as a process that consists of two distinct tasks: the localization/recognition of the anatomy of interest and its boundary delineation, as suggested in [2]. Consequently, we suggest a two-staged framework that handles these two tasks separately. For each stage we identify and evaluate potential candidate methods against ground truth. Our evaluation can be regarded as an assessment of the extent of the effectiveness of data-driven methods towards the solution of this particular problem.

The results of the experimental work presented in this chapter demonstrate that the suggested framework can provide results close to the ground truth, without any user interaction, serving as an advanced starting point of an interactive process with a small number of further interactive maneuvers. Moreover, we observe that the framework's repeatability improves when the segmentation task is tackled in two distinct stages. Parts of the work that is discussed in this chapter have also appeared in [3–5].

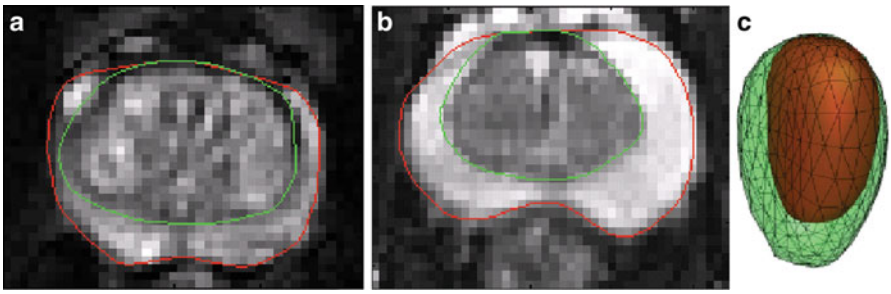


Fig. 1 Two axial midsections of the prostate of fat-suppressed T2 MRI from different individuals (the *red* and *green* contours delineate the peripheral zone and the central gland respectively) (**a**, **b**) and a three-dimensional schematic depiction of the anatomical zones of the prostate (**c**) [1] © 2006 IEEE

Background

Differential Segmentation of the Prostate

The prostate is a gland of the size and shape of a chestnut [6]. It exists only in men and is located immediately below the urinary bladder, where it surrounds a part of the urethra. Its function is the production of a slightly alkaline fluid (40 % of the volume of the semen), which assists towards the neutralization of the acidity of the vaginal tract, thus prolonging the sperm lifespan. In addition, it assists the motility of the sperm cells [6]. Benign prostatic hyperplasia (BPH) is a noncancerous enlargement of the prostate that affects 50 % of men over 60 years old [7]. Approximately a third of them will develop lower urinary tract symptoms, and a quarter of them will need to be operated. For the remaining BPH patients, drug treatment is increasingly utilized [7, 8].

The prostate is anatomically divided into the peripheral (PZ), central (CZ), transitional (TZ), and fibromuscular (FZ) zones. In BPH, the prostatic enlargement is mainly due to the volumetric increase in the TZ. Therefore, the estimation of the TZ volume and the TZ ratio (TZ volume/total prostate volume) is important for monitoring the progress of the disease and the effectiveness of drug treatments [8]. In MRI, two regions are identified: the PZ and the central gland (CG), which includes the other three anatomical zones (Fig. 1). However, in BPH, the TZ is the predominant zone in the CG, due to its expansion, and therefore TZ and CG can be considered as equivalent [1]. Differential segmentation—identifying the surfaces of both the CG and PZ—is challenging. The appearance of the central and peripheral glands varies significantly among individuals (Fig. 1). Furthermore, surrounding tissue (seminal vesicles, blood vessels, the urethra, and the bladder) present contrast challenges at different locations.

While a number of studies recently have addressed segmentation of the prostate in MR images [9–19], segmentation of separate zones has attracted rather less attention [3, 20–24]. These studies fall into two categories: those that employ algorithms trained on multiple image examples in an attempt to model in detail the morphology and shape of the different zones of the prostate and those that tackle the segmentation task interactively. Of the former group, Allen et al. [1] combine a 3-D point distribution model with a Gaussian mixture model, while others apply trained classifiers, such as an evidential C-means classifier in [20] and a Random Forest classifier used within a contextual Markov random field model in [21], or trainable graph-based models [22]. Litjens et al. [18] report that a trained linear discriminant classifier outperforms a multi-atlas approach. Interactive or semi-interactive methods [3, 24] base the segmentation on graphical representations. Given the difficulties in the segmentation task, we have taken the direction in this study of seeking to reduce the workload of interactive methods by leveraging the prior knowledge arising from a single previously segmented example.

Segmentation Propagation from One Image Example

As differential segmentation of the prostate is a challenging task, we assume that expert interaction will be required and investigate methods for minimizing the workload required to achieve a final segmentation. It can often be the case that images are acquired in a sequential (online) manner, rather than being available in a group. In this study we consider the use of a single example as a guide for further segmentations, reducing the level of intervention in subsequent cases. Forward propagation of the template segmentation results in an approximate segmentation for new cases. If this approximate segmentation is accurate, the interactive workload required is correspondingly reduced.

One of the few studies of the literature, which addresses the same problem as we formulate it in this chapter, is the study of Cootes and Taylor in [25]. They adopt a shape representation based on finite element methods (FEMs) [26, 27] when prior knowledge is based on a single image example, whereas they employ an active shape model [28] strategy when multiple image examples are available. When only a single example is available, the allowable shape variation is expressed in terms of the vibrational modes of the FEM model. As further examples are added, this artificial representation of variability is replaced by observed statistical variability.

A single image example is also employed by Rother et al. [29] in a method they call *cosegmentation*. This denotes simultaneous segmentation of the common parts of a pair of images. In order to tackle this task, they employ an algorithm, which matches the appearance histograms of the common parts of the images. At the same time, the imposition of MRF-based spatial constrains guarantees the spatial coherence of the resulting segmented regions. Results are presented for applications such as video tracking and interactive cosegmentation. Similar ideas have been reported in [30–35]. Most of these methods aim at segmenting 2-D colored natural images.

Photographs generally demonstrate good contrast between foreground and background and, due to the variation in color, simple statistics such as color histograms can offer effective discrimination of segments (e.g., [29]). However, histogram-based classification has little to offer in the context of segmentation of greyscale medical images, especially in cases where foreground and background demonstrate similar intensity variations or in case of images with complex appearance.

Active Graph Cuts (AGC) [36] leverages previous segmentations to achieve convergence in a new image in a different way. AGC, when provided with an initial cut, constructs two disjoint subgraphs from the original graph. The initial cut defines the boundary of the two subgraphs. Subsequently, the max-flow/min-cut problem is solved on each of these two subgraphs separately. The combined solution from these subgraphs provides the overall segmentation outcome in the image. We consider and evaluate AGC in our work. To the best of our knowledge, it is the first time that this algorithm is implemented and evaluated independently with respect to its performance on 3-D medical image segmentation tasks.

Atlas-based segmentation is one additional segmentation approach in which often a single image example, termed the *atlas*, is employed as the prior knowledge about the anatomy of interest [37]. An atlas constitutes a complete description of the geometrical constraints and the neighborhood relationships of the anatomy of interest and is often created by manual segmentation of one image. The segmentation of subsequent images is obtained via registration of the processed image and the atlas. Registration constitutes a procedure, which establishes a point-to-point correspondence between two images [38]. When deformable registration is employed for achieving the dense correspondence of the two images, the atlas is deformed and its labels are mapped onto the processed image, also termed *reference image* or *target image*. This process is often referred to as *warping*, *fusion,* or *matching* [38].

One issue associated with frameworks leveraging prior knowledge is the effect of the latter on their performance. If the sample that encapsulates the prior knowledge is representative of the processed population, good results are obtained. However, in the case that it consists of one image, as in this study, the results decline drastically if this image is an outlier with respect to the population. This in turn reduces the framework's repeatability. This is an issue which has attracted considerable attention in the context of atlas-based segmentation [37]. Solutions towards this problem typically involve the combination of multiple atlases into a mean atlas or alternatively the selection of a single atlas that demonstrates a high degree of similarity with the processed image from a group of atlases. A more recent approach is the multi-atlas label fusion [37]. In the context of this strategy, the segmentation suggestions from multiple atlases onto the target image are utilized as individual classifiers, which are combined via a voting scheme. These approaches improve the repeatability of atlas-based methods. However, they all employ multiple atlases. In our study, we are restricted to utilize one single example as prior knowledge. Therefore, the repeatability issue remains.

In the work presented in this chapter, we follow an approach that is driven by registration, similarly to atlas-based segmentation. However, in order to improve its repeatability, we employ a two-staged strategy, as outlined above. Each of the stages, localization/recognition of the anatomy of interest and delineation of its boundary, is tackled separately. In the first stage, the localization task is tackled via registration; in the second stage, a semi-automatic refinement of the segmentation boundary is realized via graph cuts [39, 40]. We show that adopting this strategy improves the repeatability of the framework (in comparison to the single-stage processing approach) and the sensitivity to unhelpful templates is reduced. While we address this in the context of interactive segmentation, a similar conclusion applies in "automatic" atlas-based segmentation, especially on occasions where a single atlas is employed. In addition, for each stage of our approach we identify and evaluate potential candidate methods against ground truth. Therefore, our study can also serve as a comparative performance evaluation of registration and segmentation strategies. In the next sections, we will discuss further the different components of the suggested framework.

Methods

Figure 2 summarizes our segmentation strategy, its constituent stages, and the operations performed in each of these stages, illustrated here in two-dimensions for the sake of clarity. The segmented example consists of the raw image and a binary mask.

The registration (warping) stage is followed by boundary delineation using graph cuts. As we shall discuss further in the following sections, we employ graph cuts to operate on a zone, which is created via successive erosions and dilations of the warped binary mask produced by the framework's first stage. This operation aims at the refinement of the boundary of the anatomy of interest, in cases where registration has failed to provide an accurate segmentation boundary.

Dataset

We use a dataset consisting of 22 3-D T2 fat-suppressed MR images of the prostate from individuals with BPH. T2 fat-suppressed MRI provides good contrast not only between the prostate and its surrounding tissue but also between the prostatic anatomical zones. The images were acquired using a 1.5 T Philips Gyroscan ACS MR scanner. After their acquisition, all images were manually cropped close to the prostate (Fig. 3). The ground truth for each image is a binary volumetric mask produced after averaging the manual delineation of two radiologists on the cropped images. Prior to the experiments, the intensities of all images were normalized to lie in the bounded interval [0, 255]. Lastly, all images were resampled to allow for an iso-voxel resolution and volumes of equal sizes to be created.

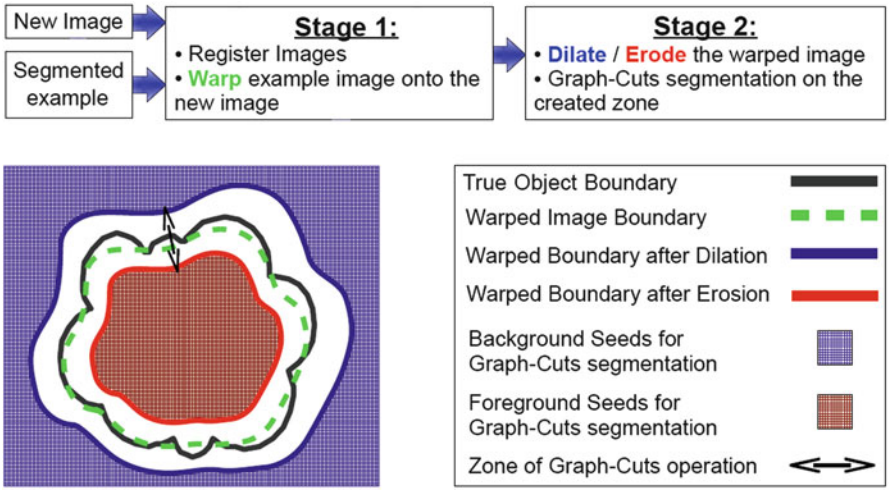


Fig. 2 Overview of the suggested framework

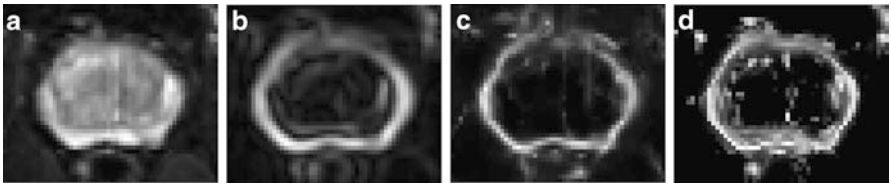


Fig. 3 An axial midsection of a T2 fat-suppressed image of the prostate (**a**), and the raw responses of a Canny (**b**), a Phase Congruency (**c**), and a SUSAN (**d**) feature detector. The settings of the parameters of each detector are outlined in the experiments section of this chapter

Localization of the Anatomy of Interest

The framework's first stage addresses the localization of the anatomy of interest. We evaluate four different registration methods and AGC. In the following paragraphs, we provide the necessary background information with respect to these techniques.

a. *Registration*

Registration is a process that establishes a point-to-point correspondence between two images. The images are considered to be identical, but one of them is treated as being corrupted by spatial distortions; therefore, they cannot be aligned in their current form. The two images are known as *target* and *floating* image. The terms *reference* or *fixed* and *template* or *moving* image respectively are also encountered in the literature [38, 41]. The aim of registration is to compute the exact geometrical transformation

that the floating image needs to suffer, in order to match the target image. For a more comprehensive review of registration and its constituent components as a framework, readers can refer to the relevant literature (e.g., [38, 41, 42]).

In the context of our framework, the example image plays the role of the template image (Fig. 2). The registration scheme computes the spatial transformation, which best aligns this image to the image to be segmented, denoted as New Image in Fig. 2. Subsequently, the spatial transformation is applied to the template image's binary mask. The deformed (warped) binary mask constitutes the segmentation suggestion of the framework's first stage. It also represents the prior knowledge with respect to the segmentation outcome, in the new image's coordinate system.

The registration methods that we assess are the B-Spline-based registration method of Rueckert et al. which employs nonrigid free-form deformations [43]; Thirion's demons registration method [44]; the deformable registration method of Glocker et al. [45], which employs MRFs and discrete optimization [46]; and the groupwise registration method of Cootes et al. [47], utilized here in a pair-wise fashion. The implementations of Kroon and Slump [48] were used for the first two registration methods, whereas the authors' implementations were provided for the latter two methods. In the next paragraphs, we highlight briefly the main components of these registration methods.

The B-Spline method of Rueckert et al. [43] employs a hierarchical transformation model, which combines global and local motion of the anatomy of interest. Global motion is described by an affine transformation, whereas local motion is described by a free-form deformation, which is based on cubic B-Splines. The overall transformation is performed within a multi-resolution setting, which reduces the likelihood of occurrence of a deformation field with invalid topology due to folding of the control points (grid points). The similarity metric employed in this method is normalized mutual information, and the optimization component is based on a gradient descent approach [49].

The underlying concept of the original demons method [44] is that every voxel of the template image is displaced by a local force, which is applied by a demon. The demons of all voxels specify a deformation field, which describes fluidlike free-form deformations; when this field is applied to the template image, the latter deforms so that it matches the reference image. The algorithm operates in an iterative multi-resolution fashion for increased robustness and faster convergence. The original demons algorithm, as presented in [44], is data driven and demonstrates analogies with diffusion models and optical flow equations. Since its original conception, several variants have emerged in the literature [50]. In our study, we use and evaluate the implementation of Kroon and Slump [48], which employs a variant suggested by Vercauteren et al. in [51]. In this variant, the authors follow an optimization approach to demons image registration; more specifically, they employ a gradient descent minimization scheme and operate over a given space of diffeomorphic spatial transformations. Diffeomorphic transformations can be inverted, which is often desirable in image registration. Lastly, Kroon and Slump

employ the joint histogram peaks as the similarity metric of their implementation, which allows for the computation of local image statistics [48].

The registration method of Glocker et al. [45], denoted as DROP (deformable image registration using discrete optimization), follows a discrete approach to deformable image registration. Similarly to the method of Ruckert et al., local motion of the anatomy of interest is modeled by cubic B-Splines. The difference, however, is that image registration is reformulated as a discrete multi-labeling problem and modeled via discrete MRFs. In addition, their optimization scheme is based on a primal-dual algorithm, which circumvents the computation of the derivatives of the objective function [46, 52]. This is due to its discrete nature. In their approach, they also follow a multi-resolution strategy, which is based on a Gaussian pyramid with several levels. Moreover, diffeomorphic transformations are guaranteed through the restriction of the maximum displacement of the control points. The authors' implementation, which is available online [53], features a range of well-known similarity metrics. In this study, the sum of absolute differences was employed.

The registration method of Cootes et al. [47], denoted as GWR, belongs to the groupwise approaches to image registration. Groupwise registration methods aim to establish dense correspondence among a set of images [47], as opposed to pair-wise approaches. In the context of groupwise registration, every image in the set is registered to the mean image, which evolves as the overall process advances. Groupwise registration is often employed in the context of automatic building of shape and appearance models from a group of images, given few annotated examples (e.g., [54]). Conversely, models of shape and appearance can assist registration, when integrated in the process, by imposing certain topological constraints in the spatial deformations. As a result of this integration, Cootes et al. follow a model-fitting approach in their registration method; for each image that is registered to the reference (mean) image, the parameters of the mean texture model are estimated, so that the texture model fits the target image. The overall aim in the process is to minimize the residual errors of the mean texture model with respect to the images of the set. This is achieved via an information theoretic framework, which is based on the minimization of description length (MDL) principle, described in [55]. Piecewise affine transformations are employed for the spatial transformations, which guarantee invertibility of the deformation field. A simple elastic shape model is employed to impose shape constraints to the transformation. Similarly to the previous methods, the registration technique of Cootes et al. follows a multi-resolution approach [47]. In our work, we utilize this method in a pair-wise fashion. We achieve this by employing the reference image to play the role of the mean image. Consequently, the texture and shape model are derived from this single image.

b. *Active Graph Cuts*

AGC [36] exploits an approximate segmentation as initialization, in order to compute the optimal final segmentation outcome via max-flow/min-cut algorithms. The authors pose no restriction on the nature of images that this algorithm may

process and they claim that convergence is achieved even when the approximate segmentation is very different from the desired one. In the context of our study, this approximate solution is provided by the surface of the segmented example image. While it does not perform registration, AGC does provide a method of propagating an initial segmentation and is a likely candidate method for the first stage of our framework. For the needs of our experimental work, we realized our own implementation of the method in MATLAB®. To the best of our knowledge, this is the first independent implementation and evaluation of this technique in the context of 3-D medical image segmentation. In the following paragraphs, we provide further details about the method.

AGC is a graph-based method; therefore, in the context of this technique, an image is represented as a graph. Its initialization, termed "initial cut", is a set of contiguous graph edges, which separates the overall graph into two subgraphs that form two independent flow networks. The vertices adjacent to the initial cut are connected to the source graph terminal. Their t-link weight is equal to the capacity of the adjacent graph edge, which is part of the initial cut. In order to solve the max-flow/min-cut problem, different algorithms may be employed. In [36], the authors suggest a preflow-push [56] approach to tackle this problem on the subgraphs. Preflow-push strategies operate locally and thus flood the network gradually. This in turn generates intermediate cuts as the algorithm progresses. However, in our work, we are rather interested in the final min-cut instead of the intermediate cuts. Therefore, we employ an implementation of a max-flow/min-cut algorithm [57], which follows the augmenting paths approach as described in [58]. The final segmentation outcome is provided by the aggregation of the solutions of the max-flow/min-cut problem on the two subgraphs.

The authors provide no instruction about the positioning of the sink graph terminal on the two subgraphs in [36]. In our work, we observed that different choices may significantly affect the segmentation outcome. For the sake of consistency, with respect to this problem, in all our experiments, the voxels at the image borders were connected to the sink graph terminal of the subgraph that lies outside the initial cut, whereas for the subgraph that is contained by the initial cut, the voxels at a fixed distance, using a distance transform, from the centroid of the initial cut were connected to the sink.

Delineation of the Anatomical Boundary

In the previous sections, we highlighted the candidate methods for the first stage of our framework: localization of the anatomy of interest in the "New Image" of Fig. 2. The second stage of our framework aims for the delineation of an accurate boundary of the anatomy of interest, given the output of the first stage as initialization.

In a previous study [57], we demonstrated that graph cuts segmentation [39, 40] offers significant advantages over several other methods in the context of

interactive segmentation. In the work presented in this chapter, we employ graph cuts in a semi-automated fashion to refine the boundary of the anatomy of interest. As shown in Fig. 2, graph cuts (GC) operates in a zone that surrounds the initial boundary defined by successive erosions and dilations of the initial binary segmentation. Erosion and dilation are morphological operations, which result into contraction and expansion of a binary object respectively [59]. The width of the zone is controlled via a user-defined parameter and depends on the number of erosions and dilations performed on the segmented image example. This is the only interaction that takes place during delineation of the anatomy of interest.

For this stage of our framework, we employ GC with a modified objective function and assess its performance as a means of accurate boundary delineation against the original GC. More specifically, we modify the objective function's boundary term, in order to enable GC to couple with feature detectors. The concept of employing feature detectors to enhance the boundary localization ability of GC is recent. A similar approach to our work [4–6] is followed by Krčah et al. [60], who employ the Hessian matrix as means of increasing the contrast at the boundaries of bones in three-dimensional CT scans. However, the GC boundary term that they employ is not the same as the one that we suggest. In our work, we couple GC with three well-known edge detectors, namely, Canny [61], phase congruency [62], and SUSAN [63]. The three GC variants are denoted as GC + C, GC + PC, and GC + S, respectively. Subsequently, we evaluate the performance of these variants with respect to their ability to provide accurate delineation of the anatomy of interest, as part of our framework's second stage. In the following paragraphs, we provide further details about our modified boundary term and the boundary detectors that we apply.

a. *Coupling Graph Cuts with Feature Detectors*

In interactive GC segmentation [39, 40], an image is represented as a graph. The user selects voxels that belong to the interior and the exterior of the object of interest, referred to as foreground and background seeds, respectively. The optimal foreground/background boundary is then obtained via global minimization of a cost function with min-cut/max-flow algorithms (e.g., [58]). Such a function is usually formulated as

$$E(A) = \lambda \cdot R(A) + B(A) \tag{1}$$

where

$$R(A) = \sum_{p \in P} R_p\left(A_p\right) \tag{2}$$

$$B(A) = \sum_{\{p,q\} \in N} B_{\{p,q\}} \cdot \delta\left(A_p, A_q\right) \tag{3}$$

and

$$\delta(A_p, A_q) = \begin{cases} 1, & \text{if} \quad A_p \neq A_q \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

$R(A)$ and $B(A)$ are the regional and boundary term of the energy function, respectively. The coefficient λ weighs the relative importance between the two terms. N contains all the unordered pairs of neighboring voxels and A is a binary vector, whose components A_p, A_q assign labels to pixels p and q in P, respectively, on a given 2-D or 3-D grid.

The regional term assesses how well the intensity of a pixel p fits a known model of the foreground or the background. These models are either known a priori or estimated by the user input, when this is sufficient. Otherwise, the regional term is weighted low relative to the boundary term or in practice $\lambda = 0$. This approach is followed in [39] as well as in this study. The boundary term encompasses the boundary properties of the configuration A, represented in the weighted graph. Each edge in this graph is usually assigned a high weight if the pixel intensity difference of its adjacent nodes is low and vice versa. The exact value of these weights is calculated with the following Gaussian function [40]:

$$B_{\{p,q\}} = K \cdot \frac{1}{\text{dist}(p,q)} \cdot \exp \frac{-(I_p - I_q)^2}{2\sigma^2} \tag{5}$$

where I_p and I_q are the intensities of two pixels p and q and dist(p,q) the Euclidean distance between them. dist(p,q) is set to 1 in case of equally spaced grids (iso-voxel volumes) when only the immediate neighbors are taken into account. Setting K to 1 leads to a Gaussian function with its peak equal to 1, which is useful for the normalization of the graph weights. σ therefore is the only free parameter, which controls the full width at half maximum of the Gaussian function.

In (5), the effect of the $|I_p - I_q|$ term is to position the min-cut at locations where neighboring voxels demonstrate high-intensity difference, which corresponds to peaks and valleys in the gradient image. This works well for images that demonstrate boundaries with good contrast between foreground and background. However, medical images are often noisy and often demonstrate weak contrast or textured boundaries which are further compromised by partial volume effects. In such challenging boundary conditions, the previous approach can face difficulties in localizing the boundary accurately. To address this problem, we suggest a modification in GC's boundary term, which allows the method to couple with feature detectors. The modified boundary term is described below.

Feature detectors typically produce a response (voxels with high grey-level intensity values) at image locations where, based on local image evidences, the likelihood for the presence of a salient feature is high. In order to allow GC to couple with feature detectors, we modify its weighting function as follows. As we wish the min-cut to occur at maxima (ridges) in the feature output, we replace the $|I_p - I_q|$ term in this function with $(R_p + R_q)/2$, where R_p and R_q is the response of

the feature detector on pixel p and q, respectively. Consequently, we have the modified boundary term:

$$B_{\{p,q\}} = \exp - \left(\varepsilon \cdot \left(R_p + R_q\right)^2\right) \tag{6}$$

where $\varepsilon = 1/8\sigma^2$. Similarly to σ in (5), ε controls the full width at half maximum of the peak of the Gaussian function. Within the following sections, we briefly describe the three well-known feature detectors that we use: Canny [61], phase congruency [62], and SUSAN [63]. Figure 3 also depicts the raw response of these feature detectors on an example T2 fat-suppressed image of the prostate.

b. *Canny Edge Detector*

The Canny edge detector was derived to be an "optimal" edge detector. Its implementation is straightforward in two and three dimensions due to the separability of the Gaussian filter, which is its main computational element [61]. The parameters of the Canny edge detector implementation are the size of the Gaussian filter and its standard deviation. Due to the fact that Canny is a gradient-based edge detector, the strength of its response at a certain image location depends on the magnitude of the gradient at this location.

c. *Feature Detection from Phase Congruency*

In [62], Kovesi employs the Fourier domain of an image to identify image features. His work is based on the local energy model, introduced by Morrone et al. [64] and Morrone and Owens [65], which suggests that humans perceive features at image locations that demonstrate maximal phase congruency in their Fourier components. Phase congruency can be calculated using log Gabor wavelets. A Gabor wavelet is a filter, which is constructed via the modulation of a Gaussian kernel function by a sinusoidal plane wave [66]. The computation of phase congruency is complex as it involves the use of multiple filters at different scales (wavelengths) and orientations. In addition, phase congruency is susceptible to noise. Therefore, a noise reduction strategy is routinely followed prior to its computation [62]. Also, due to the fact that image features are identified in the frequency domain, the strength of the phase congruency response, as a feature detector, does not depend on the gradient of the image. This is a major qualitative difference between feature detection based on phase congruency and gradient-based schemes, such as Canny detection. We computed phase congruency employing the code available from [67].

d. *The SUSAN Feature Detector*

SUSAN is an acronym, which stands for smallest univalue segment assimilating nucleus [63]. The SUSAN edge detection scheme employs a circular mask (sphere in 3-D), which is moved over the processed image. The advantage of circular masks is that they provide isotropic responses. The typical radius of the SUSAN mask is

3.4 voxels, which corresponds to a mask that covers an area of 37 pixels in 2-D and 179 voxels in 3-D. During edge detection, the nucleus of the mask is placed at each voxel of the image. Then, the brightness of each voxel within the mask is compared with the brightness of the nucleus. Those voxels that demonstrate similar brightness to the nucleus (within a user-specified tolerance) belong to the USAN area. The size of the USAN area plays an important role in this feature detection scheme. The USAN area reaches its maximum size when the mask is over image areas that demonstrate relatively uniform voxel intensity, whereas its size gets smaller when the mask approaches an edge or a corner. The SUSAN detector is devised to provide responses, when the USAN area is smaller than a predefined threshold and no response otherwise [63]. The SUSAN edge detection scheme does not need any noise reduction, it does not involve the computation of image derivatives, and it is computationally efficient. In our work, we implemented the SUSAN edge detector in MATLAB®.

Evaluation Framework

We assess the performance of our methodology with a score of classification accuracy (CA), the Tanimoto coefficient (Tc), and the maximum point to surface distance between the segmentation and the ground truth surface (MaxDist).

In order to calculate the CA metric, all the voxels are classified into true- and false-positives (TP, FP) and true- and false-negatives (TN, FN). The CA score is then defined as

$$\mathrm{CA}\ (\%) = 100 \times \frac{|\mathrm{TP}| + |\mathrm{TN}|}{|\mathrm{TP}| + |\mathrm{TN}| + |\mathrm{FP}| + |\mathrm{FN}|}\ \% \tag{7}$$

The Tc score is computed as

$$\mathrm{Tc}\ (\%) = 100 \times \frac{|\mathrm{TP}|}{|\mathrm{TP}| + |\mathrm{FP}| + |\mathrm{FN}|} \% \tag{8}$$

Finally, the MaxDist score is calculated via a 3-D distance transform. The distance is given in voxels, but since we use images with isotropic voxels, the results can be report in millimeters as well.

In the case of CA and Tc scores, accurate segmentation outcomes are represented by large values, whereas in case of the MaxDist score, accurate segmentation outcomes are represented by small values.

Experiments and Results

Localization of the Anatomy of Interest

This section concerns the performance evaluation of the four deformable registration methods discussed in the previous sections and AGC. These methods are employed to provide the localization of the anatomy of interest in a new unseen image, given a single image as an example with respect to the desired segmentation outcome. During the experiments, each image in every dataset was selected once as a template image and its ground truth surface was propagated to the remaining images of the same dataset with each of the assessed methods.

When propagation of the segmentation surface was performed via registration, for each pair of images, the spatial transformation was first computed, and then the template image's ground truth surface was warped onto the target image's space to produce the new segmentation outcome. When propagation of the segmentation surface was performed via AGC, a similar strategy was followed: each image's ground truth surface was set as the initial cut for the remaining images of each dataset, providing thus the required initialization for the AGC algorithm.

In these experiments, all registration methods were employed with their default settings. Figure 4 summarizes the results of the performance evaluation of the four registration methods and AGC. GWR provided results that demonstrated (in most cases) the best mean values of the three employed scores that quantify segmentation accuracy, with DROP providing comparable results. More specifically, GWR demonstrated mean values of the three performance scores of CA = 93.5 %, Tc = 69.6 %, and MaxDist = 5.5 mm for the total prostate segmentation task

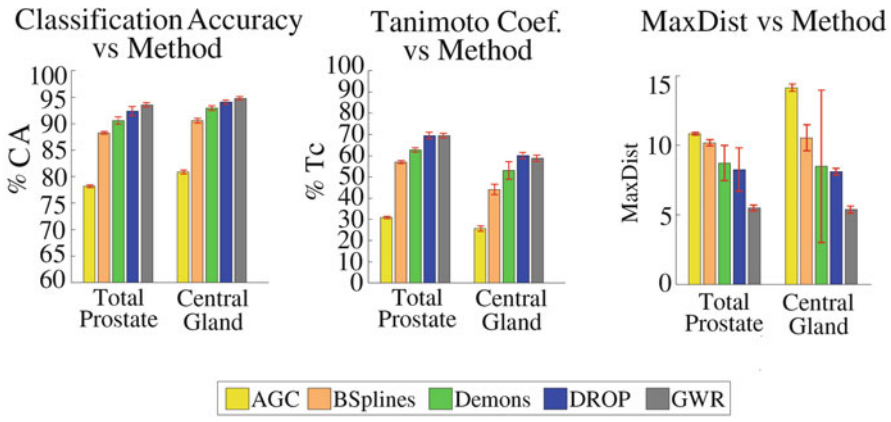


Fig. 4 Summary of the performance of the first stage's candidate methods with respect to the accuracy metrics. The error bars represent the $\pm 1.96 \times$ standard error of the mean

and CA = 94.8 %, Tc = 58.9 %, and MaxDist = 5.4 mm for the central prostatic gland segmentation task. DROP achieved mean values of the three scores of CA = 92.3 %, Tc = 69.6 %, and MaxDist = 8.3 mm for the total prostate segmentation task and CA = 94.1 %, Tc = 60.2 %, and MaxDist = 8.1 mm for the central prostatic gland segmentation task.

In terms of computational efficiency, DROP was by far the most computationally efficient method and GWR the most expensive. For instance, GWR required more than an hour to register two prostate images, whereas DROP performed the same task in few seconds. Computational efficiency is a favorable quality in the context of interactive segmentation systems. Therefore, DROP was adopted as the most appropriate method for our framework's first stage.

In our experiments, AGC consistently failed to produce plausible segmentation outcomes. This is quantitatively depicted in the results presented in Fig. 4. It is conceivable that segmentation of medical images is a challenging task for a max-flow/min-cut segmentation strategy, which is not provided with a good initialization. It is easier to understand this if we recall that AGC employs only the boundary term of a GC objective function. Therefore, in complex images, when wrong initial labeling is provided, the algorithm is susceptible to the detection of undesirable edges, thus providing erroneous segmentation outcomes.

Delineation of the Anatomical Boundary

In this section, we present the results of the evaluation of candidate methods for our framework's second stage, using different GC variants operating in a zone surrounding the boundary defined by the first stage. The voxels that lie within the eroded warped volume are selected as foreground seeds for the GC segmentation, whereas the voxels that lie outside the dilated warped volume are selected as background seeds (see Fig. 2). The zone width is a user-defined parameter that was kept constant for every dataset, to allow for unbiased experimental results. More specifically, this zone was 6 voxels wide (2 voxels outside and 4 voxels inside the boundary suggested by DROP) for segmentation of the prostatic central gland and 9 voxels wide (3 voxels outside and 6 voxels inside the DROP boundary) for the total prostate. We decided to create an asymmetric zone due to the frequent misplacement of the segmentation boundary by the registration stage outside the anatomy of interest. The width of the zone for the central gland is narrower than for the total prostate because this anatomical structure is relatively small. Consequently, a large number of erosions may leave no foreground seeds for GC to operate.

The following parameter settings were used throughout the experiments: when GC was employed with its original boundary term (5), σ was set equal to 1.5; in the case of our modified boundary term (6), ε was set equal to 0.02. The Canny edge detector used a Gaussian kernel of length 7, with standard deviation 0.7. In case of phase congruency, log Gabor filters with 6 different scales and 6 orientations were

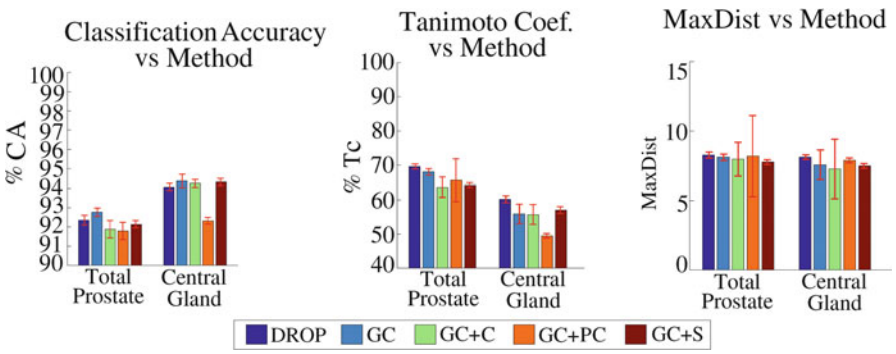


Fig. 5 Summary of the performance of the second stage's candidate methods with respect to the accuracy metrics. The error bars represent the $\pm 1.96 \times$ standard error of the mean

utilized. The minimum wavelength (smallest scale) was set to 5 voxels. Lastly, the SUSAN tolerance threshold was set to 24 (levels of grey). The output of all feature detectors was normalized to lie in the interval [0,255]. All parameters were set to these values via manual experimentation on few images.

The raw response of Canny and SUSAN edge detectors is readily computed in 3-D. However, the code employed for the computation of phase congruency in this study only tackles the task in 2-D. In order to produce an estimate of phase congruency in 3-D, the measure was computed along the three different anatomical planes, and the results were combined by selecting the maximum value from every plane for each voxel. Computation of phase congruency directly in 3-D is obviously preferred; however, such a task is nontrivial. For example, the orientations that need to be considered in 3-D are many more than in 2-D. Rajpoot et al. [68] suggest the use of the monogenic filter to tackle the computation of phase congruency in 3-D. However, when we experimented with their code, their approach produced noisier raw responses than the one we employed.

Figure 5 summarizes the results of the performance evaluation of the candidate methods for the second framework stage. The results of the DROP registration without further segmentation are also included, to allow for direct observation of the effect of the additional processing on the DROP outcome. Overall, the changes in segmentation performance with respect to accuracy due to it are small. The main effect is a slight reduction of the MaxDist error.

In the segmentation of the central gland, GC + S gave less variable results than GC. However, in the total prostate, the use of edge detectors did not seem to provide any advantage over the original GC, possibly due to the already good object/background contrast.

In the case of the central gland segmentation, the paired t-test suggests that there is no significant difference between the performances of GC + S and GC, when the accuracy is assessed with the CA metric. However, this test suggests that GC + S performs significantly better than GC ($p < 0.03$), when the Tc and MaxDist scores

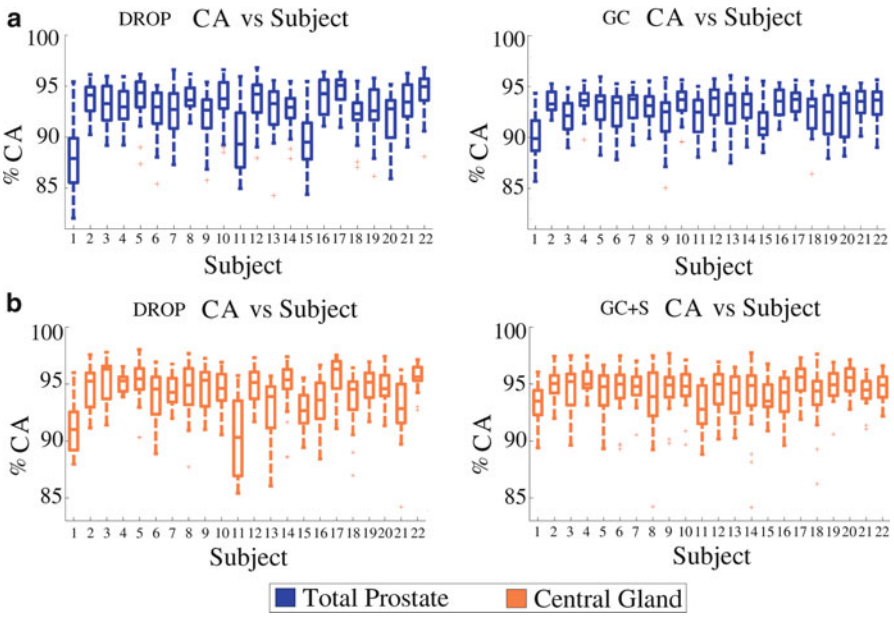


Fig. 6 Pairs of box and whisker plots depicting the repeatability of stage 1 (*left* image) and stage 2 (*right* image) for each segmentation task with respect to the CA score. The whiskers are 1.5 × the interquartile range. Values outside them are considered outliers (*red crosses*) [4] © 2011 IEEE

are used. The Wilcoxon signed-rank test [69] suggests that GC + S performs significantly better than GC ($p < 0.01$), in case of all the employed accuracy scores. The Wilcoxon signed-rank test does not assume that the compared samples are normally distributed. This statistical test may be more appropriate for our assessment, as there is no guarantee that our experimental measurements are normally distributed.

The major advantage of the additional processing step is the increase of the framework's repeatability, compared to the repeatability of the framework when we employ a single-stage processing approach based on DROP registration. Figures 6, 7, and 8 show the variation in segmentation accuracy as different examples are employed as templates, with and without application of the GC stage, using the CA, Tc, and MaxDist metrics, respectively. This improvement is clearest when the framework's performance is measured using the CA and MaxDist metric (Figs. 6 and 8). These figures suggest that in all datasets the second stage offers a reduction of the framework's dependency on the selected template. A qualitative example of the effect of the framework's second stage is also depicted in Fig. 9.

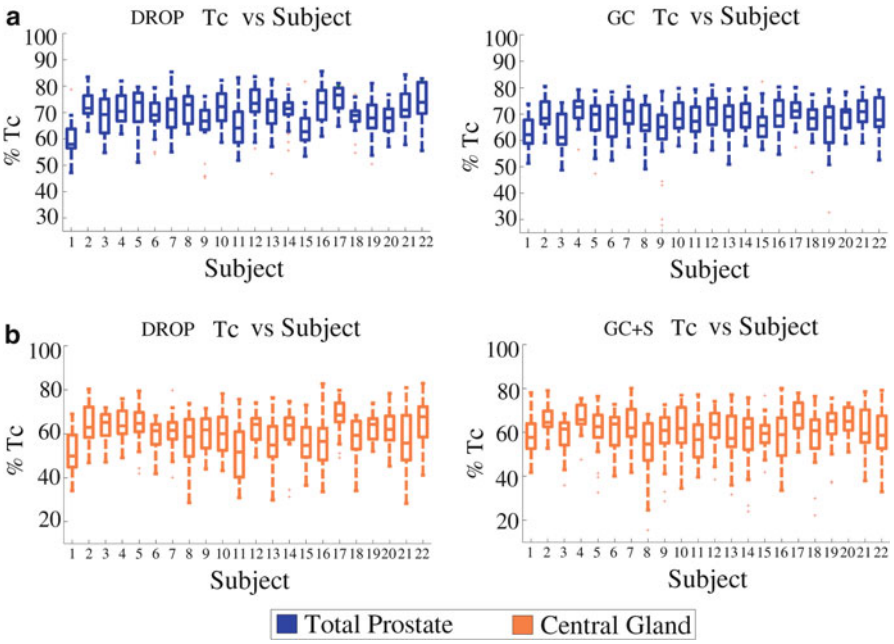


Fig. 7 Pairs of box and whisker plots depicting the repeatability of stage 1 (*left* image) and stage 2 (*right* image) for each segmentation task with respect to the Tc score. The whiskers are 1.5 × the interquartile range. Values outside them are considered outliers (*red crosses*)

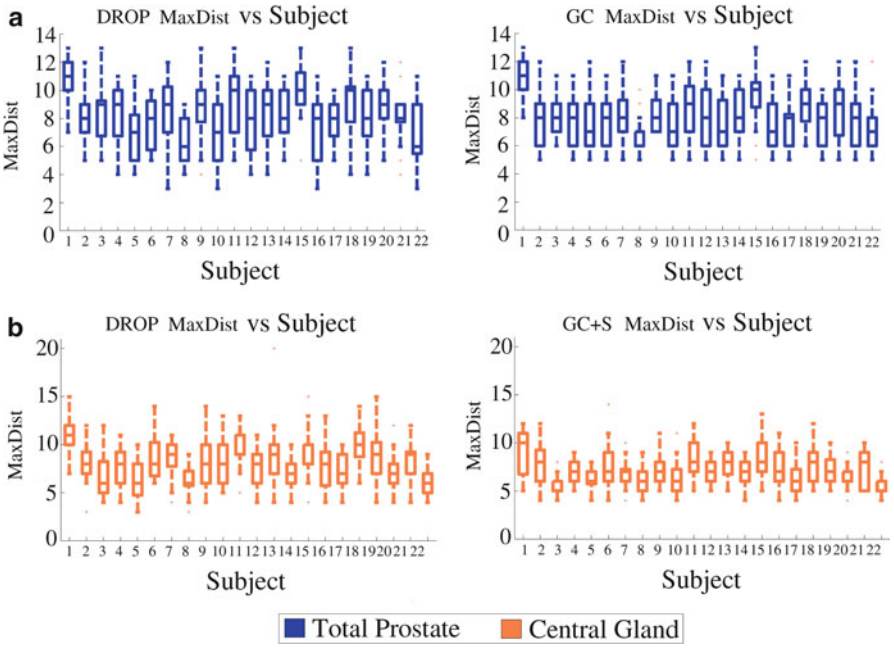


Fig. 8 Pairs of box and whisker plots depicting the repeatability of stage 1 (*left* image) and stage 2 (*right* image) for each segmentation task with respect to the MaxDist score. The whiskers are 1.5 × the interquartile range. Values outside them are considered outliers (*red crosses*)

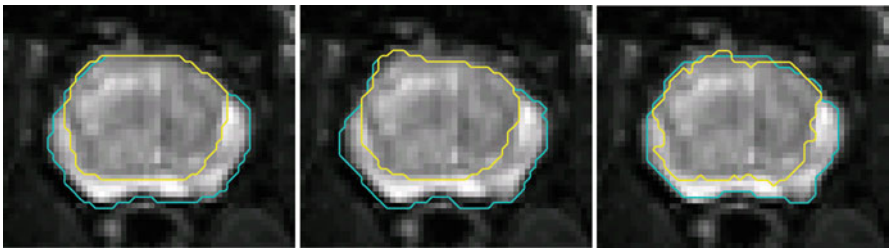


Fig. 9 An example of axial midsection of a prostate, depicting the ground truth (*left*), the segmentation outcome produced by DROP (*middle*), and GC initialized by DROP (*right*). The *yellow* and *cyan* contours delineate the central gland and the total prostate, respectively. The result of the GC+S variant is depicted for the segmentation of the central gland

Summary

In this chapter, we presented the results of a performance evaluation study of candidate methods for an interactive segmentation framework, which leverages prior knowledge from one single image example, in order to minimize the amount of required user intervention when similar images are processed. The suggested framework operates in two stages: localization (registration) followed by delineation (segmentation). The experimental results suggest that this framework can provide results close to the ground truth, without any user interaction, for a challenging segmentation task, when a deformable registration is followed by a graph cuts segmentation. These results can serve as an advanced starting point of an interactive process that can lead to the desired segmentation outcome with a small number of further interactive maneuvers.

Using segmentation of the central gland and total prostate in 3-D MR images as an example application, we show that one of the effects of the additional processing step is the decrease of the MaxDist error. While the CA and Tc scores give overall indications of agreement between the segmentation outcome and the ground truth, they are rather insensitive to local segmentation errors. The MaxDist score gives a handle on local segmentation problems, such as individual surface points being moved away from the true surface. Such cases can result in interactive workload, even if the overall CA and Tc scores are low.

In addition, while the second segmentation stage does not necessarily deliver large improvement over registration-based label propagation in individual cases, we have shown that a two-stage approach improves the framework's sensitivity to the selected template image. While we have addressed this in the context of interactive segmentation, the results of this study can be applied in "automatic" atlas-based segmentation as well, where a single image is often used as a template. Clearly, as online segmentation proceeds, knowledge from increasing numbers of segmented images can be used to inform the interactive process. Ultimately, with sufficient segmented examples, a model can be built. The phase where several

segmented images are available, but not sufficient to build a reliable model is the subject of further development of this work.

In this chapter, we also suggest a modification of the GC boundary term, which allows the method to couple with feature detectors. In our experiments, we assess the performance of GC when coupled with three well-known feature detectors against the original GC approach. Our experimental results suggest that the use of feature detectors may not provide any advantage over the original GC, when images demonstrate good object/background contrast. It may, however, improve the boundary identification ability of GC in challenging cases, where information from gradient is insufficient for the identification of the boundary of the anatomy of interest, such as in the prostate central gland. In this study, we coupled GC with feature detectors that respond to grey-level boundaries. However, detectors that identify textured edges may also be employed with our modified boundary term.

References

1. Allen PD, Graham J, Williamson DC, Hutchinson C (2006) Differential segmentation of the prostate in MR images using combined 3D shape modeling and voxel classification. In: Proceedings of the 3rd IEEE international symposium on biomedical imaging, vol 1–3, pp 410–413
2. Udupa JK, LeBlanc VR, Zhuge Y, Imielinska C, Schmidt H, Currie LM, Hirsch BE, Woodburn J (2006) A framework for evaluating image segmentation algorithms. Comput Med Imaging Graph 30(2):75–87
3. Moschidis E, Graham J (2010) Interactive differential segmentation of the prostate using graph-cuts with a feature detector-based boundary term. In: Proceedings of the 14th annual conference on medical image understanding and analysis, pp 191–195
4. Moschidis E, Graham J (2011) Propagating interactive segmentation of a single 3-D example to similar images: an evaluation study using MR images of the prostate. In: Proceedings of the 8th IEEE international symposium on biomedical imaging: from nano to macro, pp 1472–1475
5. Moschidis E, Graham J (2011) Evaluation of a framework for on-line interactive segmentation of similar 3-D images based on a single example. In: Proceedings of the 15th annual conference on medical image understanding and analysis, pp 287–291
6. Graaff VD (2001) Human anatomy, 6th edn. The McGraw-Hill Companies, Boston
7. Thorpe A, Neal D (2003) Benign prostatic hyperplasia. Lancet 361(9366):1359–1367
8. Tewari A, Shinohara K, Narayan P (1995) Transitional zone volume and transitional zone ratio: predictor of uroflow response to finasteride therapy in benign prostatic hyperplasia patients. Urology 45(2):258–265
9. Birkbeck N, Zhang J, Requardt M, Kiefer B, Gall P, Zhou SK (2012) Region-specific hierarchical segmentation of MR prostate using discriminative learning. In: MICCAI PROMISE12 Challenge
10. Gao Q, Rueckert D, Edwards P (2012) An automatic multi-atlas based prostate segmentation using local appearance-specific atlases and patch-based voxel weighting. In: MICCAI PROMISE12 Challenge
11. Vincent G, Guillard G, Bowes M (2012) Fully automatic segmentation of the prostate using active appearance models. In: MICCAI PROMISE12 Challenge
12. Kirschner M, Jung F, Wesarg S (2012) Automatic prostate segmentation in MR images with a probabilistic active shape model. In: MICCAI PROMISE12 Challenge

13. Maan B, van der Heijden F (2012) Prostate MR image segmentation using 3D active appearance models. In: MICCAI PROMISE12 Challenge
14. Ghose S, Mitra J, Oliver A, Martí R, Lladó X, Freixenet J, Vilanova JC, Sidibé D, Meriaudeau F (2012) A stochastic approach to prostate segmentation in MRI. In: MICCAI PROMISE12 Challenge
15. Malmberg F, Strand R, Kullberg J, Nordenskjöld R, Bengtsson E (2012) Smart paint—a new interactive segmentation method applied to MR prostate segmentation. In: MICCAI PROMISE12 Challenge
16. Toth R, Madabhushi A (2012) Deformable landmark-free active appearance models: application to segmentation of multi-institutional prostate MRI data. In: MICCAI PROMISE12 Challenge
17. Yuan J, Qiu W, Ukwatta E, Rajchl M, Sun Y, Fenster A (2012) An efficient convex optimization approach to 3D prostate MRI segmentation with generic star shape prior. In: MICCAI PROMISE12 Challenge
18. Litjens G, Karssemeijer N, Huisman H (2012) A multi-atlas approach for prostate segmentation in MR images. In: MICCAI PROMISE12 Challenge
19. Ou Y, Doshi J, Erus G, Davatzikos C (2012) Multi-atlas segmentation of the prostate: a zooming process with robust registration and atlas selection. In: MICCAI PROMISE12 Challenge
20. Makni N, Iancu A, Colot O, Puech P, Mordon S, Betrouni N (2011) Zonal segmentation of prostate using multispectral magnetic resonance images. Med Phys 38(11):6093–6105
21. Moschidis E, Graham J (2012) Automatic differential segmentation of the prostate in 3-D MRI using random forest classification and graph-cuts optimization. In: Proceedings of the 9th IEEE international symposium on biomedical imaging: from nano to macro, pp 1727–1730
22. Yin Y, Fotin SV, Periaswamy S, Kunz J, Haldankar H, Muradyan N, Turkbey B, Choyke P (2012) Fully automated 3D prostate central gland segmentation in MR images: a LOGISMOS based approach. In: Proceedings of SPIE medical imaging: image processing, vol 8314, pp 1–9
23. Litjens G, Debats O, van de Ven W, Karssemeijer N, Huisman H (2012) A pattern recognition approach to zonal segmentation of the prostate on MRI. In: Proceedings of the 15th international conference on medical image computing and computer-assisted intervention, vol 15, pp 413–420
24. Egger J, Penzkofer T, Kapur T, Tempany C (2012) Prostate central gland segmentation using a spherical template driven graph approach. In: Proceedings of the 5th image guided therapy workshop
25. Cootes TF, Taylor CJ (1995) Combining point distribution models with shape models based on finite element analysis. Image Vis Comput 13(5):403–410
26. Pentland A, Sclarrof S (1991) Closed-form solutions for physically based shape modeling and recognition. IEEE Trans Pattern Anal Mach Intell 13(7):715–729
27. Sclarrof S, Pentland A (1995) Modal matching for correspondence and recognition. IEEE Trans Pattern Anal Mach Intell 17(6):545–561
28. Cootes TF, Taylor CJ, Cooper DH, Graham J (1995) Active shape models—their training and application. Comput Vis Image Underst 61(1):38–59
29. Rother C, Kolmogorov V, Minka T, Blake A (2006) Cosegmentation of image pairs by histogram matching—incorporating a global constraint into MRFs, In: Proceedings of the 19th IEEE computer society conference on computer vision and pattern recognition, pp 1–8
30. Cheng DS, Figueiredo MAT (2007) Cosegmentation for image sequences. In: Proceedings of the 14th IEEE international conference on image analysis and processing, pp 635–640
31. Boiman O, Irani M (2006) Similarity by composition. In: Proceedings of the 20th annual conference on neural information processing systems, pp 177–184
32. Bagon S, Boiman O, Irani M (2008) What is a good image segment? A unified approach to segment extraction. In: Proceedings of the 10th European conference on computer vision, vol 4, pp 30–44

33. Čech J, Matas J, Perdoch M (2010) Efficient sequential correspondence selection by cosegmentation. IEEE Trans Pattern Anal Mach Intell 32(9):1568–1581
34. Batra D, Kowdle A, Parikh D, Luo J, Chen T (2010) iCoseg: interactive co-segmentation with intelligent scribble guidance. In: Proceedings of the 23rd IEEE computer society conference on computer vision and pattern recognition, pp 3169–3176
35. Vicente S, Rother C, Kolmogorov V (2011) Object cosegmentation. In: Proceedings of the 24th IEEE computer society conference on computer vision and pattern recognition, pp 2217–2224
36. Juan O, Boykov Y (2006) Active graph cuts. In: Proceedings of the 19th IEEE computer society conference on computer vision and pattern recognition, vol 1, pp 1023–1029
37. Rohlfing T, Brandt R, Menzel R, Russako DB, Maurer CR Jr (2005) Quo vadis, atlas-based segmentation? In: Suri J, Wilson DL, Laxminarayan S (eds) The handbook of medical image analysis-volume III: registration models. Kluwer Academic/Plenum Publishers, New York, NY, pp 435–486, Ch. 11
38. Fisher B, Modersitzki J (2008) Ill-posed medicine—an introduction to image registration. Inverse Problems 24:1–16
39. Boykov Y, Jolly M-P (2000) Interactive organ segmentation using graph cuts. In: Proceedings of the 3rd international conference on medical image computing and computer assisted intervention, no. 1935, pp 276–286
40. Boykov Y, Jolly M-P (2001) Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In: Proceedings of the 8th IEEE international conference on computer vision, vol 1, pp 105–112.
41. Yoo TS (2004) Insight into images: principles and practice for segmentation, registration, and image analysis. A. K. Peters Ltd, Wellesley, MA
42. Zitová B, Flusser J (2003) Image registration methods: a survey. Image Vis Comput 21(11): 977–1000
43. Rueckert D, Sonoda LI, Hayes C, Hill DL, Leach MO, Hawkes DJ (1999) Nonrigid registration using free-form deformations: application to breast images. IEEE Trans Med Imaging 18(8):712–721
44. Thirion JP (1998) Image matching as a diffusion process. Med Image Anal 2(3):243–260
45. Glocker B, Komodakis N, Tziritas G, Navab N, Paragios N (2008) Dense image registration through MRFs and efficient linear programming. Med Image Anal 12(6):731–741
46. Komodakis N, Tziritas G, Paragios N (2007) Fast approximately optimal solutions for single and dynamic MRFs. In: Proceedings of the 20th IEEE computer society conference on computer vision and pattern recognition, pp 1–8
47. Cootes TF, Twining CJ, Petrović VS, Babalola KO, Taylor CJ (2010) Computing accurate correspondences across groups of images. IEEE Trans Pattern Anal Mach Intell 32(11): 1994–2005
48. Kroon DJ, Slump CH (2009) MRI modality transformation in demon registration. In: Proceedings of the 6th IEEE international symposium on biomedical imaging, pp 963–966
49. Nocedal J, Wright SJ (2006) Numerical optimization, 2nd edn. Springer, New York
50. Gu X, Pan H, Liang Y, Castillo R, Yang D, Choi D, Castillo E, Majumdar A, Guerrero T, Jiang SB (2010) Implementation and evaluation of various demons deformable image registration algorithms on GPU. Phys Med Biol 55:207–219
51. Vercauteren T, Pennec X, Perchant A, Ayache N (2007) Non-parametric diffeomorphic image registration with the demons algorithm. In: Proceedings of the 10th international conference on medical image computing and computer assisted intervention, vol 4792, pp 319–326
52. Komodakis N, Tziritas G (2007) Approximate labeling via graph-cuts based on linear programming. IEEE Trans Pattern Anal Mach Intell 29(8):1436–1453
53. Glocker B (2012) "mrf-registration.net," [Online]. Available: http://www.mrf-registration.net/
54. Zhang P, Cootes TF (2012) Automatic construction of parts+geometry models for initializing groupwise registration. IEEE Trans Med Imaging 31(2):341–358

55. Twining CJ, Marsland S, Taylor CJ (2004) A unified information-theoretic approach to the correspondence problem in image registration. In: Proceedings of the 17th international conference on pattern recognition, vol 3, pp 704–709
56. Goldberg A, Tarjan R (1988) A new approach to the maximum-flow problem. J ACM 35(4): 921–940
57. Moschidis E, Graham J (2010) A systematic performance evaluation of interactive image segmentation methods based on simulated user interaction. In: Proceedings of the 7th IEEE international symposium on biomedical imaging: from nano to macro, pp 928–931
58. Boykov Y, Kolmogorov V (2004) An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Trans Pattern Anal Mach Intell 26(9): 1124–1137
59. Sonka M, Hlavac V, Boyle R (2008) Image processing analysis and machine vision, 3rd edn. Thomson, Toronto
60. Krčah M, Székely G, Blanc R (2011) Fully automatic and fast segmentation of the femur bone from 3D-CT images with no shape prior. In: Proceedings of the 8th IEEE international symposium on biomedical imaging: from nano to macro, pp 2087–2090
61. Canny J (1986) A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intell 8(6):679–698
62. Kovesi P (1999) Image features from phase congruency. Videre 1(3):1–27
63. Smith S, Brady MJ (1997) SUSAN-A new approach to low level image processing. Int J Comput Vis 23(1):45–78
64. Morrone M, Ross J, Burr D, Owens R (1986) Mach bands are phase dependent. Nature 324(6094):250–253
65. Morrone M, Owens R (1987) Feature detection from local energy. Pattern Recogn Lett 6:303–313
66. Lee TS (1996) Image representation using 2D Gabor wavelets. IEEE Trans Pattern Anal Mach Intell 18(10):959–971
67. Kovesi P (2012) MATLAB and octave functions for computer vision and image processing. [Online]. Available: http://www.csse.uwa.edu.au/~pk/research/matlabfns/
68. Rajpoot K, Grau V, Noble J (2009) Local-phase based 3D boundary detection using monogenic signal and its application to real-time 3-D echocardiography images. In: Proceedings of the 6th IEEE international symposium on biomedical imaging: from nano to macro, pp 783–786
69. Bland M (2000) An introduction to medical statistics, 3rd edn. Oxford University Press, Oxford

Applications of Image Analysis: Diabetic Neuropathy

50. **Application of model based image interpretation methods to diabetic neuropathy.** M. J. Byrne and J. Graham, *Proceedings of the fourth European Conference on Computer Vision, Cambridge, 1996 (vol. 2). B. Buxton and R. Cipolla (eds.). Lecture Notes in Computer Science 1065, Springer-Verlag, Berlin. pp 272-282*. doi: 10.1007/3-540-61123-1_146

Application of Model Based Image Interpretation Methods to Diabetic Neuropathy

M.J. Byrne and J. Graham

Dept. of Medical Biophysics,
University of Manchester, Oxford Road,
Manchester, M13 9PT, UK.
Email: mjb@sv1.smb.man.ac.uk

Abstract

We present two applications of model based computer vision methods to measurement of image features significant in the diagnosis of diabetic neuropathy. The first involves the location of the boundaries of nerve fascicles in light microscope images. The second involves the segmentation of capillary cell regions using electron microscope images. In each case the boundaries required are of arbitrary shape and characterised by local texture or changes in textured regions. The fascicular boundary is located using an Active Contour Model responding to a texture measure based on edge directionality. A start position for the model is automatically generated. The capillary segmentation is performed using a region–based snake responding to a weighted combination of texture measures followed by a local boundary refinement using dynamic programming. These methods show that application of various types of Active Contour Model, accompanied by appropriate starting cues, or followed by local refinements, can locate robustly positioned and intuitively correct boundaries in these images. The aim of the work is the automation of diagnostic measurements currently performed manually. We discuss the implications of automated analysis for procedures in quantitative histology.

1. INTRODUCTION

There are various symptoms and side–effects of the disease diabetes. Important among these are the effects on the nervous system. The most frequent pattern of involvement of diabetes on the nervous system is a peripheral symmetric neuropathy (non–traumatic disorder of the peripheral nerves) of the lower extremities affecting both motor and sensory functions[14].

The major effect is degeneration of the insulating myelin sheath that surrounds each nerve fibre leading to deterioration in motor and sensory functions. Biopsies are taken from patients and images obtained using light and electron microscopy. Number densities and size distributions of myelineated nerve fibres are obtained from these images. Currently these fibre measurements are made manually[2] which is a time consuming process. We have developed automated methods for nerve fibre detection in both light and electron microscope images. We do not present these methods here, although we make use of the results in section 3.3.

Nerve fibres are grouped together to form fascicles (fig.1).

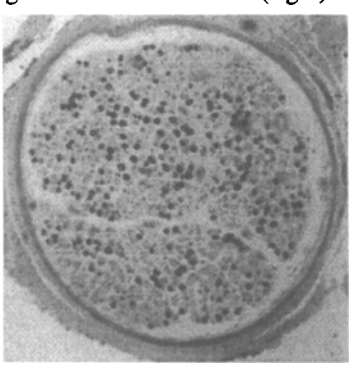


Fig.1. Example of nerve fibre fascicle

It is also necessary to locate the fascicle boundary so that an accurate measure of the fascicular area, and hence accurate fibre number densities can be obtained. In section 3 we describe a method for automatically locating the fascicular boundary using an Active Contour Model responding to an appropriate image measure with a starting cue generated as a result of the automated fibre detection.

Another symptom of interest is microangiopathy (disease of small blood vessels) with effects manifested within the nerve fibres themselves and in capillaries in the endoneurium (the interstitial connective tissue in peripheral nerves separating individual nerve fibres). The two main effects are;

i) A thickening of the basement membrane or an accumulation of basement membrane material in the capillaries (visible as an apparent contraction of the luminal area), and
ii) A proliferation of endothelial cell material together with basement membrane thickening. This manifests itself in arteries, arterioles and occasionally venules.

The structure of two endoenuerial capillaries as they appear in electron micrographs is shown in fig.2. Fig.2a is a healthy example whilst fig.2b displays a degree of diabetic neuropathy and shows these effects.

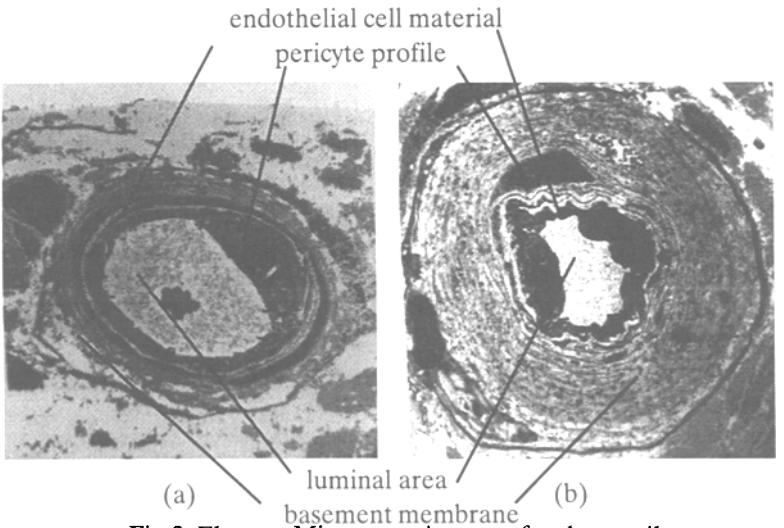


Fig.2. Electron Microscope images of endoneurail capillaries (a) normal and (b) showing neuropathy

Currently the various regions are delineated by hand[13,16]. In section 4 we will describe methods of automated segmentation of the three capillary regions using Active Contour Models followed by local boundary refinement using dynamic programming.

2. MODEL BASED METHODS

Both the fascicle and capillary images are complex. The image evidence defining the required regions involves several parameters which vary from image to image. Often the evidence is poor or missing requiring local interpolation of the data. Location and segmentation of the desired regions in both sets of images therefore require the use of some form of model based method. Statistically based methods such as Point Distribution Models (PDMs) have been sucessfully applied, as part of a constrained image search strategy[5] (Active Shape Models), to the location of poorly defined boundaries in noisy images[6]. PDMs rely on the ability to label a consistent set of landmark points representing the boundary shape in a set of training images in order to construct a statistical model of the expected boundary shape and its allowed degree of variation. In both problems described in this paper it is not possible to identify a set of landmark

points that are consistent from image to image. The lack of correspondence among training images introduces such a large degree of variability between different examples that statistics gathered on shape and appearance impose little constraint on the model. As a result, for the problems presented in this paper, Active Shape Models have not provided a useful approach to image segmentation.

2.1. SNAKES

Active Contour Models (snakes)[10] do not rely on a description of the expected region shape to constrain image search but instead impose internal constraints using a quasi–physical model to control the spacing and curvature of boundary elements. The snake combines these internal constraints with external "forces" derived from the image evidence and iteritively re–positions itself to achieve a minimum energy configuration.

In the absence of any external image forces a snake reaches equilibrium by collapsing either to a single point or line, as dictated by the internal constraints . Furthermore if the snake search is initiated too far away from the desired contour the snake will fail to converge to the correct solution. Cohen[4] added an extra inflationary term causing the snake to behave as a balloon. The inflationary term takes the form of an isotropic pressure potential resulting in an outward pressure force acting along the normal to each snake element. The balloon is inflated, expanding until trapped by strong image evidence (e.g. strong edges) but expanding through weaker evidence. The pressurised snake therefore has the advantage of being able to start its search a large distance away from the desired contour but has the disadvantage that an image feature must produce a strong response in order to overcome the snake's internal pressure force. Hence it is not always possible to segment an image adequately using a pressurised snake.

An adaptation proposed by Ivins and Porrill[9] is the statistical snake, an active region model linking the pressure term to image data within the region enclosed by the snake. An initial seed region is defined either through user interaction or through some form of cue generation. Within this region the means and variances for a suitable set of image measures are determined. These measures should be capable of distinguishing between the region of interest and those around it. The snake is allowed to expand from the boundary of the seed region until the boundary elements encounter pixels whose image measures differ significantly from those in the seed region.

An energy term for the region measure is obtained by multiplying the local change in area for each element by a goodness functional $\mathbf{G(I(x,y))}$ representing the goodness of fit of the region measure at an element of the snake positioned at $\mathbf{(x,y)}$ in an image $\mathbf{I}$. There are various choices for the goodness functional $\mathbf{G()}$. These are as follows for a region having a mean response μ and a range of allowed values within k standard deviations of μ.

Unary Pressure: The goodness functional $\mathbf{G()}$ is set to unity for pixels with image measures within the range specified by the seed region and zero for pixels outside this range.

Binary Pressure: The goodness functional $\mathbf{G()}$ is set to +1 for pixels with image measures within the range specified by the seed region and –1 for those with measures outside the range. When a snake element encounters pixels outside the seed region's range the direction of expansion at that element is reversed.

Linear Pressure. A normalised linear pressure term allows the model to reach equilibrium when its boundary elements encounter pixels at the statistical limits where the goodness functional and hence the pressure force evaluate to zero.

Several image features may be combined by use of a Mahalanobis pressure term.

3. FASCICULAR BOUNDARY LOCATION

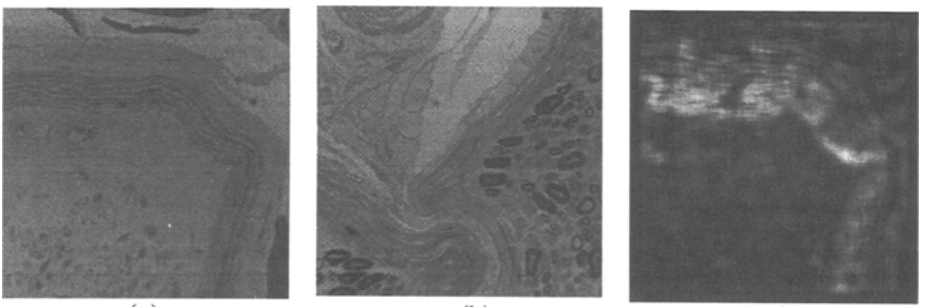
To obtain images with sufficient resolution for fibre detection using light microscopy a magnification of 40 times is required. At this magnification several fields are required to represent an entire fascicle. As a result the fibre detection is carried out on a mosaic

of images connected using cross–correlation. The composite images produced typically have dimensions of 1000–2000 x 1000–2000 pixels.

Application of active contour models to boundary detection requires the choice of a suitable image force and a method of generating a suitable starting position for the model.

3.1 CHOICE OF IMAGE FORCE

The choice of image force is determined by the fascicular boundary structure which consists of a series of closely spaced, fairly parallel lines (fig.3a,b).



(a) (b) (c)

Fig. 3. Examples of fascicle boundary structure(a,b) and result of direction algorithm (c)

However the contrast between these lines and the background is often poor and simple measures based on edge magnitude or contrast fail. To achieve robust detection we have used a texture measure based on edge frequency and edge directionality.

Image Force Algorithm: The algorithm implemented makes uses of the response of a Canny edge detector[3].

The Canny response along the fascicular boundary to consists of a number of parallel edges. Generally the response over the remainder of the image is low, except around nerve fibres, where it shows very little local directionality.

The image feature used to generate the snake's image force is the modal value of the direction of the Canny output within a local neighbourhood. The following algorithm generates an intrinsic image based on this feature.

1. Apply Canny Operator
2. Threshold Canny output to retain only salient edges
3. Quantise edge directions to 16 values
4. Calculate modal direction value within local neighbourhood
5. Retain number of responses at modal direction as pixel output

A local neighbourhood half–width of 10 pixels was empirically found to give the most robust boundary response. An example of the algorithm's output is given in fig.3c.

3.2. GENERATION OF A START POSITION

A starting cue for the snake can be obtained by making use of myelineated nerve fibres detected by our automated method. The fascicle boundary lies in a region surrounding that containing the nerve fibres. In most cases the boundary is not a great distance from the fibre region. The limit of the fibre region is calculated using the distance transform[8] of the image containing the detected nerve fibres.

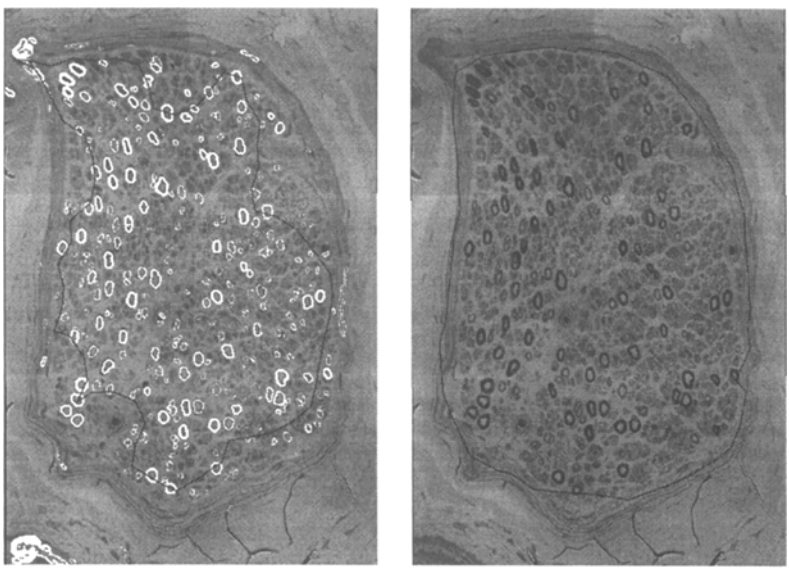
The outer boundary of the fibre region is determined by thresholding the distance transform at a range of increasing values until a single isolated contour is obtained. A single contour is typically obtained at a distance just greater than the maximum

distance between adjacent nerve fibres. This contour, after being smoothed using morphological closing, is used as the starting position of the snake. A degree of smoothing is required since the contour produced by the distance transform is extremely jagged.

3.3. RESULT OF FASCICLE SNAKE

A snake comprised of 61 elements, based on an algorithm by Williams and Shah[17] using the image force described in section.3.1, was applied to a series of sample images. In most examples the starting contour generated by the distance transform was fairly close to the actual boundary position. This allowed the snake to stabilise within 10−15 iterations achieving a close fit to the true boundary.

To assess the robustness of the boundary location with respect to the snake's starting point, an eroded version of the contour produced from the distance transform was obtained. This produced a starting point lying well within the fascicular boundary resulting in the snake having to cross regions of potentially confusing image evidence. Fig.4 shows an example of a snake using a cue eroded by 50 pixels. Fig.4a shows the detected fibres with the eroded position superimposed. Fig.4b shows the final position reached by the snake.



(a) (b)

Fig.4. Result of snake starting from eroded cue

While fig.4 demonstrates the robustness of the texture measure generating the image force, the greater search space and number of iterations required from the use of a distant starting point result in a substantial loss of efficiency. The advantage of having a starting point close to the true boundary position is in the generation of a rapid solution. In some cases exhibiting a high level of neuropathy the limit of the fibre region may be further away from the fascicle boundary requiring the use of an extended search space. The distance between the fascicle boundary and the limit of the fibre region may be a useful diagnostic measure in such cases.

4. SEGMENTATION OF CAPILLARY CELL REGIONS

The method used to segment the areas of interest is determined by the appearance of the three areas. The lumen is generally light in colour and is usually flat or shows only light texture. The endothelial cell material is dark in colour and shows a high degree

of structure. The basement membrane area is lighter in appearance and generally shows less structure with greater directionality than the endothelial cell area. At a coarse scale, across examples, the three regions can be characterised by specific textures. At a finer scale the region boundaries are characterised by a great deal of detailed structure.

4.1. APPROACH

A two stage strategy has been implemented. An initial approximation to the boundary is obtained by application of a region–based (statistical) snake starting from a user supplied seed region. The use of a region–based snake is appropriate since it makes use of the consistent texture within a region to locate the boundary. The only restriction on the snake's starting point is that it lies inside the region to be segmented. The approximate boundary obtained in this way is refined using a higher resolution search method based on dynamic programming which takes advantage of the approximation produced by the region–based snake. The dynamic programming method searches pixel by pixel near the approximate boundary using a measure based on the texture contrast between regions. This measure is capable of greater sensitivity to texture changes when applied close to the true boundary.

4.2. CHOICE OF IMAGE FORCE FOR REGION BASED SNAKE

An image measure is required that can distinguish between the various regions of interest in the capillary. A selection of texture measures was used with discriminant analysis applied to produce a weighted combination capable of producing good classification between either lumen and endothelial cell material or endothelial cell material and basement membrane area. The discriminant analysis is carried out separately for the two boundaries of interest producing a separate set of weights for each. The image measures used were:

- Local average luminance: the average grey level within a local neighbourhood.
- Gradient: [15] a measure of gradient as a function of distance between pixels using the distance dependent texture description function g(d) computed for a user specified distance d.
- Smoothness: A measure of the number of pixels within a local neighbourhood that lie within a specified grey level range of the central pixel value.
- Entropy: to distinguish between regions with little or no texture and regions with some degree of semi–random structure.
- Laws Texture Filters: Six combinations of Laws [11] texture filters were used. These are a well known set of 1D filters which can be combined to represent 2D texture primitives.

These measures were performed on the various regions of interest for a series of training images. A discriminant analysis was carried out on the measures obtained to produce a classification between the lumen and endothelial cell area and then the basement membrane and endothelial cell area. This produced a weighted combination of the region measures. A binary goodness functional based on the results of the discriminant analysis was used as the image force for the region based snake.

4.3. REFINEMENT USING DYNAMIC PROGRAMMING

Dynamic programming as a search technique has been applied to a variety of problems in machine vision[1]. An advantage of dynamic programming is that is always guaranteed to find the optimal path for a given objective function. It also compares well to other techniques such as heuristic search algorithms which depend critically upon the quality of the forward cost estimate. Its advantage as a refinement method is that the cost function is based on local measures, in contrast to the global energy function of the

snake methods. Lutkin[12] has shown that dynamic programming can be an effective method of assessing local image evidence based on an existing model boundary.

In order to refine the result of the region based snake, dynamic programming was applied to a "straightened" image constructed from single pixel spaced normals to the snake's boundary approximation. Each pixel on the straightened boundary corresponds to a node in the search graph[12]. The search then proceeds through the graph finding the best route as dictated by the cost function:

$$cost_i = \alpha\, bound_i + \beta\,(1 - (\theta_i - \theta_{i+1}))$$

where $bound_i$ is the normalised boundary response at node i and θ_i and θ_{i+1} are the angles at nodes i and i+1, and α and β are weighting constants. The cost function is designed to respond to the image evidence whilst maintaining a degree of compatibility with the initial boundary approximation. The compatibility constraint at a transition between nodes is a measure of the angle between the path from one node to the next and the direction of the approximate boundary. Since the dynamic programming is applied to a "straightened" version of the region based boundary approximation the second term in the cost function constrains the refined solution to remain close to this initial approximation.

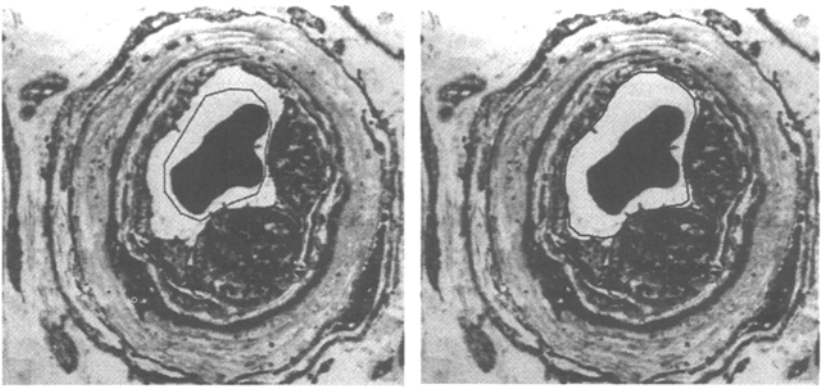
Choice of Image Measure for Dynamic Programming: The image measure used is the difference in texture between two circular regions centred along the normal to the estimated boundary position. As in section 4.2, the texture measure is a weighted combination of the region measures, the weights being determined by discriminant analysis. At a true boundary point this difference between region responses should be maximised.

4.4. RESULTS

For the segmentation of the lumen from the endothelial cell area the starting point for the snake was within the lumen. For the boundary between the basement membrane and the endothelial cell area the region based snake was positioned in basement membrane and allowed to contract inwards towards the boundary. For these experiments the endothelial cell area was not used as a starting point since its structure is less consistent than that of the other two regions.

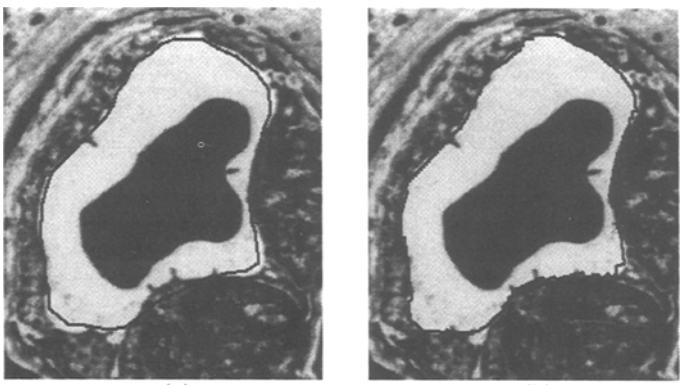
Results for the location of both boundaries are shown below. Fig.5 shows the region based snake applied to segmentation of the lumen endothelial cell area boundary. This boundary is very distinct and the region–based snake has produced a good approximation to the actual boundary position. Fig.6 shows an enlarged section of the capillary shown in fig.5 showing application of the boundary based refinement to the result of the region based snake.

Fig.7 shows the region based snake applied to locating the boundary between the basement membrane and the endothelial cell area. This boundary is less consistent than than the lumen endothelial boundary. In most places the boundary is distinct but in other places the image evidence is poor with the boundary appearing non–existent in some places. Fig.8 shows an enlarged version of fig.7 showing the detailed improvement achieved by the dynamic programming refinement.



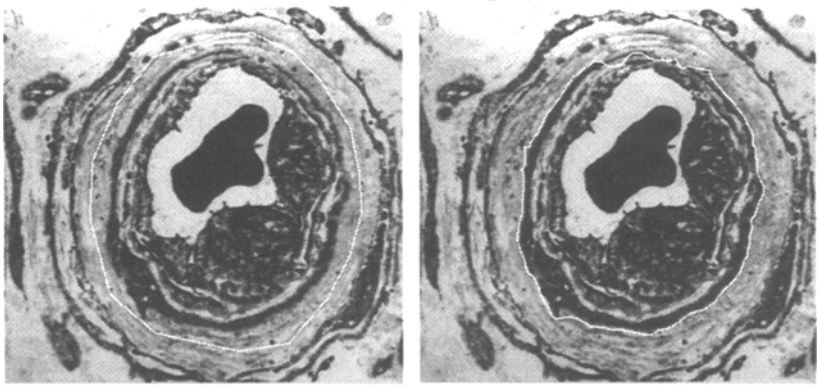
(a) (b)

Fig.5. Segmentation of lumen/endothelial cell area boundary
a)Starting position b)Result of region based snake



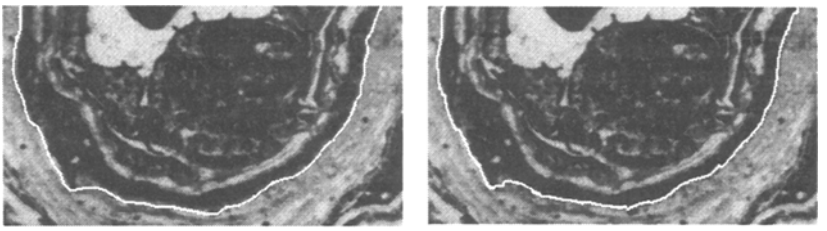
(a) (b)

Fig.6. Comparison of a)region–based snake result and
b)refinement due to dynamic programming



(a) (b)

Fig.7. Application of region based snake to location of basement membrane/endo-boundary. (a) starting point (b) result of region based snake.

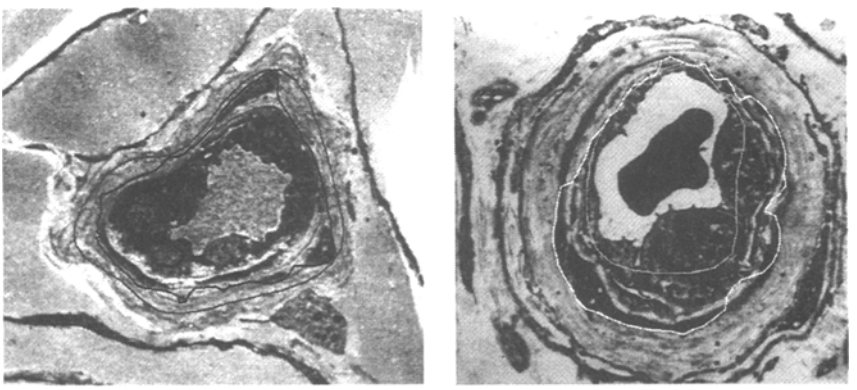


(a) (b)

Fig.8. Comparison of a)region–based snake and b)dynamic programming results.

4.5. APPLICATION OF PAIRED SNAKES

A problem encountered when segmenting the endothelial cell area from the basement membrane is the presence of small regions within the basement membrane showing similar texture to that of the endothelial cell area. Although they are small in comparison to the main body of the endothelial cell area, the region based snake can still be "trapped" by these regions (fig.9a). A snake placed within the endothelial cell area can encounter similar problems due to the inconsistent structure of this area (fig.9b).



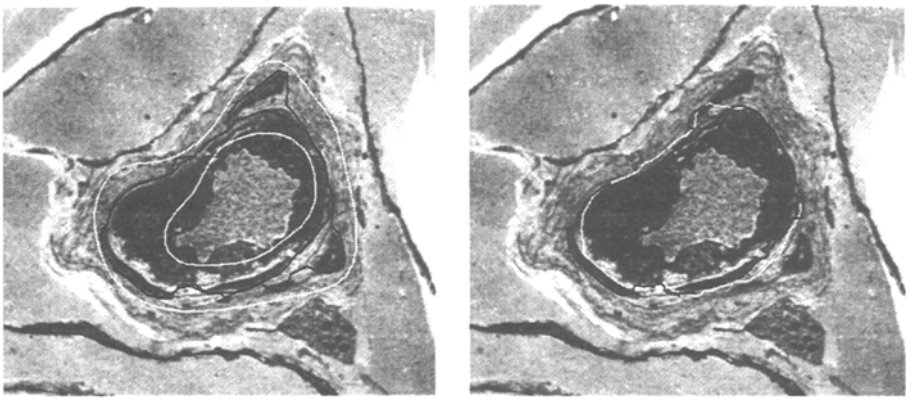
(a) (b)

Fig.9. Region based snake trapped by confusing local evidence
a)Starting from basement membrane b)Starting from endothelial cell area

This is an example of a general problem with snakes and arises from the fact that the only internal constraints on the snake are associated with smoothness. There is no way of preventing a smooth yet incorrect solution arising from confusing local evidence.

Problems with confusing evidence can be overcome by combining two or more independent assessments of the available evidence. We attempt to obtain two independent views of the evidence through the use of a pair of snakes running simultaneously from differing starting positions. In the absence of conflicting evidence both snakes would be expected to arrive at the same answer. Points where there is disagreement suggest the need for further analysis.

The snakes of section 4.4 were augmented by two further snakes initialised within the endothelial cell area, one contracting towards the lumen, the other expanding towards the basement membrane. Fig.10a shows the initial and final positions of a pair of snakes converging on the boundary between the basement membrane and the endothelial cell area. In many places the snakes arrive at an identical position. As in the example shown in fig.9a the outer snake has been trapped by the outlying regions around the endothelial area. However in these places the inner snake has generally arrived at a satisfactory solution



(a) (b)

Fig.10. a)Start(white) and end(black) positions of two region based snakes
b) Average snake position(white) and result(black) of application of dynamic programming

Our initial approach to resolving the conflicting evidence has been to use the average position of the two snakes as the model boundary for dynamic programming refinement. Fig.10b shows that the result of the refinement to be an acceptable representation of the region boundary.

5. CONCLUSIONS AND DISCUSSION

The overall aim of the work presented is the development on an automated system for measurement of diabetic neuropathy encompassing image capture and automated nerve fibre detection as well as the two applications discussed in this paper. The purpose of the fibre detection and fascicular boundary measurements is perform a study of the effects of diabetes on the number densities and size distributions of myelineated nerve fibres. This requires measurement to be made on samples from a large number of patients. Thus a suitably efficient and reliable automated system to replace the need for manual measurements is extremely desirable. Details of the diagnostic utility of the methods will be published elsewhere. The purpose of this paper is to describe the computer vision methods applied.

Fibre detection and fascicular boundary location is a fully automated process. Snake based methods have been shown to successfully locate the fascicular boundary using a starting cue generated from the results of automated fibre detection. The accuracy of boundary location is not sensitive to the effectiveness of this cue, but the fact that the starting point generated is generally close to the final position has a beneficial effect on the method's efficiency.

The capillary segmentation is intended to be as nearly automated as possible and is intended to replace manual delineation of region boundaries. The results shown in this paper have have been based on manually positioned starting points. It may be possible to generate image–based cues for this application as in the case of the fascicular boundary location. The generation of such cues has not been investigated as yet but a possible candidate might use the (usually) uniformly light luminal area.

Technically the achievement of this work has been the segmentation of structurally complex images. A principled approach to characterising texture boundaries has been taken based on trainable image features. These have been shown to be robust when used in conjunction with appropriate forms of Active Contour Model. Two problems which arise from from the use of snakes have proved particularly relevant to this work. Firstly the use of a global energy function means that they do not respond readily to detailed local boundary structure. Secondly the reliance on smoothness as the constraint on snake shape leads to a lack of shape specificity. In our case this means that confusing image evidence can lead to an incorrect solution. The first problem has been addressed through the use of an additional boundary refinement phase which takes locally detailed structure into account by using dynamic programming. The second problem has been addressed through the use of paired snakes to obtain two different views of the image evi-

dence. The use of this strategy in combination with the local refinement method produces adequate results. Further experiments will determine whether greater robustness is required. This could be achieved by more rigourous combination of the evidence. Gunn and Nixon[7], for example, have adopted an approach using paired snakes coupled together to encourage them to converge. Alternatively the dynamic programming refinement could weight the contribution made to its cost function by the approximate boundary according to the level of agreement between the paired snakes. At positions where the paired snakes disagree the local refinement could be allowed more freedom than at positions where the snakes are in agreement.

6. ACKNOWLEDGEMENTS

We would like to thank Dr. R.A. Malik of the Manchester Royal Infirmary for his assistance, and for providing the images used in this work.

7. REFERENCES

1. A.A.Amir, Using Dynamic Programming for Solving Variational Problems in Vision,*IEEE Trans. PAMI, 12(9),* 1990, pp.855–867.
2. S.T. Britland, R.J. Young, A.K. Sharma, B.F. Clarke, Acute and Remitting Painful Diabetic Neuropathy: A Comparison of Peripheral Nerve Fibre Neuropathy, *Pain, 48,* 1992, pp.316–370.
3. J. Canny, A Computational Approach to Edge Detection, *IEEE Trans. PAMI, 8(6),* 1986, pp679–698
4. L.D. Cohen, On Active Contours and Balloons, *CVGIP, Image Understanding, 53(2),* 1991, pp21–218
5. T.F. Cootes, C.J. Taylor, D.H. Cooper, J. Graham, Active Shape Models: Their Training and Application. *Computer Vision and Image Understanding 61,* 1995, pp.38–59.
6. T.F. Cootes, A. Hill, C.J. Taylor, J. Haslam. Use of Active Shape Models for Locating Structure in Medical Images, *Proc. IPMI (13),* 1993, pp33–47.
7. S.R. Gunn, M.S. Nixon, A Model–Based Dual Active Contour, *Proc. BMVC 94,* BMVC Press 1994, pp305–314.
8. RM. Haralick, L.G. Shapiro. Computer and Robot Vision Vol.1 *Addison–Wesley* 1992 pp221–223.
9. J. Ivins, J. Porrill, Statistical Snakes: Active Region Models, *Proc. BMVC 94,* BMVC Press 1994, pp377–386.
10. M. Kass, A. Witkin, D. Terzopoulos. Snakes: Active Contour Models. *Proc. 1st Intl. Conf. Computer Vision.* 1987, pp259–266.
11. K.I. Laws. Rapid Texture Identification. *SPIE Conf. on Image Processing for Missile Guidance.* vol.238 1980, pp376–380.
12. J.P. Lutkin. Interactive Segmentation of Medical Images. *Msc Thesis Manchester University,* 1994.
13. R.A. Malik, S. Tesfaye, S.D. Thompson, A. Veves, A.K. Sharma, A.J.M. Boulton, J.D. Ward. Microangiopathy in Human Diabetic Neuropathy: Relationship Between Capillary Abnormalities and the Severity of Neuropathy. *Diabetologia, 30,* 1989, pp.92–102.
14. R.A. Malik, P.G. Newrick, A.K. Sharma, A. Jennings, A.K. Ah–See, A.J.M. Boulton, J.D. Ward. Endoneurial Localisation of Microvascular Damage in Human Diabetic Neuropathy, *Diabetologia 36,* 1993 pp454–459.
15.R. Sutton, E. Hall. Texture Measures for Automatic Segmentation of Pulmonary Diseases, *IEEE Trans. Comp. c–21.* 1972, pp667–678.
16. R.G. Tilton, P.L. Hoffman, C. Kilo, J.R. Williamson. Pericyte Degeneration and Basement Membrane Thickening in Skeletal Capillaries of Human Diabetes, *Diabetes 30,* 1981, pp.326–334.
17. D.J. Williams, M. Shah. A Fast Algorithm for Active Contours and Curvature Estimation. *CVGIP: Image Undrstanding vol.55(1)* 1992, pp14–26.

51. **Exploiting weak shape constraints to segment capillary images in microangiopathy.** M. Rogers, J. Graham and R.A. Malik, *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI 2000), Pittsburgh, PA, USA, 2000, S.L. Delp, A.M. DiGioia and B. Jaramaz (eds.) (Lecture Notes in Computer Science 1935) Springer, pp 717-726*. doi: 10.1007/978-3-540-40899-4_74

Exploiting Weak Shape Constraints to Segment Capillary Images in Microangiopathy

M. Rogers[1], J. Graham[1] and R.A. Malik[2]

[1]Imaging Science and Biomedical Engineering, University of Manchester, Manchester M13 9PT, U.K.
[2]Department of Medicine, Manchester Royal Infirmary, Manchester, U.K.
mdr@server1.smb.man.ac.uk
www.isbe.man.ac.uk

Abstract. Microangiopathy is one form of pathology associated with peripheral neuropathy in diabetes. Capillaries imaged by electron microscopy show a complex textured appearance, which makes segmentation difficult. Considerable variation occurs among boundaries manually positioned by human experts. Detection of region boundaries using Active Contour Models has proved impractical due to the existence of confusing image evidence in the vicinity of these boundaries. Despite the fact that the shapes have no identifying landmarks, the weak constraints imposed by statistical shape modelling combined with genetic search can provide accurate segmentations.

1. Diabetic Nerve Capillaries

Peripheral neuropathy is an important and debilitating symptom of diabetes. Among the pathological manifestations of neuropathy is microangiopathy (disease of small blood vessels) which affects the capillaries in the endoneurium - the interstitial connective tissue in peripheral nerves separating individual nerve fibres. The two main effects are:

1. a thickening of the basement membrane or an accumulation of basement membrane material in the capillaries causing an apparent contraction of the luminal area, and
2. a proliferation of endothelial cell material together with a thickening of the basement membrane, which manifests itself in arteries, arterioles and occasionally venules.

The aetiology of the condition is unknown. Quantitative studies of the variation in shape and size of the Basement Membrane (BM), Endothelial Cell (EC) and lumen region may shed light on the progression of nerve capillary damage [1]. The structure of two endoeneurial capillaries as they appear in electron micrographs is shown in Fig. 1.

Currently segmentations of these areas for the purpose of measurement are taken by hand [1], which is a time consuming process and restricts the quantity of samples that

S.L. Delp, A.M. DiGioia, and B. Jaramaz (Eds.): MICCAI 2000, LNCS 1935, pp. 717-726, 2000.
© Springer-Verlag Berlin Heidelberg 2000

can be analysed. There is a requirement for an automated approach, both to reduce the labour required and to make the measurements more objective.

The nerve capillary structures have a complex appearance, containing no consistent structural features and wide variation in shape and structure. The image evidence defining the required boundary is extremely variable from image to image and is often highly ambiguous.

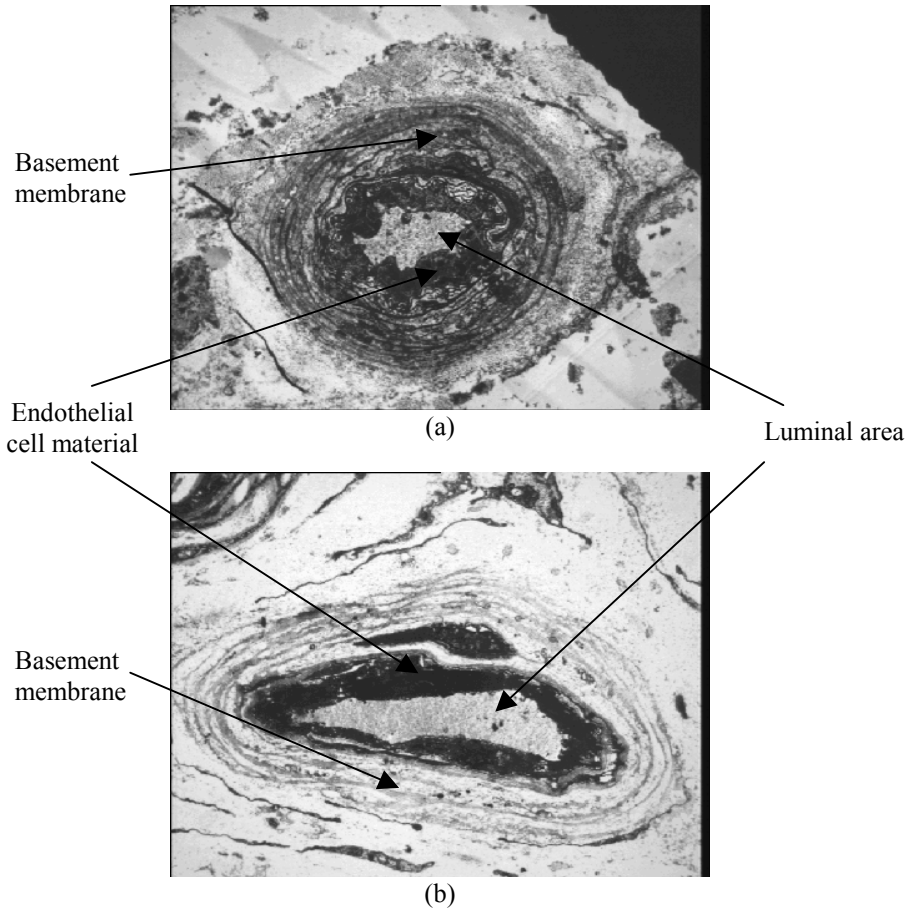


Fig. 1. Two examples of nerve capillary images showing the large variation in appearance and structure. Image (a) exhibits atypical texture around the EC/BM boundary. Image (b) shows an area of locally confusing image evidence at the top of the capillary structure.

2. Segmentation of Capillaries

Byrne and Graham [2] applied an Active Contour Model to this difficult segmentation problem and achieved encouraging results however, the active contours were initiated by manual positioning, and often became 'trapped' on confusing image evidence (see Fig. 1(b)), which provided good but incorrect local boundaries. In this study we seek to obtain more accurate segmentations by:

1. using genetic search to overcome the problem of local maxima in hill climbing methods,
2. constraining the solutions using a model of capillary shape.

There is considerable variability in the shape of capillaries so shape constraints will not be as powerful as those that can be exploited, for example, in the detection of organs in anatomical images. However, capillary shapes are not totally unconstrained either, and conducting a search within the statistical limits imposed by a training set should contribute to better solutions than might be obtained with a totally data-driven approach.

Active Shape Models (ASMs) have been applied successfully to analysis of radiological images [3,4]. The power of this method derives from the statistics of consistent landmark positions on training objects (see section 3 below). The shape representation is then manipulated by a hill climbing search mechanism. Capillaries do not have recognisable landmarks. However, it is still possible to use boundary points to represent the shape, and the resulting descriptions are easily manipulated by Genetic Algorithms (GAs).

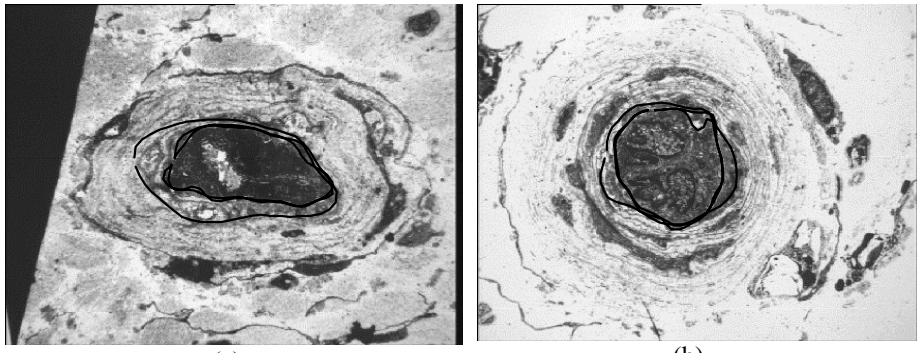
3. Materials and Methods

3.1. Data

Our data set consists of 44 electron microscope images (8-bit grey scale, 768x575 pixels) of nerve capillaries. Significant variation in the quality of the images has been introduced by inconsistencies in the image acquisition process. A set of 10 images to be used as the training set for grey level modelling has been chosen manually. The selection was made based on the perceived quality of the images.

Each capillary image has been annotated three times by experts who marked the boundaries between the various structures within the capillaries. Even for expert human annotation, the boundary positions are difficult to judge and there is considerable variation in the position of manually placed boundaries even for the cleanest images. Table 1 includes the mean point-to-line distances between different expert annotations of the same image for the 10 images in the 'good' image set. The average closest annotated boundary from any sampled point is 7.8 pixels. Many of the

differences between boundaries are small and represent variation in positioning the same perceived boundary. However, some of the larger distances represent different interpretations of the positions of the boundaries between the relevant structures, revealing a genuine ambiguity of interpretation. Fig. 2 shows examples of capillary images with two possible interpretations of the EC/BM boundary position from the three annotations.



(a) (b)

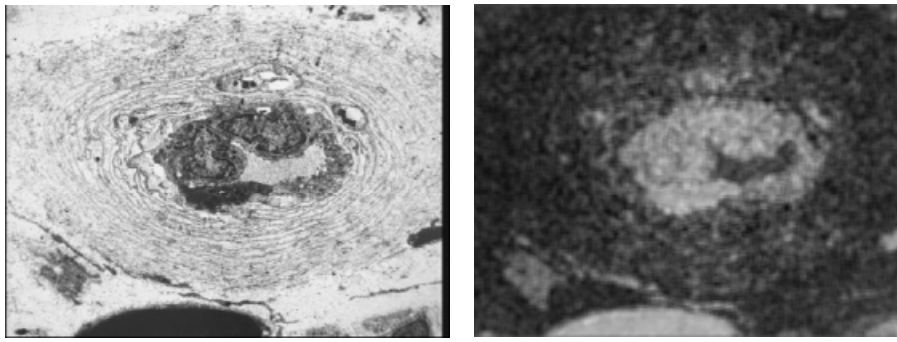
Fig. 2. Two examples of nerve capillaries with multiple expert annotations (marked as solid black lines) of different positions for the Endothelial Cell/Basement Membrane boundary

3.2. Texture Discrimination

As we wish to describe texture boundaries, we transform the grey level images into a texture image. We use the method described by Byrne and Graham [2] in which Laws texture filters [5] are applied to a set of training images to give a set of texture features. These are combined using linear discriminant analysis to provide a texture discrimination function. This function is then used to generate the texture images used in GA search. Fig. 3 shows an example of a texture image generated in this way.

3.3. Active Shape Models

Active Shape Models [6] are generated using a statistical analysis of shape and local grey level appearance over a training set of images. For a detailed description of ASM search see [7]. The training images are labelled with a set of landmark points marking consistent features throughout the set of images. Further evenly spaced model points between chosen features are often required to provide an adequate representation of the shape of a structure. The local grey level appearance is modelled over a patch at each model point.



(a) (b)

Fig. 3. An example capillary image (a) with the corresponding texture image (b) generated using the texture discrimination function.

The variation in shape across the set is described by applying Principal Component Analysis (PCA) to the landmark points, resulting in a Point Distribution Model (PDM) 7. In this way any valid example of the shape being modelled can be approximated using:

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}\mathbf{b} \tag{1}$$

where $\bar{\mathbf{x}}$ is the mean shape vector, $\mathbf{P}$ is a set of orthogonal modes of variation and $\mathbf{b}$ is a vector of shape parameters. Conversely, for any shape $\mathbf{x}$ the parameter vector $\mathbf{b}$ can be calculated:

$$\mathbf{b} = \mathbf{P}^T(\mathbf{x} - \bar{\mathbf{x}}) \tag{2}$$

The model can be used as a representation for image search. By constraining the shape to lie within a specified range of 'allowed' shapes determined by the training set, solutions can be found which are guaranteed to have 'legal' shapes.

In the capillary images there are no consistently identifiable 'landmark' points. To take account of this it was necessary to modify the ASM method. The training points were provided by sampling evenly spaced points from each boundary in the set of expert annotations. The first point on each boundary was defined to lie at one end of the major axis of the best fitting ellipse to the original boundary. Each training example contains 50 boundary points. The shape model training set contains all three annotations from each of the 44 capillary images giving 132 sets of boundary points. The first three modes of variation from the shape model built in this way are shown in Fig 4.

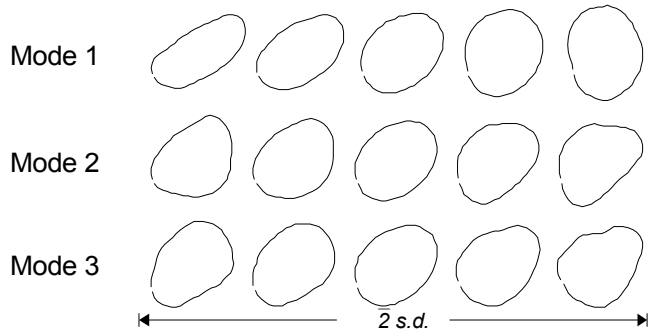


Fig. 4. First three modes of variation of the EC/BM boundary model. The model is built using 50 landmark points sampled from each of 132 expertly annotated boundaries.

The grey level appearance in the capillary images is very variable across the entire set, and this can lead to a model that fits well to flat image data. In order to produce good models capable of discriminating between regions, a subset of 10 images with good discrimination between endothelial area and basement membrane were manually chosen to build the grey level models. An area of 20x3 pixels around each landmark was modelled. Since the model points do not correspond to any particular anatomical feature, there is no reason to believe that the grey level appearance at one model point should be different from that at any other point. Any difference between the grey level models is a feature of the small sample rather than any real property of the data. Therefore a single local model, calculated from all model points was used for all points in the PDM.

A further adjustment to the standard ASM grey level modelling scheme was introduced to utilize our prior knowledge of how the texture images were formed. An ASM was built and an iteration of search was carried out using the original landmarks as a starting point. This has the effect of finding the local best fit of the profile model to the image data, and relocating landmarks closer to areas of high texture gradient. The new point positions after search are then used to build a new grey level model that will find areas of high texture gradient more robustly. This is only appropriate as we want image search to identify areas of high texture gradient as good model matches. The process can be thought of as removing small errors from the annotated landmark positions. Comparative results between a standard ASM with a combined grey level model and an ASM built in this manner are given in section 4.

3.4. Genetic Algorithms

In the standard implementation ASMs perform a hill-climbing search. Small areas of locally difficult texture evidence in the capillary images cause many local maxima to be present in an objective function of the quality of fit of an ASM (see Fig. 5). It is necessary to apply a search method capable of overcoming local maxima to correctly determine the global solution. Genetic Algorithms [8] are a commonly used stochastic

search mechanism which can find good solutions in the presence of numerous local maxima in the objective function. In general, the fittest individuals of a population of candidate boundaries tend to reproduce and survive to the next generation.

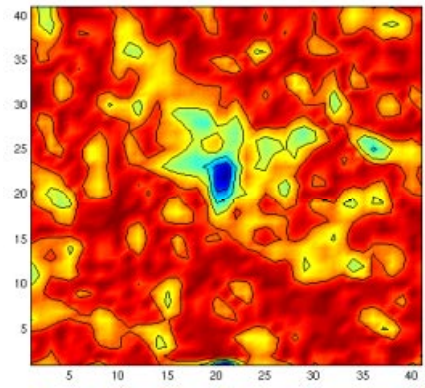


Fig. 5. A plane through the search space defined by the objective function (eqn. 3), produced by varying translation parameters shows many local maxima. Axis values are in pixels.

We use as the objective function:

$$f = \sum_i \exp\left(\frac{-m_i}{2}\right) \quad (3)$$

where m_i is the Mahalanobis distance for the fit of model point i from the mean position. Hill *et al.* [9] show that convergence of GAs is accelerated by performing an iteration of hill-climbing ASM search at the end of each generation. We can think of the GA locating hills in the objective function and the ASM reaching the top of them. This approach has been shown to speed up GA convergence [9].

4. Results

GA searches were run on the set of 10 texture processed nerve capillary images that had been used to build the ASM grey level model in a leave-one-out cross validation, using a population size of 100 individuals and a maximum of 25 generations. The individual with the highest fitness from the final population was taken as the boundary found by the search. Result boundaries were evaluated against the expert annotations available for the corresponding image by calculating the point-to-line distance for each landmark point. Several annotated boundaries exist for each image giving a set of point-to-line distances for each landmark. Table 1 gives details of the point-to-line distances obtained for each of the 10 image searches.

Image	*A* E_{bm}	*B* E_{bm}	*C* E_{bm}
1	5.98	6.76	10.34
2	26.19	25.48	6.71
3	15.28	14.52	8.22
4	6.73	3.84	5.86
5	12.60	12.94	10.18
6	5.97	5.43	6.23
7	6.32	5.65	6.60
8	16.70	14.79	6.91
9	6.42	4.94	6.48
10	129.3	84.62	8.72
Average	23.15	17.90	7.62

Table 1. Search results from leave-one-out GA searches. Column *A* shows results from GA search with a standard ASM with a combined grey level model; column *B* shows results using an ASM with optimised landmarks and a combined grey level model and column *C* shows differences between expert annotations.

	%	E_{bm}	$E_{b\sigma}$
Successful (E_{bm}<35)	72.73	17.14	12.57
Failure (E_{bm}>35)	27.27	62.99	28.53
Total	100	29.64	32.30

Table 2. Robustness results from the set of 44 capillary images, both combined and seperated into successful and unsuccessful search categories. E_{bm} is the mean point-to-line distance of each landmark to the closest single annotation, $E_{b\sigma}$ is the standard deviation of the values

The searches of all the images have found good approximations to the annotated boundaries except for image number 10 which has a far larger error than the rest of the set for all three methods. Of the two GA experiments, the performance of the ASM with an optimized grey level model is slightly better throughout the set with a mean point-to-line error E_{bm} of 17.90 against the standard ASM error of 23.15. Ignoring the results from image 10, average E_{bm} errors become 11.36 and 10.48 for the standard and optimised landmark models respectively, which are comparable to the error in expert annotation. Fig. 6 shows examples of GA search results together with the closest expert annotation.

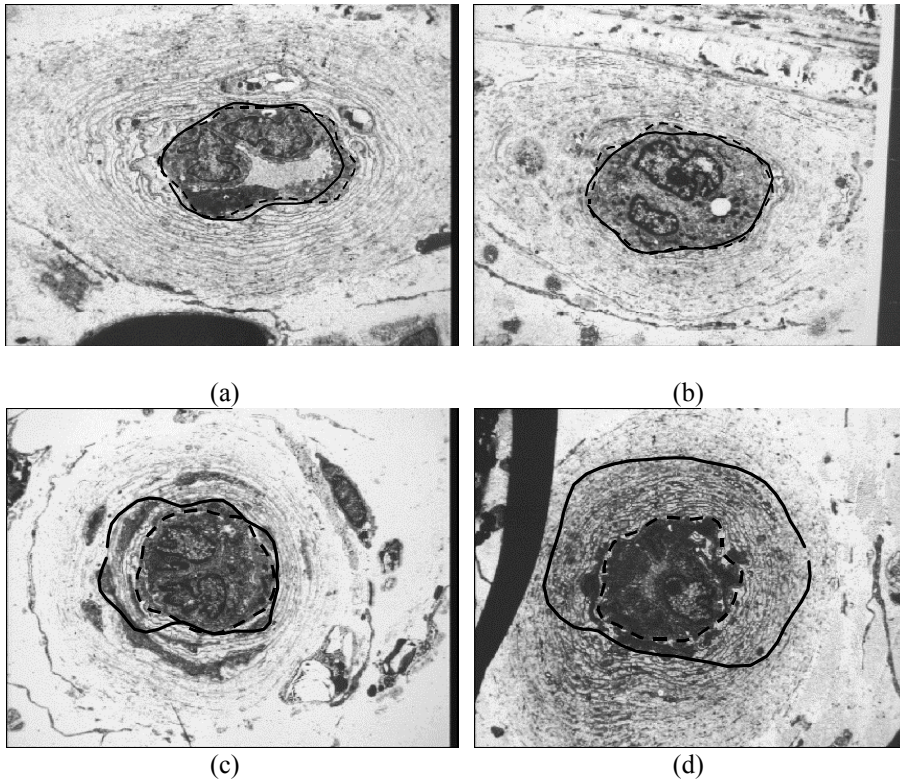


Fig. 6. GA search results: solid lines are automatic segmentations and dashed lines show the closest expert annotation. (a) shows the result for image 1 from Table 1; (b) shows the result for a successful search from the robustness set of 44 images; (c) show a search result that has been influenced by local image evidence; (d) shows the results from image 10 in Table 1 which is an example of a failed search.

Table 2 presents an extension of the evaluation in which leave-one-out tests were carried out on the entire set of 44 usable capillary images. In each case 43 images were used to train the shape model, but the training of the texture model was limited to the 'good' set of 10 images. The search results have been classified into successful searches, with a mean point-to-line distance less than 35 pixels, and unsuccessful searches. The GA search technique is successful in 72.7% of cases. The overall point-to-line error throughout the entire set was 29.6 pixels.

5. Conclusions and Discussion

Actives Shape Models manipulated by GA search have been shown to produce results that have comparable accuracy to human experts. However, the method is inadequately robust at its current state of development; just over 25% of the searches

converge on totally misleading evidence. Inspection of the images on which the search failed suggests that this occurred in cases where the textured appearance of the boundary is significantly different from those in the model, including image 10 in the 'good' set. Fig. 1(a) gives an example of such an image with unusual texture appearance. This indicates that the most likely approach to improving robustness is in development of more adaptive texture models. It is not yet clear whether the very wide variation in texture in some images is a genuine feature of the capillary images or if it has been caused by some error in the imaging process. Analysis of a larger image set will allow investigation of this.

Acknowledgements

We would like to thank Dave Walker of the Department of Medicine, Manchester Royal Infirmary for his assistance in annotating the capillary images used in this work.

References

1. R A Malik, S Tesfaye, S D Thompson, A Veves, A K Sharma, A J M Boulton, J D Ward. Microangiopathy in Human Diabetic Neuropathy: Relationship between Capillary Abnormalities and the Severity of Neuropathy. *Diabetologia*, vol. 30, pp. 92-102, 1989.
2. M J Byrne and J Graham. Application of Model Based Image Interpretation Methods to Diabetic Neuropathy. *Proceedings of European Conference on Computer Vision*, vol. 2, pp. 272-282, 1996.
3. T F Cootes, A Hill, C J Taylor, J Haslam. The Use of Active Shape Models for Locating Structures in Medical Images. *Image and Vision Computing* vol.12, no.6, pp.355-366, 1994.
4. T F Cootes, A Hill, C J Taylor, Medical Image Interpretation Using Active Shape Models: Recent Advances. *14th International Conference on Information Processing in Medical Imaging*, pp. 371-372, 1995.
5. K I Laws. Textured Image Segmentation. *University of Southern California*, 1980.
6. T F Cootes, C J Taylor, D H Cooper and J Graham. Active shape models – their training and application. In *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38-55, 1995.
7. T F Cootes, C J Taylor, D H Cooper and J Graham. Image search using trained flexible shape models. *Advances in Applied Statistics*, pp. 111-139, 1994.
8. L Davis. Genetic Algorithms and Simulated Annealing. *Pitman, London*, 1987.
9. A Hill, T F Cootes and C J Taylor. A Generic System for Image Interpretation Using Flexible Templates. *BMVC*, pp. 276-285, 1992.

52. **Dual-model detection of nerve fibres in corneal confocal microscopy images.** M.A. Dabbah, J. Graham, I Petropoulos, M. Tavakoli and R.A. Malik, *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI) 2010, Beijing China, Part 1, T. Jiang, N. Navab, J.P.W. Pluim, M. A. Viergever, eds. (Lecture Notes in Computer Science 6361, Springer, Heidelberg) pp300-307, 2010*. doi: 10.1007/978-3-642-15705-9_37

Dual-Model Automatic Detection of Nerve-Fibres in Corneal Confocal Microscopy Images

M.A. Dabbah[1], J. Graham[1], I. Petropoulos[2], M. Tavakoli[2], and R.A. Malik[2,⋆]

[1] Imaging Sciences and Biomedical Engineering (ISBE),
The University of Manchester, Oxford Rd,
Manchester, M13 9PT, UK
{m.a.dabbah,jim.graham}@manchester.ac.uk
[2] Cardiovascular Research Group, The University of Manchester,
46 Grafton St., Manchester, M13 9NT, UK
{ioannis.petropoulos,mitra.tavakoli,rayaz.a.malik}@manchester.ac.uk

Abstract. Corneal Confocal Microscopy (CCM) imaging is a non-invasive surrogate of detecting, quantifying and monitoring diabetic peripheral neuropathy. This paper presents an automated method for detecting nerve-fibres from CCM images using a dual-model detection algorithm and compares the performance to well-established texture and feature detection methods. The algorithm comprises two separate models, one for the background and another for the foreground (nerve-fibres), which work interactively. Our evaluation shows significant improvement ($p \approx 0$) in both error rate and signal-to-noise ratio of this model over the competitor methods. The automatic method is also evaluated in comparison with manual ground truth analysis in assessing diabetic neuropathy on the basis of nerve-fibre length, and shows a strong correlation ($r = 0.92$). Both analyses significantly separate diabetic patients from control subjects ($p \approx 0$).

1 Introduction

Diabetic Peripheral Neuropathy (DPN) is one of the most common long-term complications of diabetes. The accurate detection and quantification of DPN are important for defining at-risk patients, anticipating deterioration, and assessing new therapies. Current methods of detecting and quantifying DPN, such as neurophysiology, lack sensitivity, require expert assessment and focus only on large nerve-fibres whereas the earliest signs of neuropathy are likely to be found among small nerve-fibres. On the other hand, small nerve-fibre damage is currently assessed using skin/nerve biopsy, which is highly invasive and is not suitable for repeated investigations.

However, recent research [15,10,8] using Corneal Confocal Microscopy (CCM) suggests that this non-invasive, and hence reiterative, test might be an ideal surrogate endpoint for human diabetic neuropathy. These studies demonstrate that measurements made by CCM accurately quantify corneal nerve fibre morphology. The measurements reflect the severity of DPN and relate to the extent of

⋆ This work is supported by a JDRF scholar grant 17-2008-1031.

T. Jiang et al. (Eds.): MICCAI 2010, Part I, LNCS 6361, pp. 300–307, 2010.
© Springer-Verlag Berlin Heidelberg 2010

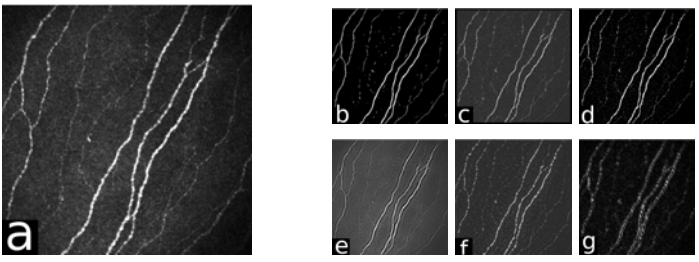


Fig. 1. An illustration of the methods' responses. (a) the CCM image, (b) Dual-model, (c) Linop, (d) Hessian, (e) 2D Gabor, (f) Monogenic and (g) DTCWT.

intra-epidermal nerve-fibre loss seen in skin biopsy. However, the major limitation preventing extension of this technique to wider clinical practice is that analysis of the images using interactive image analysis is highly labour-intensive and requires considerable expertise to quantify nerve-fibre pathology. To be clinically useful as a diagnostic tool, it is essential that the measurements be extracted automatically.

The first critical stage in analysis of CCM images (an example is shown in Figure 1(a)) is the detection of nerve-fibres. This is challenging as the nerve-fibres often show poor contrast in the relatively noisy images. The literature on this topic is not extensive, although the problem has a superficial similarity to other, more widely investigated, applications, such as detection of blood-vessels in retinal images. Ruggeri *et al.* [17] describe a heuristic method that was adapted from retinal analysis. In [2] we conducted a preliminary comparison of methods for contrast enhancement of nerve-fibres, comparing a Gabor wavelet with a well-established line detector.

This paper presents a dual-model algorithm for automatic detection and measurement of nerve-fibres in CCM images. Using a 2D Gabor wavelet and a Gaussian envelope, the dual-model of foreground (nerve-fibres) and background is constructed and applied to the original CCM image. The detection relies on estimating the correct local and dominant orientation of the nerve-fibres. Identifying low-contrast fibrous structures is a commonly encountered problem in several areas of investigation. Examples include mammography, retinopathy, angiography and detection of asbestos fibres. A number of methods have been developed and successfully applied in these applications. We evaluate our dual-model in comparison with some of these methods and with appropriate well-established feature detectors. While our analysis focuses on CCM images, our results suggest that the this may be an appropriate contrast enhancement method in other application domains. In addition to the evaluation of the nerve-fibre detection responses, we have also evaluated the clinical utility of the method by a comparison with manual analysis.

2 Linear-Structure and Feature Detection

A method of linear structure detection (Line Operator - Linop), originally developed for detection of asbestos fibres [4] has also been shown to be effective

in detecting ducts in mammograms [18]. Linop exploits the linear nature of the structures to enhance their contrast by computing the average intensity of pixels lying on a line passing through the reference pixel for multiple orientations and scales. The largest values are chosen to be corresponding to the line, the strength of which is determined by the difference with the average intensity of the similarly oriented square neighbourhood.

In a preliminary study [2], we use the 2D Gabor filter [9] to detect nerve-fibres in CCM images. The filter is a band-pass filter that consists of a sinusoidal plane wave with a certain orientation and frequency, modulated by a Gaussian envelope. This spatial domain enhancement is based on the convolution of the image with the even-symmetric Gabor filter that is tuned to the local nerve-fibre orientation.

Frangi *et al.* [6] used a multiscale decomposition of the Hessian matrix to detect and measure blood vessels in Digital Subtraction Angiography images. They derived a discriminant function based on the eigenvalues and eigenvectors that has maximum response for tube-like structures. The external energy is used to attract the curve towards points which have a high likelihood of lying on a central vessel axis.

The Dual-Tree Complex Wavelet Transform (DTCWT) [11] is an extension of the Discrete Wavelet Transform (DWT), which provides a sparse representation and characterisation of structures and texture of the image at multiresolutions. The DTCWT utilises two DWT decompositions (trees) with specifically selected filters that gives it the properties of approximate shift-invariance and good directionality. The key feature of the DTCWT operation lies in the differences between the filters in the two trees.

The Monogenic signal [5] (a variant of a 2D analytic signal) is an extension of the analytic signal using quaternionic algebra in an attempt to generalise the method so it is capable of analysing intrinsically 2D signals e.g. structures within images. The Monogenic signal is based on the Riesz transform, which is a 2D generalization of the Hilbert transform used in the conventional analytic signal. The Monogenic signal is defined as the combination of the original signal and the Riesz-transformed one in the algebra of quaternions.

3 Dual-Model Nerve-Fibre Detection

In order to quantify the CCM images the nerve-fibres have to be detected. These captured images of nerve-fibre structures could suffer from several types of corruption due to some acquisition conditions, and nerve-fibres may appear faint due to small size or being only partly in the focus plane. Therefore, a nerve-fibre contrast enhancement algorithm is needed to exploit the linear structure of the nerve-fibres and distinguish them from the background noise. All of the methods described in the previous section are capable of providing this enhancement. In the next section we describe our approach.

3.1 Nerve-Fibre Contrast Enhancement

For this purpose the foreground model $\mathcal{M}_{\mathcal{F}}$ is an an even-symmetric and real-valued Gabor [9,3] wavelet and the background model $\mathcal{M}_{\mathcal{B}}$ is a two-dimensional Gaussian envelope.

$$\mathcal{M}_{\mathcal{F}}{}^{(x_\theta,y_\theta)} = \left(\cos\left(\tfrac{2\pi}{\lambda}x_\theta + \phi\right)\right) e^{\left\{-\frac{1}{2}\left(\frac{x_\theta^2}{\sigma_x^2}+\frac{\gamma^2 y_\theta^2}{\sigma_y^2}\right)\right\}} \tag{1}$$

$$\mathcal{M}_{\mathcal{B}}{}^{(x_\theta,y_\theta)} = \alpha e^{\left\{-\frac{1}{2}\left(\frac{x_\theta^2}{\sigma_x^2}+\frac{\gamma^2 y_\theta^2}{\sigma_y^2}\right)\right\}} \tag{2}$$

$$x_\theta = \; x\cos\theta + y\sin\theta \tag{3}$$

$$y_\theta = -x\sin\theta + y\cos\theta \tag{4}$$

The x and y axes of the dual-model coordinate frame x_θ and y_θ are defined by a rotation of θ, which is the dominant orientation of the nerve-fibres in a particular region within the image (see Section 3.2). This dual-model is used to generate the positive response $\mathcal{R}_{\mathcal{P}} = \mathcal{M}_{\mathcal{F}} + \mathcal{M}_{\mathcal{B}}$ and the negative response $\mathcal{R}_{\mathcal{N}} = \mathcal{M}_{\mathcal{F}} - \mathcal{M}_{\mathcal{B}}$ that are applied to the original CCM image and can be represented as in Equations (5) and (6) respectively.

$$\mathcal{R}_{\mathcal{P}}{}^{(x_\theta,y_\theta)} = \left[\cos\left(\tfrac{2\pi}{\lambda}x_\theta + \phi\right) + \alpha\right] e^{\left\{-\frac{1}{2}\left(\frac{x_\theta^2}{\sigma_x^2}+\frac{\gamma^2 y_\theta^2}{\sigma_y^2}\right)\right\}} \tag{5}$$

$$\mathcal{R}_{\mathcal{N}}{}^{(x_\theta,y_\theta)} = \left[\cos\left(\tfrac{2\pi}{\lambda}x_\theta + \phi\right) - \alpha\right] e^{\left\{-\frac{1}{2}\left(\frac{x_\theta^2}{\sigma_x^2}+\frac{\gamma^2 y_\theta^2}{\sigma_y^2}\right)\right\}} \tag{6}$$

The equations of $\mathcal{R}_{\mathcal{P}}$ and $\mathcal{R}_{\mathcal{N}}$ assume that the Gaussian envelope of both responses are identical i.e. they have the same variances $\sigma^2_{(x,y)}$ and the same aspect ratio γ. The magnitude of the Gaussian envelope α defines the threshold in which a nerve-fibre can be distinguished from the background image. The value of α can be set empirically to control sensitivity and accuracy of detection. The wavelength λ defines the frequency band of the information to be detected in the CCM image. Its value might be computed for a subregion within the image that has significant variability of nerve-fibre width. However for simplicity, λ is chosen to be a global estimate of the entire image based on empirical results.

This in turn enhances the nerve-fibres that are oriented in the dominant direction, and decreases anything that is oriented differently by increasing the contrast between the foreground and the noisy background, whilst effectively reducing noise around the nerve-fibre structure as shown in Figure 1(b). This pixel-wise operation adjusts the models to suit the local neighbourhood characteristics of the reference pixel at $f^{(i,j)}$ by modifying the parameters of the foreground and background models. The dot products of the models and the reference pixel's neighbourhood (Equations 7 and 8) are then combined to generate the final enhanced value of this particular reference pixel $g^{(i,j)}$ (Equation 9).

$$\Gamma_p^{(i,j)} = \left\langle f_\omega^{(i,j)}, \mathcal{R}_{\mathcal{P}} \right\rangle \tag{7}$$

$$\Gamma_n^{(i,j)} = \left\langle f_\omega^{(i,j)}, \mathcal{R}_{\mathcal{N}} \right\rangle \tag{8}$$

$$g^{(i,j)} = \frac{\Gamma_p^{(i,j)}}{1 + e^{\left(-2k\Gamma_n^{(i,j)}\right)}} \tag{9}$$

The neighbourhood area of the reference pixel is defined by the width ω. The sharpness of the transition of the enhanced image value at a particular pixel $g^{(i,j)}$ is controlled by k. A larger k amounts to a sharper transition when $\Gamma_n = 0$.

3.2 Nerve-Fibre Orientation Estimation

In CCM images, the nerve-fibres flow in locally consistent orientations everywhere. In addition, there is a global orientation that dominates the general flow. This orientation field describes the coarse structure of nerve-fibres. Using the least mean square (LMS) algorithm [7], the local orientation of the block centred at certain pixel is computed as in [16].

Since the orientations vary at a slow rate, a low-pass Gaussian filter is applied globally in order to further reduce errors at near-nerve-fibre and non-nerve-fibre regions. The LMS produces a stable smooth orientation field in the region of the nerve-fibres; however when applied on the background of the image, i.e. between fibres, the estimate is dominated by noise due to the lack of structure and uniform direction.

4 Experimental Results and Analysis

The evaluation has been conducted on a database of 525 CCM images captured using the HRT-III[1] microscope from 69 subjects (20 controls and 49 diabetic patients). The resolution is $1.0417\mu m$ and the field of view is $400 \times 400\mu m^2$ of the cornea. For each individual, several fields of view are selected manually from near the centre of the cornea that show recognisable nerve-fibres. Using the Neuropathy Disability Score (NDS) [1], 48 patients were categorised into four groups according to severity of neuropathy (asymptomatic: $0 \geq$NDS≤ 2 ($n = 26$), mild: $3 \geq$NDS≤ 5 ($n = 9$), moderate: $6 \geq$NDS≤ 8 ($n = 10$) and severe: $9 \geq$NDS≤ 10 ($n = 3$)).

The performance of all methods is obtained by validating the extracted nerve-fibres in comparison with an expert manual delineation using *CCMetrics*[2]. Only the raw response of each method is taken into account without any further post-processing operations or shade correction methods as shown in Figure (1). Binary images are obtained by a simple uniform thresholding operation that is followed by a thinning operation to achieve a one-pixel-wide skeleton image.

4.1 Comparison of Nerve-Fibre Detection Methods

Three measures have been used in order to quantify the evaluation: the false-positive (FPR), the true-positive (TPR) and the equal-error rate (EER), which is the average of optimal FPR and false-negative rate at minimal difference between both. The measurements are taken by comparing the generated skeleton at different threshold intervals of the methods' responses with the manually delineated "ground-truth". A tolerance of $\pm 3.141\mu m$ (3 pixels) was allowed in

[1] Heidelberg Engineering Inc., modified to acquire corneal confocal images.

[2] CCMetrics is a purpose built interactive graphical interface which helps experts to manually delineate nerve-fibres in CCM images.

determining coincidence between the ground-truth and the detected nerve-fibres. The Peak Signal to Noise Ratio (PSNR) is also used to evaluate the performance of all methods. The PSNR is computed with respect to the mean squared error of the detected nerve-fibres from the manual delineation. The practical implementations of the Hessian, the DTCWT and the Monogenic signal were obtained from public domain sources [12,14,13], while the rest are implemented by our research group.

The EER and PSNR values for all the methods are presented in the box-plots in Figure 2 and Table 1. Each data point in Figure 2 corresponds to the evaluation on one of the 525 CCM images in the database. The dual-model shows lower EER and higher PSNR than all other methods (Table 1). These improvements are statistically significant ($p \approx 0$ using three different non-parametric tests). The table also shows that the standard deviations of both EER and PSNR are low for the dual-model, which indicates a more stable and robust behaviour. The closest competitor is Linop. The methods designed for linear structures perform rather better on this test than the more generic DTCWT and Monogenic signal methods.

The superior performance of the dual-model is borne out by the ROC curves of Figure 2, in which the dual model shows improved detection at all operation points.

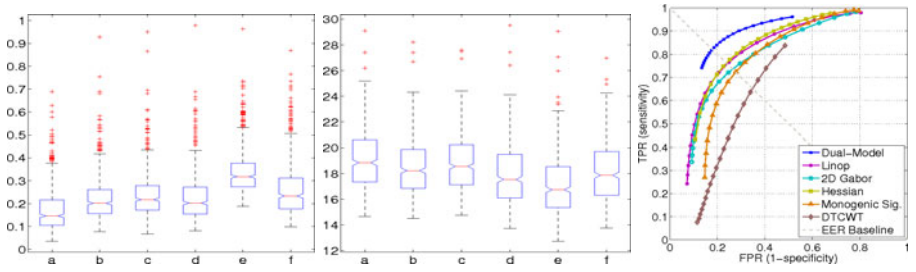


Fig. 2. From left to right, the box-plots of the EER and the PSNR are shown for all methods. The ROC curves are presented at the far right. The box-plots indicate the upper and the lower quartiles as well as the median (the bar) of the EER and PSNR values respectively; whiskers show the extent of the rest of the data while crosses indicate outliers for (a) dual-model, (b) Linop, (c) 2D Gabor, (d) Hessian, (e) DTCWT and (f) Monogenic.

Table 1. A comparison of mean EER and PSNR and their standard deviations

	Dual-Model	Linop [4]	2D Gabor [2,9]	Hessian [6]	DTCWT [11]	Monog. [5]
$EER_{[\%]}$	17.79± 10.58	22.65± 10.76	24.15± 10.74	23.14± 11.53	34.17± 10.43	26.50± 12.58
$PSNR_{[dB]}$	19.08± 2.16	18.51± 2.09	18.80± 2.11	17.93± 2.27	17.00± 2.23	18.11± 2.20

4.2 Assessment of Clinical Utility Results

In previous studies, using manual measurement of nerve-fibres, several features have been used to quantify the CCM images, including nerve-fibre length (NFL):

the total length of nerve-fibres measured in an image, nerve-fibre density: the total number of nerve-fibres per unit area and branch density: the number of fibre branches per unit area. Of these nerve-fibre length proved to be the most discriminating, and we use this measure here to compare automated with manual measurement of the nerve-fibre images.

The box-plots in Figure 3 show a strong similarity between the manual and the automated analysis. However the scale of the NFL has slightly changed from $(3.68-33.91)$ for the manual analysis to $(5.67-26.53)$ for the automated analysis. ANOVA analysis results in a p-value for discrimination among these groups which is slightly higher for the automated than the manual analysis, though both are significant $(p \approx 0)$. The automated NFL measurements show a very strong correlation $(r = 0.92)$ with the manual NFL values, which indicates that the automated system is successfully identifying the correct nerve-fibres. The coefficient of variation $c_v = \frac{\sigma}{\mu}$ of the manual analysis is 0.34, reducing for the automated analysis to 0.29, which indicates more reliability and robustness of the results.

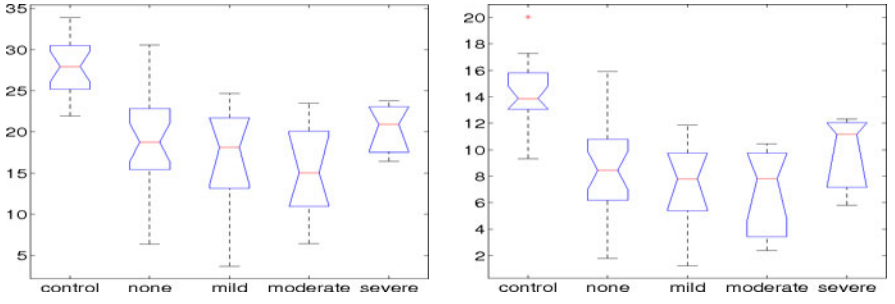


Fig. 3. Box-plots showing the NFL scores for each of the NDS groups calculated manually (left) and automatically (right)

5 Conclusion

The analysis of CCM images requires the identification of fibre-like structures with low contrast in noisy images. This is a requirement shared by a number of imaging applications in biology, medicine and other fields. A number of methods have been applied in these applications, and we have compared some of these, and more generic methods with a dual-model detection algorithm devised for this study. The comparison used a large set of images with manual ground truth. In terms of both error-rates (pixel misclassification) and signal-to-noise ratio, the dual model achieved highest performance. It seems reasonable to propose that this filter is likely to prove equally useful in applications of a similar nature. The question of the clinical utility of the method was also addressed in this paper. The evaluation has shown that the automatic analysis is consistent with the manual ground truth with a correlation of $(r = 0.92)$. Similarity in grouping control and patient subjects between manual and automated analysis was also achieved with $(p \approx 0)$. Therefore, it is sound to conclude that the automated analysis, which can be much quicker, is a potentially more reliable and practical alternative to

manual analysis due to its consistency and immunity to the inter/intra-observer variabilities.

References

1. Abbott, C.A., Carrington, A.L., Ashe, H., Bath, S., Every, L.C., Griffiths, J., Hann, A.W., Hussein, A., Jackson, N., Johnson, K.E., Ryder, C.H., Torkington, R., Ross, E.R.E.V., Whalley, A.M., Widdows, P., Williamson, S., Boulton, A.J.M.: The north-west diabetes foot care study: incidence of, and risk factors for, new diabetic foot ulceration in a community-based patient cohort. Diabetic Medicine 19(5), 377–384 (2002)
2. Dabbah, M.A., Graham, J., Tavakoli, M., Petropoulos, Y., Malik, R.A.: Nerve fibre extraction in confocal corneal microscopy images for human diabetic neuropathy detection using gabor filters. In: Medical Image Understanding and Analysis (MIUA), pp. 254–258 (July 2009)
3. Daugman, J.G.: Two-dimensional spectral analysis of cortical receptive field profiles. Vision Research 20(10), 847–856 (1980)
4. Dixon, R.N., Taylor, C.J.: Automated asbestos fibre counting. In: Machine Aided Image Analysis, pp. 178–185. Institute of Physics, London (1979)
5. Felsberg, M., Sommer, G.: The monogenic signal. IEEE Transactions on Signal Processing 49(12), 3136–3144 (2001)
6. Frangi, A.F., Niessen, W.J., Vincken, K.L., Viergever, M.A.: Multiscale vessel enhancement filtering. In: Wells, W.M., Colchester, A.C.F., Delp, S.L. (eds.) MICCAI 1998. LNCS, vol. 1496, pp. 130–137. Springer, Heidelberg (1998)
7. Hong, L., Wan, Y., Jain, A.: Fingerprint image enhancement: algorithm and performance evaluation. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(8), 777–789 (1998)
8. Hossain, P., Sachdev, A., Malik, R.A.: Early detection of diabetic peripheral neuropathy with corneal confocal microscopy. The Lancet 366(9494), 1340–1343 (2005)
9. Jain, A.K., Farrokhnia, F.: Unsupervised texture segmentation using gabor filters. Pattern Recognition 24(12), 1167–1186 (1991)
10. Kallinikos, P., Berbanu, M., O'Donnell, C., Boulton, A., Efron, N., Malik, R.: Corneal nerve tortuosity in diabetic patients with neuropathy. Investigative Ophthalmology & Visual Science 45(2), 418–422 (2004)
11. Kingsbury, N.: Complex wavelets for shift invariant analysis and filtering of signals. Applied and Computational Harmonic Analysis 10(3), 234–253 (2001)
12. Kingsbury, N.: Dual-Tree Complex Wavelet Transform Pack (June 2002), `http://www-sigproc.eng.cam.ac.uk/~ngk/`
13. Kovesi, P.: An implementation of Felsberg's monogenic filters (August 2005), `http://www.csse.uwa.edu.au/~pk/research/matlabfns/`
14. Kroon, D.J., Schrijver, M.: Hessian based Frangi Vesselness filter (October 2009), `http://www.mathworks.co.uk/`
15. Malik, R.A., Kallinikos, P., Abbott, C.A., van Schie, C.H.M., Morgan, P., Efron, N., Boulton, A.J.M.: Corneal confocal microscopy: a non-invasive surrogate of nerve fibre damage and repair in diabetic patients. Diabetologia 46(5), 683–688 (2003)
16. Rao, A.R.: A taxonomy for texture description and identification. Springer, New York (1990)
17. Ruggeri, A., Scarpa, F., Grisan, E.: Analysis of corneal images for the recognition of nerve structures. In: IEEE Conference of the Engineering in Medicine and Biology Society (EMBS), pp. 4739–4742 (September 2006)
18. Zwiggelaar, R., Astley, S., Boggis, C., Taylor, C.: Linear structures in mammographic images: Detection and classification. IEEE Transactions on Medical Imaging 23(9), 1077–1086 (2004)

53. **Automatic analysis of diabetic peripheral neuropathy using multi-scale quantitative morphology of nerve fibres in corneal confocal microscopy imaging**. M.A. Dabbah, J. Graham, I.N. Petropoulos, M. Tavakoli, R.A. Malik, *Med. Image Anal 15(5): 738-747 (2011)*. doi:10.1016/j.media.2011.05.016

Medical Image Analysis 15 (2011) 738–747

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Automatic analysis of diabetic peripheral neuropathy using multi-scale quantitative morphology of nerve fibres in corneal confocal microscopy imaging

M.A. Dabbah [a,*], J. Graham [a,c], I.N. Petropoulos [b], M. Tavakoli [b], R.A. Malik [b]

[a] Imaging Sciences and Biomedical Engineering (ISBE), The University of Manchester, Stopford Building, Oxford Rd., Manchester M13 9PT, UK
[b] Cardiovascular Research Group, The University of Manchester, 46 Grafton St., Manchester M13 9NT, UK
[c] School of Computer Science, The University of Manchester, Kilburn Building, Oxford Road, Manchester M13 9PL, UK

ARTICLE INFO

Article history:
Available online 13 June 2011

Keywords:
Corneal confocal microscopy
Diabetic neuropathy
Curvilinear structures
Image quantification

ABSTRACT

Diabetic peripheral neuropathy (DPN) is one of the most common long term complications of diabetes. Corneal confocal microscopy (CCM) image analysis is a novel non-invasive technique which quantifies corneal nerve fibre damage and enables diagnosis of DPN. This paper presents an automatic analysis and classification system for detecting nerve fibres in CCM images based on a multi-scale adaptive dual-model detection algorithm. The algorithm exploits the curvilinear structure of the nerve fibres and adapts itself to the local image information. Detected nerve fibres are then quantified and used as feature vectors for classification using random forest (RF) and neural networks (NNT) classifiers. We show, in a comparative study with other well known curvilinear detectors, that the best performance is achieved by the multi-scale dual model in conjunction with the NNT classifier. An evaluation of clinical effectiveness shows that the performance of the automated system matches that of ground-truth defined by expert manual annotation.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

According to numerous clinical reports (DiabetesUK, 2010), diabetes is among the most challenging chronic health problems. For example, in the UK it is estimated that one in twenty people has diabetes, whether diagnosed or undiagnosed, and by 2025 four million people will have the condition. Damage to the peripheral nerves (diabetic peripheral neuropathy, DPN) is one of the commonest long-term complications of diabetes occurring in at least 50% of patients with diabetes (Boulton, 2005). As a consequence, about one in six diabetic patients have chronic painful neuropathy, compared to one in 20 non-diabetic subjects (Daousi et al., 2004). It is the main initiating factor for foot ulceration, Charcot's neuroarthropathy and lower extremity amputation. As 80% of amputations are preceded by foot ulceration, an effective means of detecting and treating neuropathy would have a major medical, social and economic impact. The development of new treatments to slow, arrest or reverse this condition is of paramount importance but is presently limited due to difficulties with end points employed in clinical trials (Dyck et al., 2007). Therefore accurate detection and quantification of DPN are important to define at-risk patients, anticipate deterioration, and assess new therapies. Current methods are unsatisfactory, lacking sensitivity and requiring expert assessment, and focus only on large fibres (neurophysiology) or are invasive (skin/nerve biopsy). Unfortunately, diabetic neuropathy lacks a non-invasive surrogate for nerve damage (Tesfaye et al., 2010).

Recent research (Malik et al., 2003; Kallinikos et al., 2004; Hossain et al., 2005) using corneal confocal microscopy (CCM) suggests that this non-invasive, and hence reiterative, test might be an ideal surrogate endpoint for human diabetic neuropathy. The establishment of CCM as a surrogate for early diagnosis and an early biomarker for diabetic neuropathy could identify those at risk and prompt more intense intervention including improved glycaemic, blood pressure and lipid control. Furthermore a sensitive surrogate endpoint would significantly lower hurdles to the development of disease-modifying therapeutics by enhancing the capacity to test therapeutic efficacy. The major advance of CCM is the entirely non-invasive and relatively rapid (≈2 min) acquisition of images of small nerve fibres in patients. However, the major limitation preventing extension of this technique to wider clinical practice is that analysis of the images using interactive image analysis is highly labour-intensive and requires considerable expertise to quantify nerve pathology. To be clinically useful as a diagnostic tool, it is essential that the measurements be extracted automatically.

If an automatic CCM image analysis system is to be applied clinically, especially to define early degeneration or regeneration, then a key step is the automatic detection of low-contrast nerve fibres

* Corresponding author.
E-mail address: m.a.dabbah@manchester.ac.uk (M.A. Dabbah).

1361-8415/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.media.2011.05.016

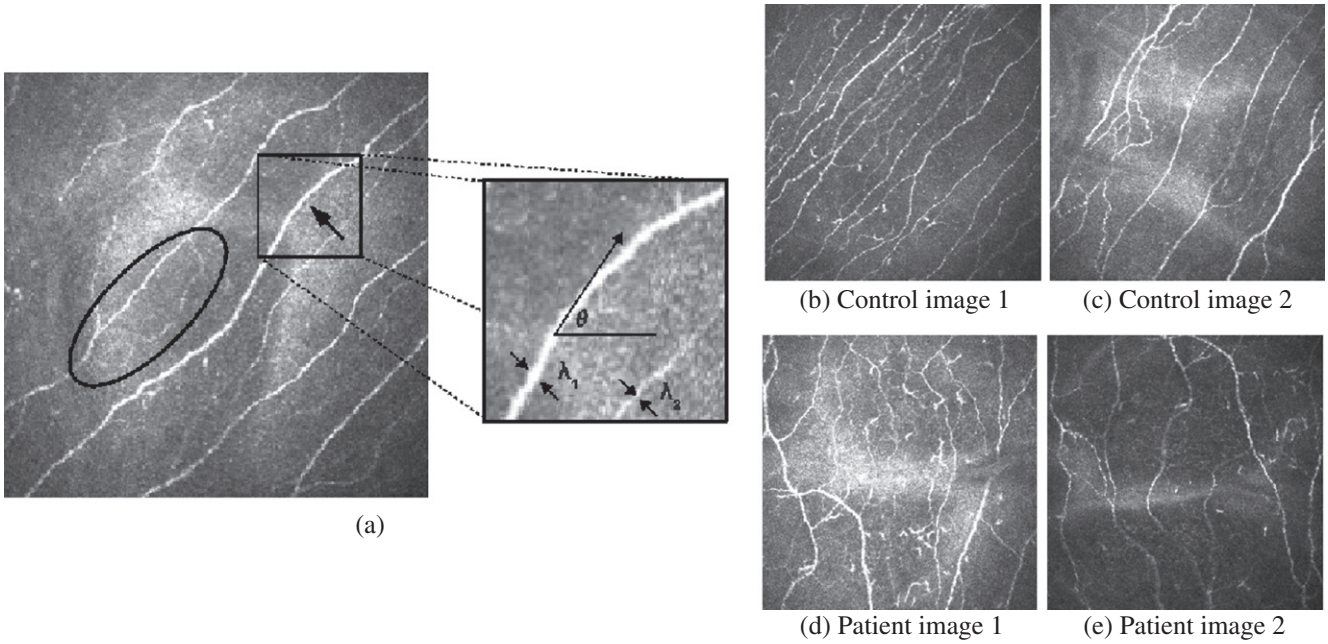


Fig. 1. (a) An example of CCM image and nerve-fibre characteristics. (b)–(e) Samples of CCM image from controls and patients, showing the effects of different imaging artefacts and neuropathy status.

among image noise (see Fig. 1). The literature on this topic is not extensive, although the problem has a superficial similarity to other, more widely investigated, applications, such as detection of blood-vessels in retinal images. Ruggeri et al. (2006) and Scarpa et al. (2008) describe a heuristic method that was adapted from retinal analysis. A number of methods have been developed to enhance the contrast of such linear structures. In a previous study (Dabbah et al., 2009), we used the 2D Gabor filter (Jain and Farrokhnia, 1991) to detect nerve fibres in CCM images. The filter is a band-pass filter that consists of a sinusoidal plane wave with a certain orientation and frequency, modulated by a Gaussian envelope. This spatial domain enhancement is based on the convolution of the image with the even-symmetric Gabor filter that is tuned to the local nerve-fibre orientation. We subsequently extended this to form a dual-model detector (Dabbah et al., 2010), see Section 4.

The automated system of analysing CCM images presented in this paper is an extension of our previous single scale dual-model fibre detector (Dabbah et al., 2010). The new detection algorithm uses the dual-model property in a multi-scale framework to generate feature vectors from localised information at every pixel. These vectors are then used to classify pixels using random forests (RF) (Breiman, 2001) and neural networks (NNT) (Moller, 1993).

In the remainder of the paper we introduce CCM imaging, the image characteristics and the metrics that have been used to quantify the nerve morphology by interactive image analysis (Sections 2 and 3). We describe the single-scale dual model filter (Dabbah et al., 2010) and its extension to multiple scales with pixel classification (Sections 4–6). In Section 7 we describe a comparative evaluation showing the improved performance of the multi-scale version over not only the single-scale filter but a number of other multi-scale detectors. We also demonstrate that the automatically detected fibres result in morphometric features equivalent to those generated by expert interactive analysis.

2. Corneal confocal microscopy

The cornea is one of the body's most innervated tissues. The sub-basal nerve plexus runs parallel to the surface of the cornea in the Bowman's membrane, lying between the outer epithelial layer and the stroma. Bowman's layer is about 8–12 μm thick, and the nerves may be imaged by confocal microscopy using either a white-light source or a laser source. In this study laser confocal microscopy was used.[1] Typical images are shown in Fig. 1.

2.1. CCM for imaging diabetic peripheral neuropathy

Recent studies suggest that small unmyelinated c-fibres may be the earliest to be damaged in diabetic neuropathy (Umapathi et al., 2007; Loseth et al., 2008; Malik et al., 2005). The only techniques which allow a direct examination of unmyelinated nerve fibre damage are those of sural nerve biopsy with electron microscopy (Malik et al., 2005, 2001), and the skin-punch biopsy (Novella et al., 2001; Singleton et al., 2001; Sumner et al., 2003), but both are invasive procedures. However, our previous studies in patients with diabetic neuropathy have shown that CCM can be used to quantify early small nerve fibre damage and accurately quantify the severity of diabetic neuropathy (Malik et al., 2003; Kallinikos et al., 2004). Moreover, we have shown that corneal nerve damage assessed using CCM relates to the severity of intra-epidermal nerve fibre loss in foot skin biopsies (Quattrini et al., 2007) and the loss of corneal sensation (Tavakoli et al., 2007) in diabetic patients. CCM also detects early nerve fibre regeneration following pancreas transplantation in diabetic patients (Mehra et al., 2007). Recently we have also shown that CCM detects nerve fibre damage in patients with Fabry disease (Tavakoli et al., 2009) and idiopathic small fibre neuropathy (Tavakoli et al., 2010a) in the presence of normal electro-physiology and quantitative sensory testing (QST). CCM offers considerable potential as a surrogate marker, and hence as an endpoint for clinical trials in diabetic neuropathy (Tavakoli et al., 2010b; Hossain et al., 2005).

[1] The Heidelberg Retina Tomograph (HRT-III) confocal scanning laser ophthalmoscope developed by Heidelberg Engineering Inc. The instrument can be converted into a corneal confocal microscope using a microscope lens which is attached to the standard lens.

2.2. Nerve fibre quantification

Nerve fibres in CCM images appear as bright linear structures that flow in a predominant direction everywhere. However nerve fibres have their independent local orientation θ (Fig. 1). They also have different dimensions of length and diameter λ. Longer nerve fibres with larger diameter are considered to be the main trunks while nerve fibres branching from the main trunks are considered to be secondary nerve fibres (or nerve branches) as shown in the square and the ellipse of Fig. 1 respectively.

Previous analyses of CCM images have used manual delineation of the nerve fibres by experts (Malik et al., 2003; Kallinikos et al., 2004; Hossain et al., 2005). These studies have shown promising results in distinguishing control and patient groups using features such as nerve-fibre length (NFL), nerve-fibre density (NFD), nerve-branch density (NBD) and tortuosity (NFT) of nerve fibres. Abnormal subjects usually have fewer nerve fibres than normal subjects and more tortuous structures as shown in Fig. 1. This in turn affects the quantified metric, that may provide a diagnosis of the neuropathy.

NFL, which we return to in Section 7, is defined as the total length of all nerve fibres visible in the CCM image per square mm. The total length is computed by tracing all the nerve fibres and nerve-branches in the image. This number is then divided by the area of the field-of-view provided by the microscope to produce the NFL (mm/mm^2) value.

2.3. Artefacts

Although the process of capturing the images is relatively short and quick, saccadic eye movement is faster, which could result in motion or blurring effects of the nerve fibres. As shown in the image samples of Fig. 1, the nerve fibres may also appear very faint due to differences of depth. The same nerve fibre could appear and disappear several times as it moves in and out of the focus plane. This movement will also affect the visual diameter and the brightness of the fibre. Since the cornea is a transparent spherical structure, illumination artefacts arise that result in low-frequency variation in image brightness and contrast. As shown in Fig. 1d, CCM images also contain small bright structures (usually cells) that are not nerve fibres, which add to the challenge of identifying nerve fibres.

3. Linear-structure and feature detection

Detection of curvilinear structures is a requirement in several applications of medical image analysis. A method of linear structure detection (Line Operator – LinOp), originally developed for detection of asbestos fibres (Dixon and Taylor, 1979) has also been shown to be effective in detecting ducts in mammograms (Zwiggelaar et al., 2004). LinOp exploits the linear nature of the structures to enhance their contrast by computing the average intensity of pixels lying on a line passing through the reference pixel for multiple orientations and scales. The largest values are chosen to correspond to the line, the strength of which is determined by the difference with the average intensity of the similarly oriented square neighbourhood.

Frangi et al. (1998) used a multiscale decomposition of the Hessian matrix to detect and measure blood vessels in Digital Subtraction Angiography images. They derived a discriminant function based on the eigenvalues and eigenvectors that has maximum response for tube-like structures. The external energy is used to attract the curve towards points which have a high likelihood of lying on a central vessel axis.

The Monogenic signal (Felsberg and Sommer, 2001) is a 2D generalization of the analytic signal, widely used in time-domain signal processing. There are several possible ways of extending this approach to multiple dimensions. The Monogenic signal approach makes use of the Riesz transform, and results in separating the signal into local amplitude (or "structure" corresponding approximately to image intensity) and local phase (corresponding to local changes). It has been used in extracting structure information (such as edge and ridge) from images in several medical image analysis applications (Pan et al., 2006; Ali et al., 2008).

In a preliminary study (Dabbah et al., 2009), we used the 2D Gabor filter (Jain and Farrokhnia, 1991) to detect nerve fibres in CCM images. This spatial domain enhancement is based on the convolution of the image with the even-symmetric Gabor filter that is tuned to the local nerve-fibre orientation.

4. Single-scale dual-model enhancement

All of the methods described in Section 3 are potential means of enhancing the linear nerve structures in the face of the image corruption outlined in Section 2.3. In Dabbah et al. (2010) we reported on the performance of the single-scale dual-model detector in comparison with these methods. We showed that the detectors specifically designed for detection of linear structures performed better than more general feature detectors, such as the Monogenic filter. In particular the single-scale dual-model detector was superior to all of them. In this section we briefly describe the algorithm.

4.1. Nerve-fibre contrast enhancement

The dual model consists of separate models of foreground and background, which adapt to local image conditions to cope with slowly varying illumination artefacts. The foreground model $\mathcal{M}_{\mathcal{F}}$ is an even-symmetric and real-valued Gabor (Jain and Farrokhnia, 1991; Daugman, 1980) wavelet and the background model $\mathcal{M}_{\mathcal{B}}$ is a two-dimensional Gaussian envelope.

$$\mathcal{M}_{\mathcal{F}}(x_\theta, y_\theta) = \left(\cos\left(\frac{2\pi}{\lambda}x_\theta + \phi\right)\right) \cdot \exp\left\{-\frac{1}{2}\left(\frac{x_\theta^2}{\sigma_x^2} + \frac{\gamma^2 y_\theta^2}{\sigma_y^2}\right)\right\} \quad (1)$$

$$\mathcal{M}_{\mathcal{B}}(x_\theta, y_\theta) = \alpha \exp\left\{-\frac{1}{2}\left(\frac{x_\theta^2}{\sigma_x^2} + \frac{\gamma^2 y_\theta^2}{\sigma_y^2}\right)\right\} \quad (2)$$

$$x_\theta = x\cos\theta + y\sin\theta \quad (3)$$

$$y_\theta = -x\sin\theta + y\cos\theta \quad (4)$$

The x and y axes of the dual-model coordinate frame x_θ and y_θ are defined by a rotation of θ, which is the dominant orientation of the nerve fibres in a particular region within the image (see Section 4.2). λ and ϕ are the wavelength and the phase of the sinusoidal signal modulated by the 2D Gaussian envelope with x-axis variance σ_x^2 and y-axis variance σ_y^2. The aspect ratio of the Gaussian kernel is defined by γ and its magnitude is α. This dual-model is used to generate the positive response $\mathcal{R}_{\mathcal{P}} = \mathcal{M}_{\mathcal{F}} + \mathcal{M}_{\mathcal{B}}$ and the negative response $\mathcal{R}_{\mathcal{N}} = \mathcal{M}_{\mathcal{F}} - \mathcal{M}_{\mathcal{B}}$ that are applied to the original CCM image and can be represented as in Eqs. (5) and (6) respectively.

$$\mathcal{R}_{\mathcal{P}}(x_\theta, y_\theta) = \left[\cos\left(\frac{2\pi}{\lambda}x_\theta + \phi\right) + \alpha\right] \cdot \exp\left\{-\frac{1}{2}\left(\frac{x_\theta^2}{\sigma_x^2} + \frac{\gamma^2 y_\theta^2}{\sigma_y^2}\right)\right\} \quad (5)$$

$$\mathcal{R}_{\mathcal{N}}(x_\theta, y_\theta) = \left[\cos\left(\frac{2\pi}{\lambda}x_\theta + \phi\right) - \alpha\right] \cdot \exp\left\{-\frac{1}{2}\left(\frac{x_\theta^2}{\sigma_x^2} + \frac{\gamma^2 y_\theta^2}{\sigma_y^2}\right)\right\} \quad (6)$$

The equations of $\mathcal{R}_{\mathcal{P}}$ and $\mathcal{R}_{\mathcal{N}}$ assume that the Gaussian envelope of both responses are identical, *i.e.* they have the same variances $\sigma_{(x,y)}^2$ and the same aspect ratio γ. The magnitude of the Gaussian envelope α defines the threshold in which a nerve fibre can be distin-

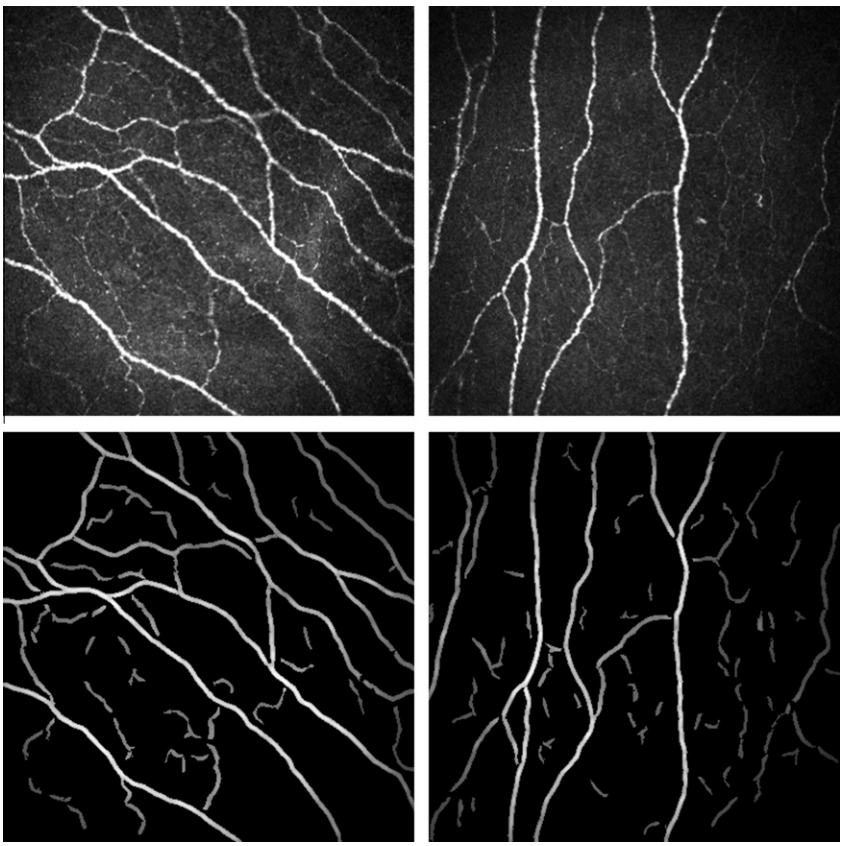


Fig. 2. An illustration of the single-scale dual-model detector (Dabbah et al., 2010). The images in the top row are the original CCM images, and their response is shown in the bottom row.

guished from the background image. The value of α can be set empirically to control sensitivity and accuracy of detection. The wavelength λ defines the frequency band of the information to be detected in the CCM image. Its value might be computed for a sub-region within the image that has significant variability of nerve-fibre width. However for simplicity, λ is chosen to be a global estimate of the entire image based on empirical results.

This in turn enhances the nerve fibres that are oriented in the dominant direction, and decreases anything that is oriented differently by increasing the contrast between the foreground and the noisy background, whilst effectively reducing noise around the nerve-fibre structure as shown in Fig. 2. This pixel-wise operation adjusts the models to suit the local neighbourhood characteristics of the reference pixel at $\mathbf{I}(i,j)$ by modifying the parameters of the foreground and background models. The dot products of the models and the reference pixel's neighbourhood (Eqs. (7) and (8)) are then combined to generate the final enhanced value of this particular reference pixel $\mathbf{g}^{(i,j)}$ (Eq. (9)).

$$\Gamma_p^{(i,j)} = \langle \mathbf{I}_\omega(i,j), \mathcal{R}_\mathcal{P} \rangle \tag{7}$$

$$\Gamma_n^{(i,j)} = \langle \mathbf{I}_\omega(i,j), \mathcal{R}_\mathcal{N} \rangle \tag{8}$$

$$\mathbf{g}^{(i,j)} = \frac{\Gamma_p^{(i,j)}}{1 + e^{\left(-2k\Gamma_n^{(i,j)}\right)}} \tag{9}$$

The neighbourhood area, $\mathbf{I}_\omega(i,j)$, of the reference pixel (i,j) is defined by the width ω. $\mathcal{R}_\mathcal{P}$ and $\mathcal{R}_\mathcal{N}$ are the responses from Eqs. (5) and (6). $\langle \cdot, \cdot \rangle$ is the dot product operator. The sharpness of the transition of the enhanced image value at a particular pixel $\mathbf{g}^{(i,j)}$ is controlled by k. A larger k amounts to a sharper transition when $\Gamma_n = 0$.

4.2. Nerve-fibre orientation estimation

In CCM images, the nerve fibres flow in locally consistent orientations. In addition, there is a global orientation that dominates the general flow. This orientation field describes the coarse structure of nerve fibres. Using the least mean square (LMS) algorithm (Hong et al., 1998), the local orientation of the block centred at a certain pixel is computed as in Rao (1990).

Since the orientations vary at a slow rate, a low-pass Gaussian filter is applied globally in order to further reduce errors at near-nerve fibre and non-nerve fibre regions. The LMS produces a stable smooth orientation field in the region of the nerve fibres; however when applied on the background of the image, *i.e.* between fibres, the estimate is dominated by noise due to the lack of structure and uniform direction.

4.3. Nerve fibre extraction

The response image is a map of the confidence at each pixel that it corresponds to a nerve fibre. The sharp transition of the dual-model between background and foreground has resulted in useful characteristics in the response image, Fig. 2. Well-defined nerve fibres are more likely to appear as connected structures, while noise and small undesired curvilinear structures will also be detected but usually manifested as ill-defined and disoriented small fragments. This makes the extraction of nerve fibres a trivial task, and the separation of noise and information becomes easier in the post-processing stages.

The coordinates of each detected nerve fibre are considered to be the central pixel along the width of the detected objects that appear as thick ridges flowing across the image. Hence, after the noise

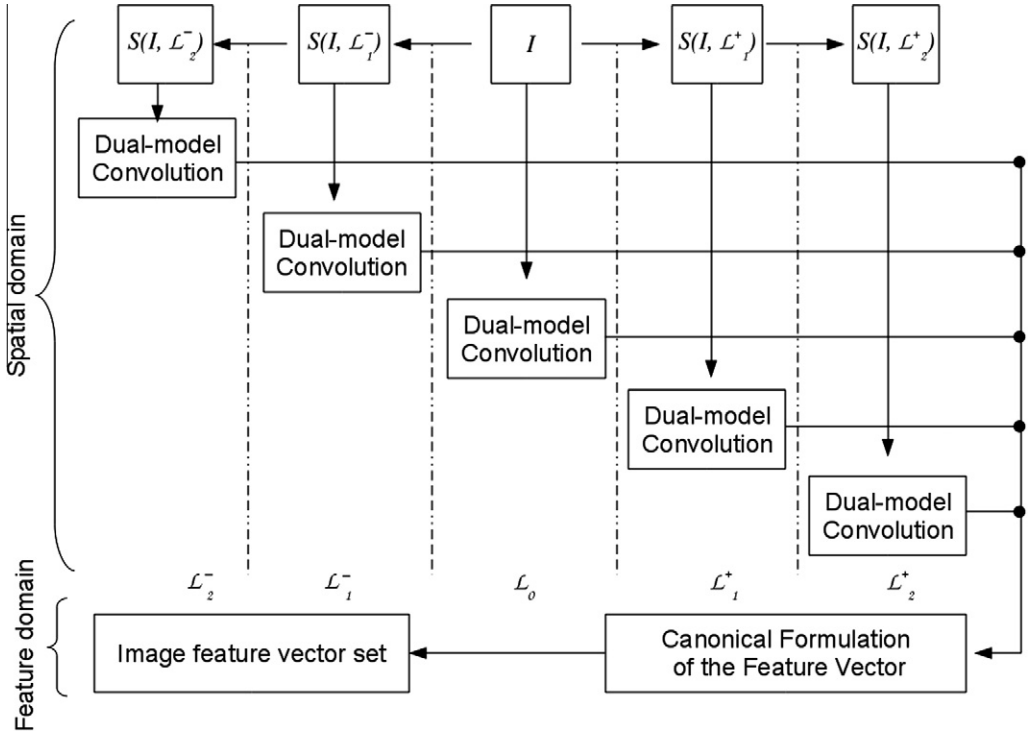


Fig. 3. A conceptual diagram illustrating the operation of the multi-scale dual-model detection algorithm. The images are convolved with the adaptable dual-model algorithm at different scales and then the responses are combined in the feature space to generate a feature vector for every pixel in the image. The $S(\cdot)$ is the scaling function.

(small fragments) is removed in post-processing, the response images are converted to binary images using a global threshold. The remaining large fragments represent the detected nerve fibres and are thinned using the method of Zhang and Suen (1984) to obtain the skeleton image (*i.e.* the one-pixel wide line).

5. Multi-resolution dual-model enhancement

The single resolution detector described in Section 4 makes use of local orientations calculated on a regional basis and operates with a single wavelength parameter for the Gabor filter, thereby assuming a single width for all fibres. In this section we extend the model to multiple resolutions using a scale pyramid as shown in Fig. 3. We also calculate responses over a range of orientations, selecting the most appropriate scale and orientation of the response by pixel classification. There are three parameters of the Gabor filter that can vary in scale: λ, the wavelength of the sinusoid and σ_x and σ_y, the widths of the Gaussian envelope. To explore this scale space efficiently, we make use of the single-scale results, choosing values of λ, σ_x and σ_y at the original image scale to be the values used in the single-scale detector. Keeping these values constant we create a pixel pyramid by sub-sampling (with smoothing) and super-sampling (by interpolation) the original image. While super-sampling the image adds no new information to the pyramid, it has the effect of reducing the wavelength and Gaussian widths of the Gabor filter relative to the size of the image structure.

5.1. Image pyramid

Let us denote $\mathcal{L}$ as a vector set of different scale (re-sampling levels) parameters. Each level l represent a set of estimated parameters used in the dual-model detection. The spatial frequency of the image structure (nerve fibres) in l is defined by the λ.

$$\mathcal{L} \triangleq \{l_L^-, l_{L-1}^-, \ldots, l_0, \ldots, l_{L-1}^+, l_L^+\} \in \mathbb{R}^3 | L \in \mathbb{Z}^+ \tag{10}$$

$$l_k^{\pm} \triangleq \{\lambda, \sigma_x, \sigma_y\} \mid k = 1, 2, \ldots, 2L+1 \tag{11}$$

For example $l_1^-(\lambda)$ defines the wavelength of the Gabor filter's sinusoid at the super-sampled level 1. $l_2^-(\sigma_x)$ defines the Gaussian spread in x of the Gabor filter at the sub-sampled level 2, etc. In our implementation $L = 2$ and the pixel sampling is doubled (halved) between levels. At each level, eight values of orientation (θ) are explored. The specific values of λ, σ_x and σ_y at level l_0 were defined empirically in the single-scale detector to be $\lambda = 9$, $\sigma_x = 4$ and $\sigma_y = 3$.

5.2. Feature vector extraction

In order to generate the feature vector of each CCM image **I** we use the transform $\mathcal{T} : \mathbb{R}^{M\times N} \rightarrow \mathbb{R}^{M\times N\times O\times S}$, where $M \times N$ are the dimensions of the image, O is the number of orientations used and S is the number of levels in the pyramid ($2L + 1$). Analogous to the single-scale dual-model detection algorithm in Section 4, transform $\mathcal{T}$ consists of two models: foreground model $\mathcal{M}_{\mathcal{F}}(x_\theta, y_\theta, l_k^{\pm})$ and background model $\mathcal{M}_{\mathcal{B}}(x_\theta, y_\theta, l_k^{\pm})$. The difference between these models and those of the single scale (Eqs. (1) and (2)) is that they are a function of the different scales defined by the pyramid $\mathcal{L}$. Also, all orientations are computed at every pixel unlike the single model where orientation is locally estimated. Hence there are no equivalents of Eqs. (3) and (4) in this case.

$$\mathcal{M}_{\mathcal{F}}(x_\theta, y_\theta, l_k^{\pm}) = \cos\left(\frac{2\pi}{\lambda_k} x_\theta + \phi\right) \cdot \exp\left\{-\frac{1}{2}\left(\frac{x_\theta^2}{\sigma_{xk}^2} + \frac{\gamma^2 y_\theta^2}{\sigma_{yk}^2}\right)\right\} \tag{12}$$

$$\mathcal{M}_{\mathcal{B}}(x_\theta, y_\theta, l_k^{\pm}) = \alpha \cdot \exp\left\{-\frac{1}{2}\left(\frac{x_\theta^2}{\sigma_{xk}^2} + \frac{\gamma^2 y_\theta^2}{\sigma_{yk}^2}\right)\right\} \tag{13}$$

The adaptation of these two models across the complete range of scales and orientations defined by the pyramid $\mathcal{L}$ should cover all of the relevant feature space of the nerve fibres. By convolving them

with the images to generate foreground and background responses $\mathcal{R}_{\mathcal{F}}(\theta,\lambda)$ and $\mathcal{R}_{\mathcal{B}}(\theta)$, and finding the difference $\mathcal{G}_i$ between these responses we can generate the feature vector $\mathcal{F}$ that describes the CCM image $\mathbf{I}$.

$$\mathcal{R}_{\mathcal{F}}(\theta,l_k^{\pm}) = \mathbf{I} * \mathcal{M}_{\mathcal{F}}(x_\theta, y_\theta, l_k^{\pm}) \tag{14}$$

$$\mathcal{R}_{\mathcal{B}}(\theta,l_k^{\pm}) = \mathbf{I} * \mathcal{M}_{\mathcal{B}}(x_\theta, y_\theta, l_k^{\pm}) \tag{15}$$

$$\begin{aligned}\mathcal{G}_i &= \mathcal{R}_{\mathcal{F}}(\theta,l_k^{\pm}) - \mathcal{R}_{\mathcal{B}}(\theta,l_k^{\pm}) \\ &= \mathbf{I} * \left(\cos\left(\frac{2\pi}{\lambda_k}x_\theta + \phi\right) - \alpha\right)\cdot \\ &\quad \exp\left\{-\frac{1}{2}\left(\frac{x_\theta^2}{\sigma_{xk}^2} + \frac{\gamma^2 y_\theta^2}{\sigma_{yk}^2}\right)\right\}\end{aligned} \tag{16}$$

$$\mathcal{F} = \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_{O\times S}\} \tag{17}$$

α is the threshold parameters that is equivalent to the same parameter in the single-scale detector (Eqs. (5) and (6)). However, in this multi-scale algorithm the logistic transition (Eq. (9)) is replaced by the classification step of the generated feature vector $\mathcal{F}$ in order to make the final decision.

5.3. Canonical form of the feature vector

Unlike the single-scale detector, the interpretation of the response is not trivial. Applying the transform $\mathcal{T}$ using the pyramid $\mathcal{L}$ generates longer feature vectors which raises the questions of how to interpret the response in the best possible way.

Since these feature vectors are associated with certain orientations, frequencies and local regions of the image, the specific sequence of features in the feature vectors is dependent on the order which these features are formulated. For example the order of the feature vector provides information about the local orientation of the fibre. This is useful to know, but irrelevant to classifying the pixel as belonging to the fibre or non-fibre classes. We need to generate the features in a canonical form, which means that similar pixels have similar feature vectors. In this case we wish the feature vectors to be orientation invariant. This can be achieved by assigning the first sample of the vector to its maximum value, corresponding to the predominant orientation, and then circularly shifting all samples by this offset.

$$\mathbf{f} \triangleq \mathcal{F}(i,j) = \{\mathcal{G}_1(i,j), \mathcal{G}_2(i,j), \ldots, \mathcal{G}_{O\times S}(i,j)\} \tag{18}$$

$$\tau = t \mid \arg\max_{f_t}(\mathbf{f}) \tag{19}$$

$$\mathbf{f} \leftarrow \zeta(\mathbf{f} - \tau) \tag{20}$$

where ζ is the cyclic shift function; $\mathbf{f}$ is the pixel-wise feature vector of $\mathcal{F}$ at (i,j). τ is the number of shift cycles defined by the maximum value of the vector $\mathbf{f}$. This guarantees that responses of foreground and background models are canonically aligned in the newly formed feature vector and independent of the particular orientation of the models.

6. Nerve fibre classification

We consider three possible ways of using the feature vector $\mathcal{F}$ to assign pixels (i,j) to the foreground or background classes.

6.1. Maximum projection

One simple way of interpreting the feature vector of each CCM image is by considering the maximum value of a particular sample among all different frequencies and orientations. Following the cyclic shift the first feature in the feature vector $\mathcal{F}$ has the maximum value.

$$\mathbf{I}(i,j)_{Enh} = \begin{cases} f_t & \text{if } f_t \geqslant 0 | t = 0 \\ 0 & \text{Otherwise} \end{cases} \tag{21}$$

The scale and the orientation of this maximum value of $\mathbf{f}$ is taken to be the frequency and orientation at a particular pixel (i,j) of the detected nerve fibre in the enhanced image $\mathbf{I}_{Enh}$. Although this method is effective, efficient and easy to implement, it discards the rest of the sample responses at other orientations and scales, hence neglecting the possibility that combinations of features may be useful in correctly classifying pixels.

6.2. Scaled conjugate gradient neural network

We assign pixels to fibre or non-fibre classes by means of a multi-layer perceptron neural network trained using the conjugate gradient descent method. Conjugate gradient methods (CGM) (Fletcher, 1975) are general purpose second order techniques that help minimise functions of several variables using the second derivatives of the function. They generally find a better way to a minimum than a first order technique (such as standard backpropagation), by proceeding in the direction which is conjugate to the directions of the previous steps of the error function. Thus the minimisation performed in one step is not partially undone by the next, as is the case with standard backpropagation and other gradient descent methods. The traditional CGM uses the gradient to compute a search direction. It then uses a line search algorithm to find the optimal step size along a line in the search direction (Johansson et al., 1991).

The scaled conjugate gradient algorithm (SCG), developed by Moller (1993), was designed to avoid the time-consuming line search. This algorithm combines the model-trust region approach used in the Levenberg–Marquardt algorithm with the conjugate gradient approach in order to numerically estimate the second derivatives (Hessian matrix) and scale the step size. Its success in large-scale problems does not depend on the user dependent parameters *learning rate* and *momentum constant* as in Rumelhart et al. (1986). The number of input and output nodes are determined by the feature vector, while the number of hidden nodes was empirically set at 50 to represent the variation in the classification space and is defined during the learning procedure.

6.3. Random forest classifier

The random forests (RF) machine learning algorithm is a classifier (Breiman, 2001) that encompasses bagging (Breiman, 1996) and random decision forests (Amit and Geman, 1997; Ho, 1998). RF became popular due to its simplicity of training and tuning while offering a similar performance to boosting. It is a large collection of decorrelated decision trees, which are ideal candidates to capture complex interaction structures in data. RF is supposed to be resistant to over-fitting of data if individual trees are sufficiently deep.

Consider a RF collection of tree predictors $h(\mathbf{x};\psi_u)$, $u = 1,\ldots,U$, where $\mathbf{x}$ is a random sample of d-dimensions associated to random vector $\mathbf{X}$ and ψ_u independent identically distributed random vectors. Given a dataset of N samples, the bootstrap training sample of tree $h(\mathbf{x};\psi_u)$ is used to grow the tree by recursively selecting a subset of random dimensions $\hat{d}$ such that $\hat{d} \ll d$ and picking the best split of each node based on these variables. Unlike conventional decision trees, pruning is not required.

$$\hat{c} = \text{majority vote}\{C_u(\mathbf{x})\}_1^U \tag{22}$$

To make a prediction for a new sample $\mathbf{x}$, the trained RF could then be used for classification by majority vote among the trees of the RF as shown in Eq. (22), where $C_u(\mathbf{x})$ is the class prediction of the uth RF tree. The important parameters of the RF classifier were set as

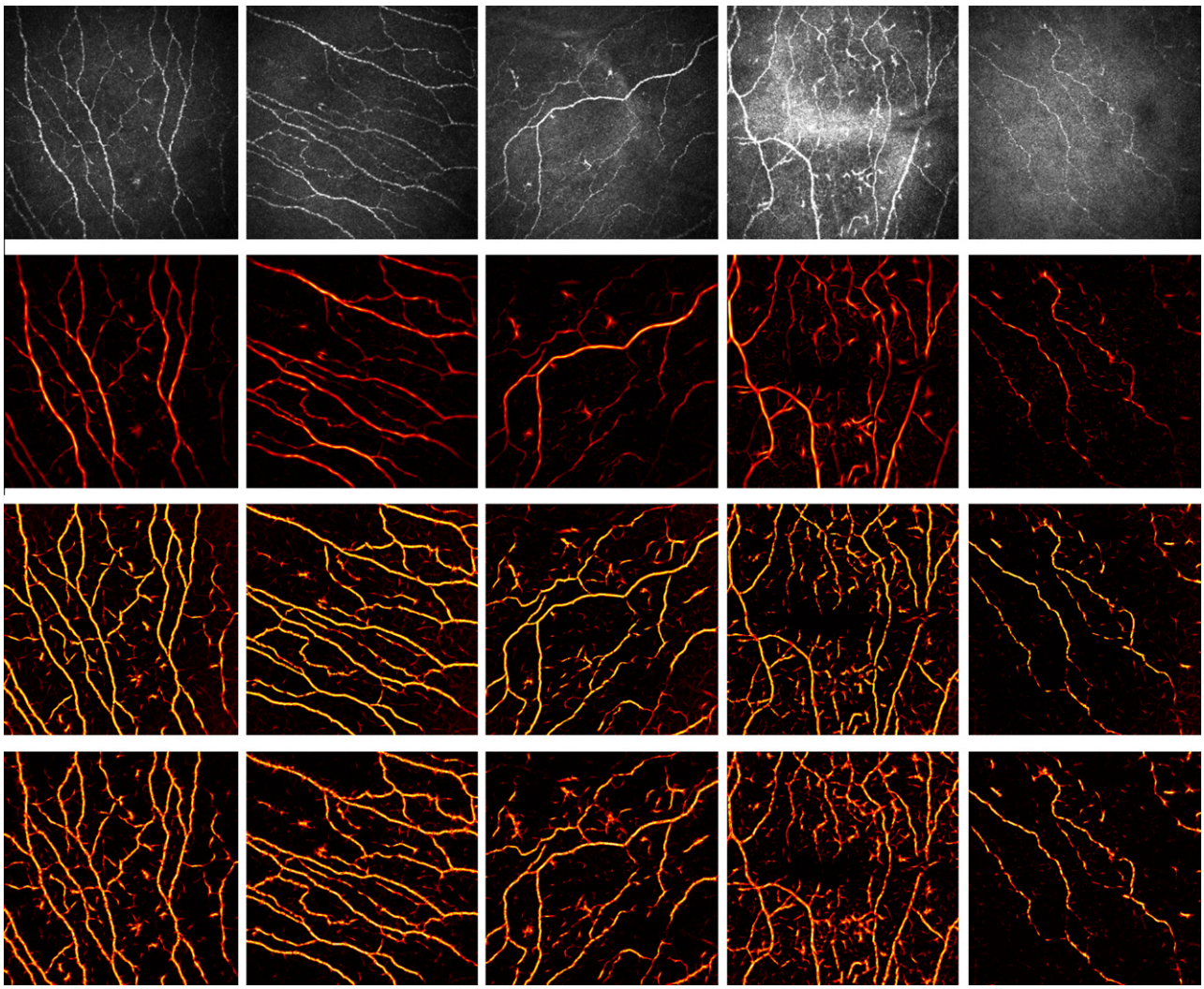


Fig. 4. An illustration of the multi-scale dual-model detection responses when using different pixel classification methods. The first row consists of the original CCM images. The following rows contain the response images when using maximum response, NNT and RF respectively. Responses are presented as heat maps, where brighter colours correspond to higher values. The best response is given when using NNT. The classifier successfully learnt the right balance of sensitivity and specificity (see Section 7). The RF has a far greater sensitivity than the maximum response but its higher sensitivity results in noisier detection. The improved response is most visible in regions of the image where the signal to noise ratio is low. The very bright region at the centre of the image in column 4 is an extreme example of a low-frequency illumination artifact. It is not clear visually whether any fibres are present there. The NNT and RF detectors identify more fibres with greater confidence.

follows in this case. The number of trees in the forest should be sufficiently large to ensure that each input class receives a number of predictions: set to 1000. The number of variables randomly sampled at each branch: set to 5.

7. Detection results and analysis

7.1. Database and experimental settings

The evaluation is conducted on a database of 521 CCM images captured using the HRT-III microscope from 68 subjects (20 controls and 48 diabetic patients). The images have a size of 384×384 pixels, 8-bit grey levels and are stored in BMP format. The resolution is 1.0417 μm and the field of view is $400 \times 400\ \mu m^2$ of the cornea. For each individual, several fields of view are selected manually from near the centre of the cornea that show recognisable nerve fibres. Other than the processing inherent in the filters (described above), no additional preprocessing was applied to the images.

Using the neuropathy disability score (NDS) (Abbott et al., 2002), the patients were categorised into four groups according to severity of neuropathy (non-neuropathic: $0 \geqslant NDS \leqslant 2 (n = 26)$, mild: $3 \geqslant NDS \leqslant 5 (n = 9)$, moderate: $6 \geqslant NDS \leqslant 8 (n = 10)$ and severe: $9 \geqslant NDS \leqslant 10 (n = 3)$).

7.2. Nerve fibre detection performance

The evaluation of detecting nerve fibres is conducted against ground-truth data which has been generated by clinical experts using *CCMetrics*.[2] Each nerve fibre and branch is traced to generate a single-pixel wide line along the fibre centre, from which the parameters NFL, NFD, NBD and tortuosity can be derived. In automatic detection, the response images are thresholded and then thinned to one-pixel wide lines. These lines are then compared pixel by pixel to the ground-truth, a true positive being scored if the detected pixel is within a three-pixel (3.14 μm) tolerance of ground truth and a false positive if it is outside this tolerance. The evaluation is quantified in terms of true-positive rate (TPR or sensitivity) and false-positive rate (FPR or 1-specificity) defined at the operational

[2] *CCMetrics* is a purpose built interactive graphical interface which helps experts to manually delineate nerve fibres in CCM images.

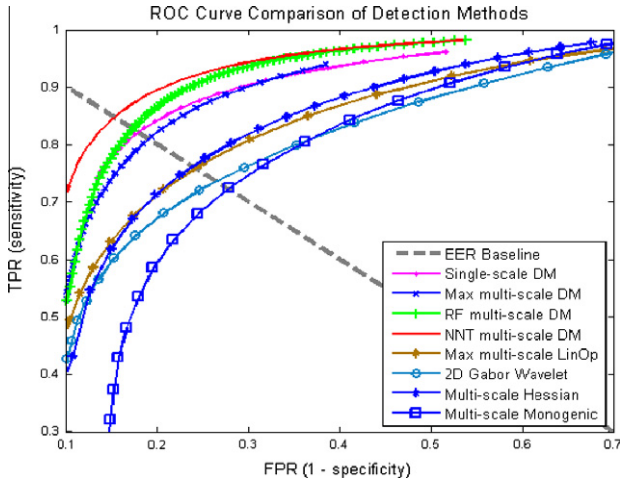


Fig. 5. ROC curves of nerve fibre detection for all different methods including the RF and NNT pixel classifiers of the multi-scale dual model. As shown the NNT has achieved the best performance followed by the RF classification. The single-scale dual-model algorithm has marginally outperformed the maximum response method.

point of the equal error rate (EER). The training of the NNT and RF was based on a single CCM image with a ground-truth delineation. Once the classifier is trained using this single image, it is applied on the entire database and the results are obtained through a comparison with the ground-truth delineation of every test image.

The single-scale methods (Gabor wavelet and single scale dual model) were evaluated against their single-scale response, while the multi-scale methods (LinOp, Hessian and Monogenic filters) were evaluated against their maximum response. Fig. 4 shows the response images in different CCM images arising from each of the three methods of pixel classification *i.e.* maximum response, NNT and RF. Both the RF and NNT classifiers are more sensitive than the maximum response method.

In our earlier study (Dabbah et al., 2010) we compared the single scale dual-model detector with the comparator methods described in Section 3, some of which are specifically designed to detect curvilinear structures, while others are more general feature detectors. In that comparison we used single-scale instantiations of all detectors, though some have multi-scale implementations. The dual model produced the best ROC curves and EER classification rates. Here we repeat the evaluation using multi-scale versions of all detectors. Fig. 5 shows the resulting ROC curves. The single-scale dual model detector is also included for comparison.

The single scale dual-model produces a better response than the multi-scale versions of the other methods. The maximum projection version of the multi-scale dual model produced slightly worse results than the single-scale version, while both the RF and NNT versions generated improved results, more so in the case of the NNT classifier.

This may be due to the fact that the orientation estimate in the single scale model are locally smoothed, whereas those in the multi-scale, maximum response, model are not, and therefore subject to noise variations. The NNT and RF classifications are less sensitive to noisy orientation estimates because all orientations across scale are contributing to the solution.

Due to the second order derivative components in the Hessian and the Monogenic methods their responses are very sensitive to the background noise. LinOp and the 2D Gabor methods, on the other hand were less sensitive to noise, but tended to include too much background.

Fig. 6 provides a visual illustration of the responses of several of the detectors in the comparison. Fig. 5 and Table 1 provide quanti-

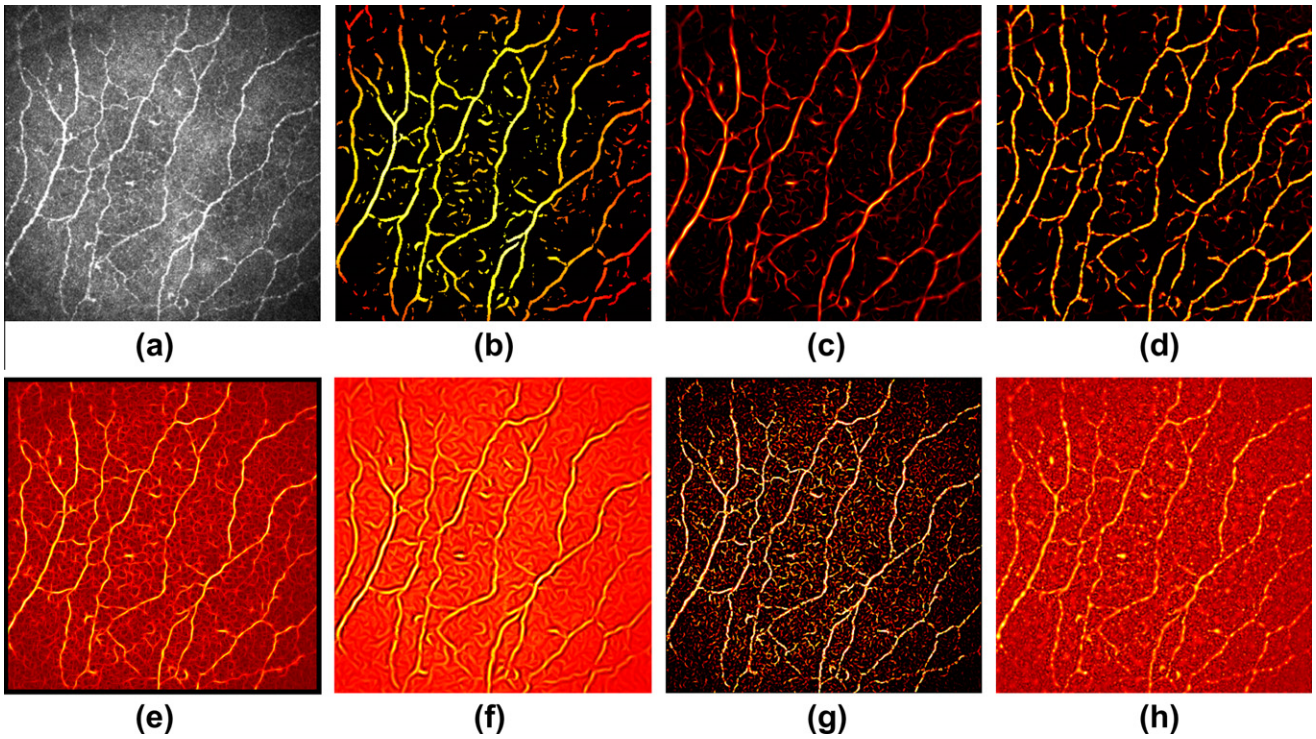


Fig. 6. A visual comparison between the responses of different detection methods for the original CCM image in (a). (b) is the single-scale dual-model response, (c) is the maximum response method for the multi-scale dual-model and (d) is its NNT counterpart. (e) is the LinOp response, (f) is the 2D Gabor Wavelet response, (g) the Hessian matrix response and (h) is the Monogenic signal response. The multi-scale dual model with NNT classification has the best performance followed by the single-scale dual-model. The Hessian and Monogenic responses suffer from a greater sensitivity to noise due the second derivative components in the algorithms. The LinOp and the 2D Gabor responses struggled to suppress the background. Responses are presented as heat maps, where brighter colours correspond to higher values.

tative confirmation of the qualitative results shown in Figs. 4 and 6. The maximum response output of the multi-scale dual model achieves superior performance to the maximum response outputs of the Hessian and Monogenic filters, and matches the performance of the multi-scale LinOp. The multi-scale dual model using NNT pixel classification achieves the highest performance in detecting nerve fibres. It achieves highest sensitivity and specificity at the EER of 15.44%. We did not set out to conduct a comparison between the two classifiers used, rather to show that the classification method is capable of producing useful results. Using the particular (empirically set) parameters for these classifiers and this data set, the RF is more sensitive than the NNT, resulting in a noisier response (Fig. 4). The measured error rates for the NNT classifier shown in Table 1 (significant at the $p < 0.05$ level) emphasise the superior performance achieved by the NNT classifier, here.

7.3. Clinical utility using nerve-fibre length analysis

In studies using interactive measurements, nerve-fibre length (NFL), was shown to be the most sensitive of the CCM metrics to the presence of neuropathy as assessed by the current clinical techniques. Hence it is also used here to evaluate the similarity of the automatic analysis to the manual analysis. Fig. 7 shows the distribution of NFL measurements in NDS groups made interactively by experts (a) and automatically (b). The manually and automatically generated NFL distributions are very similar and strongly correlated (Fig. 7c) with $r = 0.95$. They are both statistically significant in separating between the NDS groups: for the manual analysis ($p = 0.03 \times 10^{-6}$), while the automatic has ($p = 0.68 \times 10^{-6}$). However as shown in the scatter plot Fig. 7c this statistical significance is not enough for classification of individual cases due to the overlapped distributions. This could be as a result of the limitation in using the NDS score, which is used as a diagnostic score and unstable for individual analysis. This result however could be improved by utilising the potential of the automatic analysis in utilising further metrics such as nerve fibre width.

8. Conclusion

The analysis of CCM images requires the identification of fibre-like structures with low contrast in noisy images. This is a requirement shared by a number of imaging applications in biology, medicine and other fields, and a number of methods have been developed and used in these various applications. In the present work we present a new multi-scale dual-model method to detect corneal nerve fibres in CCM images and we compare this with some more generic methods. In our evaluation the multi-scale dual-model with the NNT pixel classification has outperformed all other methods and obtained the lowest EER at 15.44%. A point worth noting is that the additional performance was achieved at the expense of a very small training burden. A single annotated image was used to provide training data for both the RF and NNT classifiers. This is a potentially important issue in the practical implementation of the method.

The clinical utility of the method was also evaluated by comparing our automatic detection against expert manual annotation of the images. We demonstrate equivalent results with the manual analysis which has previously demonstrated encouraging clinical performance for the stratification of neuropathic severity. Here we have used the NDS score, which is widely used clinically and is adequate for defining the clinical severity of neuropathy to assess the correspondence between manual and automatic detection. However, the NDS may not be adequate for a thorough assessment of clinical utility because it does not detect small fibre damage. Hence as CCM can evaluate small fibre damage, any assessment of the clinical utility of this test may be limited. As noted in Section 2.1, the accepted gold standard for defining small fibre pathology can only be achieved by either nerve or skin biopsy, both of which are invasive and highly labour-intensive assessments. We are currently collecting a data set that will enable us to evaluate the CCM metrics with measures of loss of nerve fibres in skin biopsies.

In conclusion the automated analysis produces equivalent results to manual analysis, while being a quicker and potentially

Table 1
A comparison of mean EER, its standard deviations, TPR (sensitivity) and FPR (1-specificity) of all detection methods. The table clearly shows that the multi-scale dual model with NNT classification results in the lowest error rate.

	Max	NNT	RF	Dual model	LinOp	2D Gabor	Hessian	Monogenic
EER (μ)	0.2056	0.1544	0.1746	0.1779	0.2265	0.2415	0.2314	0.2650
EER (σ)	0.1806	0.1083	0.1176	0.1058	0.1076	0.1074	0.1153	0.1258
TPR (sensitivity)	0.8135	0.8478	0.8290	0.8172	0.766	0.7212	0.7773	0.7240
FPR (1-specificity)	0.1940	0.1533	0.1747	0.1758	0.2489	0.2467	0.2527	0.2782

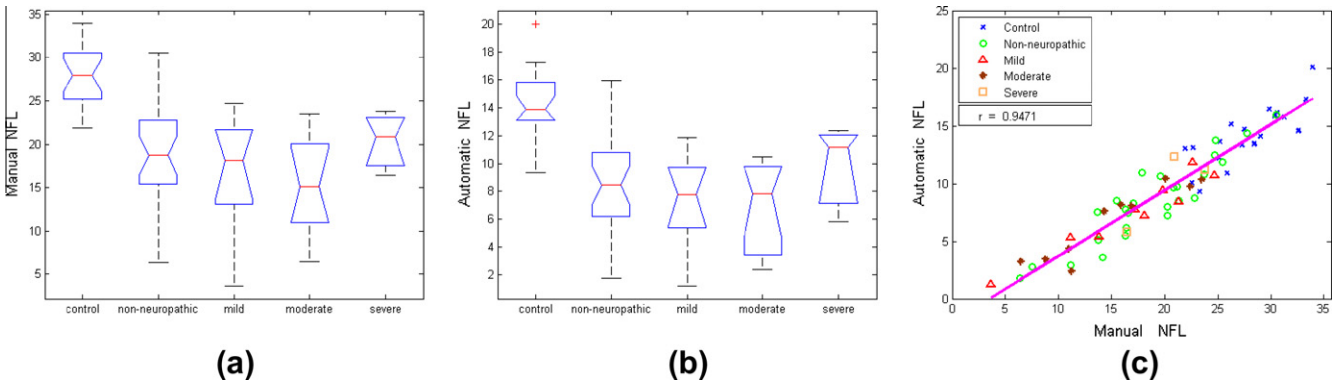


Fig. 7. A comparison between the manually and automatically obtained NFL in groups of different severity of neuropathy, as judged by NDS score. The manual (a) and automatic (b) boxplots show strong similarity. Both are statistically significant ($p \approx 0$) in separating the NDS groups detailed in Section 7.1. The scatter plot in (c) shows the strong correlation between them ($r = 0.95$) and demonstrates the overlap between the groups according to the NDS categories.

more reliable and practical alternative due to its consistency and immunity to inter/intra-observer variability. The multi-scale detection method used here could, of course, be applied in other contexts as the detection of curvilinear structures is a requirement in a number of applications. The method is generic, requiring only the establishment of appropriate parameters for λ, σ_x and σ_y at the resolution of the original image. The empirical values used in this application are quoted in Section 5.1.

Acknowledgements

This work is supported by a Juvenile Diabetes Research Foundation (JDRF) scholar Grant 17-2008-1031 and subject to patent application (UK Patent Application No. 1005905.3).

References

Abbott, C.A., Carrington, A.L., Ashe, H., Bath, S., Every, L.C., Griffiths, J., Hann, A.W., Hussein, A., Jackson, N., Johnson, K.E., Ryder, C.H., Torkington, R., Ross, E.R.E.V., Whalley, A.M., Widdows, P., Williamson, S., Boulton, A.J.M., 2002. The north-west diabetes foot care study: incidence of, and risk factors for, new diabetic foot ulceration in a community-based patient cohort. Diabetic Medicine 19 (5), 377–384.

Ali, R., Gooding, M., Christlieb, M., Brady, M., 2008. Advanced phase-based segmentation of multiple cells from brightfield microscopy images. In: The IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI'08), pp. 181–184.

Amit, Y., Geman, D., 1997. Shape quantization and recognition with randomized trees. Neural Computation 9, 1545–1588.

Boulton, A.J., 2005. Management of diabetic peripheral neuropathy. Clinical Diabetes 23 (1), 9–15.

Breiman, L., 1996. Bagging predictors. Machine Learning 24, 123–140.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Dabbah, M.A., Graham, J., Tavakoli, M., Petropoulos, Y., Malik, R.A., 2009. Nerve fibre extraction in confocal corneal microscopy images for human diabetic neuropathy detection using Gabor filters. In: Medical Image Understanding and Analysis (MIUA), pp. 254–258.

Dabbah, M.A., Graham, J., Tavakoli, M., Petropoulos, Y., Malik, R.A., 2010. Dual-model automatic detection of nerve-fibres in corneal confocal microscopy images. The International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 6361, 300–307.

Daousi, C., MacFarlane, I.A., Woodward, A., Nurmikko, T.J., Bundred, P.E., Benbow, S.J., 2004. Chronic painful peripheral neuropathy in an urban community: a controlled comparison of people with and without diabetes. Diabetic Medicine 21 (9), 976–982.

Daugman, J.G., 1980. Two-dimensional spectral analysis of cortical receptive field profiles. Vision Research 20 (10), 847–856.

DiabetesUK, 2010. Diabetes in the UK 2010: key statistics on diabetes. <http://www.diabetes.org.uk/>.

Dixon, R.N., Taylor, C.J., 1979. Automated asbestos fibre counting. In: Machine Aided Image Analysis. Institute of Physics, London, pp. 178–185.

Dyck, P.J., Norell, J.E., Tritschler, H., Schuette, K., Samigullin, R., Ziegler, D., Bastyr, E.J., Litchy, W.J., O'Brien, P.C., 2007. Challenges in design of multicenter trials: endpoints assessed longitudinally for change and monotonicity. Diabetes Care 30, 2619–2625.

Felsberg, M., Sommer, G., 2001. The monogenic signal. IEEE Transactions on Signal Processing 49 (12), 3136–3144.

Fletcher, R., 1975. Practical Methods of Optimization. John Wiley & Sons.

Frangi, A.F., Niessen, W.J., Vincken, K.L., Viergever, M.A., 1998. Multiscale vessel enhancement filtering. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 130–137.

Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (8), 832–844.

Hong, L., Wan, Y., Jain, A., 1998. Fingerprint image enhancement: algorithm and performance evaluation. IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (8), 777–789.

Hossain, P., Sachdev, A., Malik, R.A., 2005. Early detection of diabetic peripheral neuropathy with corneal confocal microscopy. The Lancet 366 (9494), 1340–1343.

Jain, A.K., Farrokhnia, F., 1991. Unsupervised texture segmentation using Gabor filters. Pattern Recognition 24 (12), 1167–1186.

Johansson, E.M., Dowla, F.U., Goodman, D.M., 1991. Back-propagation learning from multi-layer feed-forward neural networks using the conjugate gradient method. International Journal of Neural Systems 2 (4), 291–302.

Kallinikos, P., Berbanu, M., O'Donnell, C., Boulton, A., Efron, N., Malik, R., 2004. Corneal nerve tortuosity in diabetic patients with neuropathy. Investigative Ophthalmology & Visual Science 45 (2), 418–422.

Loseth, S., Stalberg, E., Jorde, R., Mellgren, S., 2008. Early diabetic neuropathy: thermal thresholds and intraepidermal nerve fibre density in patients with normal nerve conduction studies. Journal Neurology 255, 1197–1202.

Malik, R., Tesfaye, S., Newrick, P., Walker, D., Rajbhandari, S., Siddique, I., Sharma, A., Boulton, A., King, R., Thomas, P., Ward, J., 2005. Sural nerve pathology in diabetic patients with minimal but progressive neuropathy. Diabetologia 48, 578–585.

Malik, R., Veves, A., Walker, D., Siddique, I., Lye, R., Schady, W., Boulton, A., 2001. Sural nerve fibre pathology in diabetic patients with mild neuropathy: relationship to pain, quantitative sensory testing and peripheral nerve electrophysiology. Acta Neuropathologica 101, 367–374.

Malik, R.A., Kallinikos, P., Abbott, C.A., van Schie, C.H.M., Morgan, P., Efron, N., Boulton, A.J.M., 2003. Corneal confocal microscopy: a non-invasive surrogate of nerve fibre damage and repair in diabetic patients. Diabetologia 46 (5), 683–688.

Mehra, S., Tavakoli, M., Kallinikos, P.A., Efron, N., Boulton, A.J.M., Augustine, T., Malik, R.A., 2007. Corneal confocal microscopy detects early nerve regeneration after pancreas transplantation in patients with type 1 diabetes. Diabetes Care 30 (10), 2608–2612.

Moller, M.F., 1993. A scaled conjugate gradient algorithm for fast supervised learning. Neural Networks 6, 525–533.

Novella, S., Inzucchi, S., Goldstein, J., 2001. The frequency of undiagnosed diabetes and impaired glucose tolerance in patients with idiopathic sensory neuropathy. Muscle Nerve 24, 1229–1231.

Pan, X.-B., Brady, M., Highnam, R., Declerck, J., 2006. The use of multi-scale monogenic signal on structure orientation identification and segmentation. In: Astley, S., Brady, M., Rose, C., Zwiggelaar, R. (Eds.), Digital mammography, Lecture Notes in Computer Science, vol. 4046. Springer, Berlin/Heidelberg, pp. 601–608.

Quattrini, C., Tavakoli, M., Jeziorska, M., Kallinikos, P., Tesfaye, S., Finnigan, J., Marshall, A., Boulton, A.J.M., Efron, N., Malik, R.A., 2007. Surrogate markers of small fiber damage in human diabetic neuropathy. Diabetes 56 (8), 2148–2154.

Rao, A.R., 1990. A Taxonomy for Texture Description and Identification. Springer-Verlag, New York, Inc., New York, USA.

Ruggeri, A., Scarpa, F., Grisan, E., 2006. Analysis of corneal images for the recognition of nerve structures. In: IEEE Conference of the Engineering in Medicine and Biology Society (EMBS), pp. 4739–4742.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning internal representations by error propagation. In: Rumelhart, D.E., McClelland, J.L. (Eds.), Parallel Distributed Processing: Exploration in the Microstructure of Cognition. MIT Press, pp. 318–362.

Scarpa, F., Grisan, E., Ruggeri, A., 2008. Automatic recognition of corneal nerve structures in images from confocal microscopy. Investigative Ophthalmology and Visual Science (49), 4801–4807.

Singleton, J., Smith, A., Bromberg, M., 2001. Increased prevalence of impaired glucose tolerance in patients with painful sensory neuropathy. Diabetes Care 24, 1448–1453.

Sumner, C., Sheth, S., Griffin, J., Cornblath, D., Polydefkis, M., 2003. The spectrum of neuropathy in diabetes and impaired glucose tolerance. Neurology 60, 108–111.

Tavakoli, M., Kallinikos, P.A., Efron, N., Boulton, A.J.M., Malik, R.A., 2007. Corneal sensitivity is reduced and relates to the severity of neuropathy in patients with diabetes. Diabetes Care 30 (7), 1895–1897.

Tavakoli, M., Marshall, A., Pitceathly, R., Fadavi, H., Gow, D., Roberts, M.E., Efron, N., Boulton, A.J., Malik, R.A., 2010a. Corneal confocal microscopy: a novel means to detect nerve fibre damage in idiopathic small fibre neuropathy. Experimental Neurology 223 (1), 245–250.

Tavakoli, M., Marshall, A., Thompson, L., Kenny, M., Waldek, S., Efron, N., Malik, R.A., 2009. Corneal confocal microscopy: a novel noninvasive means to diagnose neuropathy in patients with fabry disease. Muscle & Nerve 40 (6), 976–984.

Tavakoli, M., Quattrini, C., Abbott, C., Kallinikos, P., Marshall, A., Finnigan, J., Morgan, P., Efron, N., Boulton, A.J.M., Malik, R.A., 2010b. Corneal confocal microscopy: a novel non-invasive test to diagnose and stratify the severity of human diabetic neuropathy. Diabetes Care 33, 1792–1797.

Tesfaye, S., Boulton, A.J., Dyck, P.J., Freeman, R., Horowitz, M., Kempler, P., Lauria, G., Malik, R.A., Spallone, V., Vinik, A., Bernardi, L., Valensi, P., on behalf of the Toronto Diabetic Neuropathy Expert Group, 2010. Diabetic neuropathies: update on definitions, diagnostic criteria, estimation of severity, and treatments. Diabetes Care 33 (10), 2285–2293.

Umapathi, T., Tan, W., Loke, S., Soon, P., Tavintharan, S., Chan, Y., 2007. Intraepidermal nerve fiber density as a marker of early diabetic neuropathy. Muscle Nerve 35, 591–598.

Zhang, T.Y., Suen, C.Y., 1984. A fast parallel algorithm for thinning digital patterns. Communications of the ACM 27 (3), 236–239.

Zwiggelaar, R., Astley, S., Boggis, C., Taylor, C., 2004. Linear structures in mammographic images: detection and classification. IEEE Transactions on Medical Imaging 23 (9), 1077–1086.

54. **An automatic tool for quantification of nerve fibres in corneal confocal microscopy images.** X. Chen, J Graham, M.A. Dabbah, I.N. Petropoulos, M. Tavokoli, R.A. Malik, *IEEE Trans. Biomedical Engineering (in press).*

    This paper is accepted for publication in IEEE Transactions on Biomedical Engineering. The preprint included here appears on the IEEE explore website (http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=7484747&filter%3DAND%28p_IS_Number%3A4359967%29)
    at the time of writing and was downloaded from there.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2016.2573642, IEEE Transactions on Biomedical Engineering

An Automatic Tool for Quantification of Nerve Fibres in Corneal Confocal Microscopy Images

Xin Chen, Jim Graham, Mohammad A. Dabbah, Ioannis N. Petropoulos, Mitra Tavakoli, Rayaz A. Malik

***Abstract*— *Objective*: We describe and evaluate an automated software tool for nerve fibre detection and quantification in corneal confocal microscopy (CCM) images, combining sensitive nerve-fibre detection with morphological descriptors. *Method*: We have evaluated the tool for quantification of Diabetic Sensorimotor Polyneuropathy (DSPN) using both new and previously published morphological features. The evaluation used 888 images from 176 subjects (84 controls and 92 patients with Type 1 diabetes). The patient group was further subdivided into those with (n=63) and without (n=29) DSPN. *Results*: We achieve improved nerve-fibre detection over previous results (91.7% sensitivity and specificity in identifying nerve-fibre pixels). Automatic quantification of nerve morphology shows a high correlation with previously reported, manually measured, features. ROC analysis of both manual and automatic measurement regimes resulted in similar results in distinguishing patients with DSPN from those without: AUC of about 0.77 and 72% sensitivity-specificity at the equal error rate point. *Conclusion*: Automated quantification of corneal nerves in CCM images provides a sensitive tool for identification of DSPN. Its performance is equivalent to manual quantification, while improving speed and repeatability. *Significance*: Corneal confocal microscopy is a novel in-vivo imaging modality that has the potential to be a non-invasive and objective image biomarker for peripheral neuropathy. Automatic quantification of nerve morphology is a major step forward in the early diagnosis and assessment of progression, and, in particular, for use in clinical trials to establish therapeutic benefit in diabetic and other peripheral neuropathies.**

***Index Terms*— Diabetic Sensorimotor Polyneuropathy, Computer Aided Diagnosis, Corneal Confocal Microscopy, Image Analysis, Nerve Fibre Quantification**

I. Introduction

DIABETIC sensorimotor polyneuropathy (DSPN) is one of most common long term complications of diabetes. Up to 50% of diabetic patients suffer from it [1], and it is estimated that about one in six diabetic patients have chronic painful neuropathy [2]. Several methods are currently used to quantify neuropathy, including clinical scoring of symptoms, quantitative sensory testing, nerve conduction measurements and microscopic measurement of intra-epidermal nerve-fibre density (IENFD) in skin biopsy samples. These methods have their advantages and limitations. Thus, whilst symptoms and signs are directly relevant to the patient and are easily recorded, they are subjective resulting in poor repeatability [3]. Neurophysiology is more objective; however it only assesses large fibres, which constitute a tiny proportion of all the nerve fibres present in a peripheral nerve and has also been shown to have limited reproducibility [4]. The quantification of IENFD in skin biopsies is objective, but is clearly invasive and requires considerable expertise in assessment. There is a need for a rapid, non-invasive assessment that is truly quantitative and assesses small nerve fibres, which are more likely to be involved in neuropathy [5, 6].

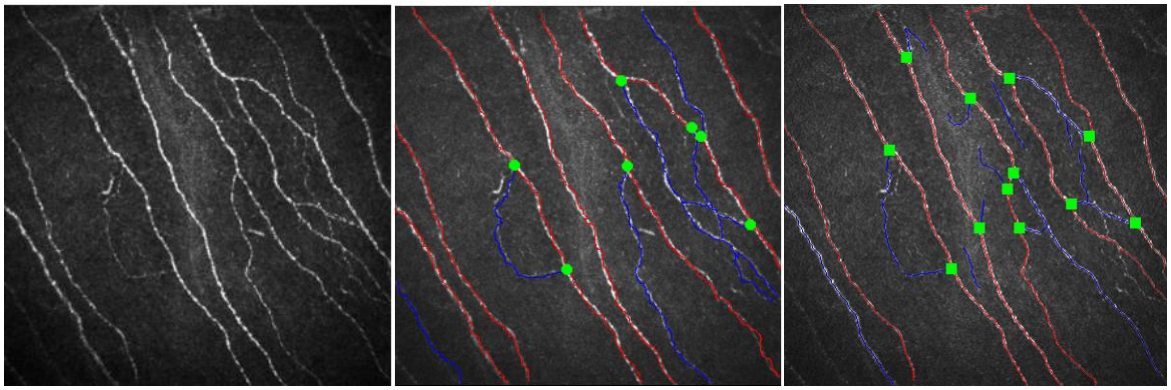
Corneal confocal microscopy (CCM) images of nerve fibres are captured from the sub-basal plexus immediately above Bowman's membrane of the cornea by an in-vivo laser confocal microscope. Fig. 1a shows an example image. One of the advantages of CCM is the entirely non-invasive and relatively rapid (about 2 minutes) acquisition of images of small nerve fibres and other corneal structures. Clinical studies [7] have shown that CCM is capable of making quantitative assessment of DSPN and has the potential to be an ideal surrogate endpoint. It has also recently been shown to have a predictive ability in identifying diabetic patients at risk of developing DSPN [8] and has been used in several clinical intervention studies showing nerve-fibre repair [9-11]. Interactive analysis has been used to derive measurements from these images, such as corneal nerve-fibre length (CNFL), corneal nerve-fibre density (CNFD) and corneal nerve branch density (CNBD) [12, 13] (Fig. 1). CNFL is defined as the total length of all nerve fibres visible in the CCM image per square millimetre. CNFD and CNBD are the number of the major nerves (red lines in Fig. 1b) per square millimetre and the number of primary branches emanating from those major nerve trunks (green dots in Fig. 1b) per square millimetre respectively. Although an association has been demonstrated between these quantitative features and the severity of DSPN [7] in cross sectional studies, the manual analysis suffers from the usual problems of being labour-intensive and subjective and therefore raises considerable difficulties, particularly when undertaking longitudinal follow-up studies [14].

- *J. Graham is with the Centre for Imaging Sciences, the University of Manchester, UK. E-mail: jim.graham@manchester.ac.uk*
- *X. Chen is now at the Division of Imaging Sciences and Biomedical Engineering, Kings College London, UK. E-mail: xin.chen@kcl.ac.uk*
- *M. Dabbah is currently at Roke Manor Research Ltd. Romsey, UK*
- *I. Petropoulos, M. Tavakoli and Rayaz Malik are with Centre for Endocrinology & Diabetes, Institute of Human Development, Manchester, UK. I. Petropoulos and Rayaz Malik are also with Weill Cornell Medical College in Qatar, Division of Medicine, Doha, Qatar.*

Copyright (c) 2016 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

0018-9294 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2016.2573642, IEEE Transactions on Biomedical Engineering



(a) (b) (c)

Fig. 1. (a) Original CCM image. (b) Manually quantified CCM image. (c) Automatically quantified CCM image. Red lines represent main nerve fibres, blue lines are branches and green spots indicate branch points on the main nerve trunks. Refer to online coloured version.

Consequently the quantification results show poor reproducibility, especially in CNBD [15]. For the technology to be clinically useful, the analysis of images needs to be done automatically.

Here we describe a fully automatic nerve-fibre detection and quantification system. Fig. 1a indicates that the appearance of nerve fibres in CCM images covers a wide contrast range, with some fibres appearing very faint on a noisy background, whilst other, larger, fibres show strong contrast. A number of studies have presented methods of detecting similar linear structures in different types of images e.g. the detection of blood vessels in retinal images [16], and the detection of curvilinear structure in mammograms [17]. Previous studies aimed at automatic fibre detection in CCM images include Scarpa et al. [18] who described a method for tracing nerve fibres based on automatically initialised seed points, and Holmes et al. [19]who identified fibres based on ridge points. Sindt et al. [20] detected several types of objects visible in CCM images, including dendritic immune cells and wing cells in addition to nerve fibres. Dabbah et al. [21] presented a method of fibre detection based on a multi-scale Gabor filter with responses trained using a neural network. The best detection performances in various applications are achieved using methods based on machine learning, in which features are derived from training images [16, 17, 21].

Following fibre detection, it is required to extract individual fibres, identify branches and quantify appropriate features for classification. A number of studies have investigated the quantification of a variety of image features, describing the morphology of nerve fibres delineated either manually or automatically [13, 19-21]. These studies have shown the relationship between several features, including those listed above, and neuropathic status. None of them, however, has addressed the question of diagnosis of individual subjects.

We have previously described our image filter for enhancing nerve-fibre pixels [21] and reported clinical results of applying this system to DSPN [22]. This paper describes the development of the fibre detection method into a complete tool for measurement of nerve-fibre morphology to act as a diagnostic aid, making three specific contributions over our earlier publications: (1) we compare our fibre detector [21] with another, successful, linear feature descriptor and demonstrate the best reported performance in detecting nerve-fibre pixels in CCM images. (2) The detailed algorithms for quantification of morphometric features are presented for the first time (only CNFL was used in [21]), including the established features (CNFD, CNFL, CNBD) and new features: Corneal Nerve-Fibre Width Histogram (CNFWH) and Corneal Nerve-Fibre Orientation Histogram (CNFOH). (3) Finally, we report a technical validation of the proposed system based on CCM images obtained from 84 control subjects and 92 type 1 diabetic patients, which, to our knowledge, is the largest dataset in the literature for DSPN diagnosis of individuals.

II. Methods

A. CCM Images and Manual Measurement

CCM images (Fig. 1(a)) were captured from all participants using the Heidelberg Retina Tomograph Rostock Cornea Module (HRT-III) as described in [13]. The image dimensions are 384×384 pixels with the pixel size of 1.0417μm. During the CCM scan, images captured from all corneal layers and six sub-basal images from the right and left eyes were selected for analysis. Criteria for image selection were depth, focus position and contrast. A single experienced examiner, masked from the outcome of the medical and peripheral neuropathy assessment, manually quantified images of all study participants using purpose-written proprietary software (CCMetrics: M. A. Dabbah, Imaging Science, University of Manchester) to delineate main fibres, branch fibres and branch points (red lines, blue lines and green dots respectively in Fig. 1b). The reproducibility and reliability of manual annotation are reported in [15].The specific parameters measured in each frame were ENREF_18: CNFD, CNFL and CNBD, as described in section I in accordance with our previously published protocol [13].

B. Automated CCM Measurement

The automated CCM measurement process consists of two main steps: nerve-fibre detection and nerve-fibre quantification.

1) Nerve-Fibre Detection:

In this and similar applications [16, 17], methods based on machine learning have been reported to outperform others in detection of curvilinear features. The machine learning method normally consists of two key elements, feature description and classifier training on a set of samples.

0018-9294 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2016.2573642, IEEE Transactions on Biomedical Engineering

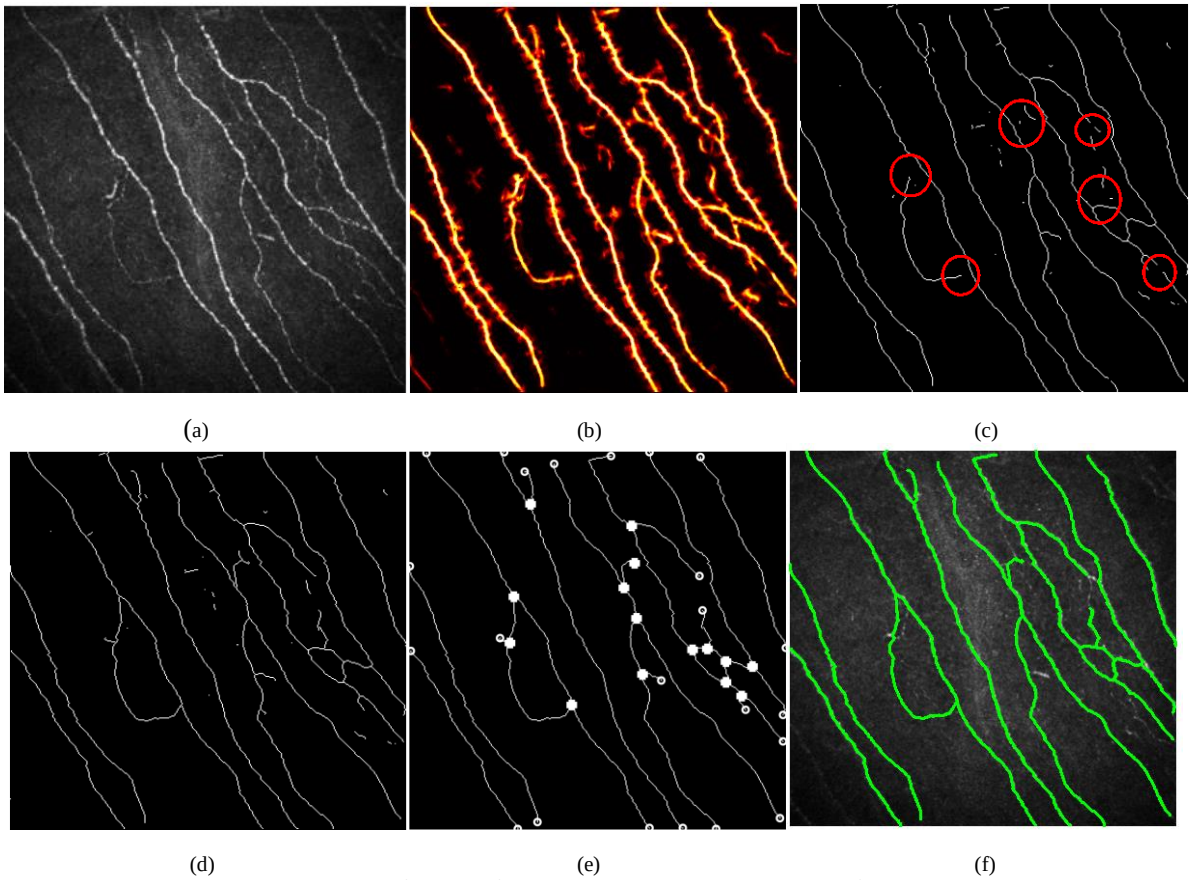


Fig. 2. (a) Original CCM image (b) Response image after nerve-fibre detection and denoising (c) Nerve-fibre skeleton with highlighted weak connection segments (d) Nerve-fibre skeleton after assessment of weak connections. (e) Automatically detected end points (hollow circles) and intersection points (solid circles). (f) Final detected nerve fibres.

For the feature description process, we have implemented and adapted two of the most successful methods [17, 21] for representing curvilinear structures. Dabbah et al. [21] proposed a multi-scale "dual-model filter" (DMF) that combines a foreground model based on a Gabor wavelet with a Gaussian background model that scales the output according to the level of noise. In our implementation, we apply the DMF at eight orientations (suggested in [21]) and at four levels of an image pyramid. Each level is a down-sampled (with smoothing) version of its immediate higher level by a factor of 2. The Gabor wavelet and Gaussian filter covered orientations from 0° to 180° and a range of fibre widths that we found to be sufficient for the CCM images in our study. The DMF method results in 32-dimensional vectors (8 orientations × 4 scale pyramid levels) to describe features at each pixel location.

Berks et al. [17] described a system that used the dual-tree complex wavelet transform (DWT) [23] for detection of linear structures in mammograms. The DWT combines the outputs of two discrete transforms, using real wavelets differing in phase by 90°, to form the real and imaginary parts of complex coefficients. It provides a directionally selective representation with approximately shift-invariant coefficient magnitudes and local phase information. As in the DMF method, the DWT is applied to a four-level image pyramid. Additionally, the DWT is performed at six different orientations (±15°, ±45°, ±75°, used in [17]) at each pyramid level. The six sub-bands are then multiplied by {i, -i, i, -1, 1, -1} respectively, so that the phase at the centre of the impulse response of each wavelet is zero. Finally, to achieve 180° rotational symmetry, any coefficient with negative imaginary part is replaced with its complex conjugate. Hence from coarse level to fine level of the image pyramid, the DWT results in a 48-element feature vector (4 level image pyramid × 6 orientation × 2 magnitude and phase) for each selected pixel location. Both of these detectors outperformed competitors in their respective domains.

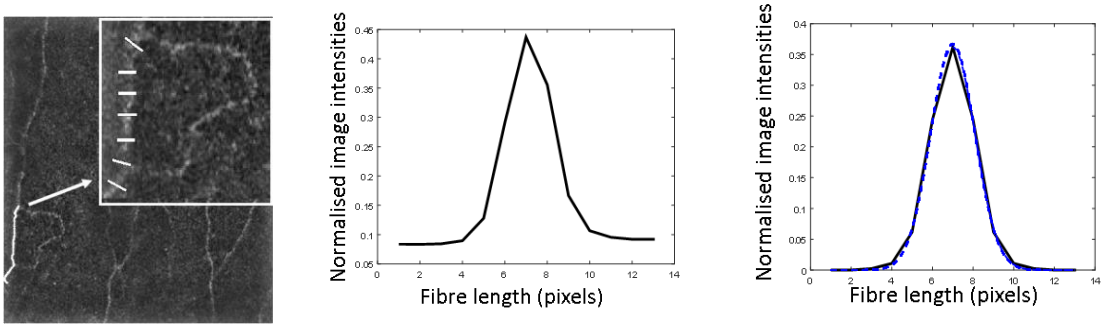
In this study we have implemented both detectors in the form proposed by the original authors (number of pyramid levels and orientations) as these produced feature vectors of similar dimension. We then subjected them to a comparative analysis in detecting nerve fibres.

For classifier training, the feature descriptors and their corresponding fibre/non-fibre labels from a set of training samples were used as the inputs to a classifier, which took the form of either a neural network or random forest [24]. The trained classification model was then used for classifying fibre/non-fibre pixels in unseen CCM images.

CCM images have a fairly high level of background noise (see Fig. 2a), which at a fine scale have similar contrast to nerve fibres at random orientations. These may be detected by the trained detectors and are removed by a further denoising step, which iteratively diminishes pixels that are not consistent with the dominant direction over a localised region. The output response image after denoising is shown in Fig. 2b. The

0018-9294 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2016.2573642, IEEE Transactions on Biomedical Engineering



(a) (b) (c)

Fig. 3. (a) Original CCM image with a highlighted segment, a selection of orthogonal profile lines are indicated on the enlarged inset. Profiles are calculated at each pixel along the segment. (b) Average of all the profile lines along the whole fibre segment. (c) The symmetric profile of (b) is firstly calculated, and then normalised (Solid line). A Gaussian distribution is fitted for nerve-fibre width estimation (broken line). The final width equals 2.5 times the RMS width (σ) of the fitted Gaussian curve.

evaluation and comparison of different combinations of the two feature descriptors and the two classifiers, before and after denoising, are presented in section III.

Based on the denoised image, a threshold is then applied to generate a binary image. The optimum threshold value is determined by the training and validation experiments described in section III-A. The binary image is then filtered by morphological operators to fill small gaps within nerve fibres and link adjacent structures. The binary structures are thinned to obtain a one-pixel wide skeleton (Fig. 2c). Branch and end points, identified by counting the neighbours of each skeleton point, are each assigned a unique label. For some regions, the evidence for nerve fibres is too weak (as highlighted in Fig. 2c) to be detected by a global threshold. However, the undetected pixels may be important in determining the nerve-fibre connectivity. Hence, for each end point, we extrude 30 pixels along the fibre orientations. The orientation of nerve fibres at each pixel location can be estimated using the second eigenvalue of the Hessian matrix of the response image. If an intersection with another fibre is detected and the average probability from the response image of the extruded pixels is sufficiently high (> 0.2), the extruded line is retained, otherwise it is eliminated (Fig. 2d). Subsequently, independent small segments and short branches that are less than 15 pixels long are removed, and the intersection points (solid circles) and end points (hollow circles) are calculated again as shown in Fig. 2e. The final binary skeleton, as shown in Fig. 2f, is used for total nerve-fibre quantification, described in the next section.

2) Nerve-Fibre Quantification

Fig. 2f shows that the output of fibre detection consists of several networks of interconnected line segments. In order to produce similar results to the manual CNFD, CNFL and CNBD, it is important to identify the main fibres within these networks and the branch points along the main fibres. To connect the appropriate fibre segments together, we generate four N × N matrices (MI, ML, MW and MO) to store the fibre intensity, fibre length, fibre width (described later in this section) and fibre orientation information respectively for each segment. N is the total number of branch and end points. If the i^{th} and j^{th} end/branch points are connected by a segment, the intensity, width, length and orientation information will be saved at the [i, j] location of the corresponding matrices; if they are not connected, these elements are zero. The matrices of intensity (MI), length (ML) and width (MW) are symmetric, as the elements at [i, j] and [j, i] should be identical. The [i, j] and [j, i] elements in the orientation matrix MO represent the respective orientations of the opposite ends of the fibre segment.

Identification of the main nerve fibres starts with the most prominent segments: those with greatest length and width. These are identified by multiplying the corresponding elements of MW and ML to produce a new matrix MA. The segments are considered in sequence according to the corresponding values of MA in descending order. There are normally two candidate segments that intersect with the current segment at a branch point. The candidate segments are ranked for the length, orientation difference, intensity and width parameters respectively. The candidate with the highest summed rank is chosen to connect with the current segment. The process continues till an end point is reached. The relevant entry in MA is set to zero and the process continues until no non-zero elements remain in MA. Finally, a list of connected fibres is obtained. Only the fibres with length greater than a threshold are kept as the main fibres. Fig. 1b and 1c respectively show the manual and automatic quantification results of the CCM image in Fig. 1a. The red lines show the principal nerve fibres, which are counted to produce CNFD. The blue lines indicate the secondary nerve fibres, which together with the principal fibres make up CNFL. The green dots are the branch points from the main fibres that are used for CNBD calculation.

Besides the CNFD, CNFL and CNBD features that are readily measured in the manual analysis, automatic quantification is able to calculate additional features. These additional CCM features include the total corneal nerve-fibre area per mm^2 (CNFA), the corneal nerve-fibre width histogram (CNFWH) and the corneal nerve-fibre orientation histogram (CNFOH). These can be calculated if the width and orientation at each nerve-fibre location is known. The orientation is calculated by the Hessian method referred to in section II-B-1. The nerve-fibre width estimation for a particular segment is illustrated in Fig. 3. Fig. 3a shows a highlighted example nerve-fibre segment along with a

0018-9294 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2016.2573642, IEEE Transactions on Biomedical Engineering

magnified version. At each nerve-fibre location, an intensity profile line of length 13 pixels (larger than the thickest fibre) is extracted perpendicular to the nerve-fibre orientation, as indicated by the short straight lines in Fig. 3a. The profiles corresponding to a fibre segment are averaged along the length of the segment to generate a representative profile for the segment, which is then further averaged (Fig. 3b) with its symmetrically inverted profile, smoothed by a three pixel length average filter and normalised (Fig. 3c). Finally a Gaussian curve is fitted to the normalised profile curve (Fig. 3c). The final width of that segment is calculated as 2.5 (empirically determined) times the RMS width of the fitted Gaussian curve. CNFWH is the number of occurrences of different fibre widths in the range between 1 to 8 pixels, at 0.2 pixels interval (36 bins). The CNFA is calculated as sum of fibre width × fibre length of all the fibre segments in mm^2. The CNFOH is the number of occurrences of different fibre orientations in the range between 0° to 179°, at 5 degree interval (36 bins).

III. Materials and Evaluation

We performed the model training and testing processes on two independent datasets. Dataset 1 contains 200 CCM images which were randomly selected from healthy volunteers and subjects who were diagnosed with type 1 diabetes. This dataset was used for model training and validation for parameter optimisations. The testing stage was conducted on an independent dataset 2 that contained 888 images captured from 176 subjects (84 controls and 92 diabetic patients). The subjects were divided into 3 groups: control (n=84), type 1 diabetic patient with no neuropathy (n=63) and diabetic patients with neuropathy (n=29). The Toronto Diabetic Neuropathy Expert Group (TC) [6] recommendation was followed to define an individual to have DSPN if he/she met both of the following criteria: (1) Abnormal nerve conduction – A peroneal motor nerve conduction velocity of <42 m/s; (2) a symptom or sign of neuropathy, defined as ONE of the following: (a) diabetic neuropathy symptom (DNS) [25] of 1 or more out of 4, (b) neuropathy disability score (NDS) [26] of 3 or more out of 10. These features, along with a number of other clinical and physiological parameters, were measured for each subject [22].

Following the description in section II-A, all images from both dataset 1 and dataset 2 were acquired by the same procedure. They were all manually segmented by a trained clinician (INP). CNFD, CNFL and CNBD were also measured manually in each of the images using the CCMetrics annotation tool (denoted as MCNFD, MCNFL, and MCNBD, respectively).

A. Evaluation of Nerve-Fibre Detection

For the evaluation of nerve-fibre detection, we firstly trained and validated the models based on dataset 1 using a two-fold cross validation. The 200 images in dataset 1 were randomly divided into two groups with 100 images each. Each set served for parameter setting and training based on the other half as a test set. The roles were then reversed. We performed this two-fold cross validation on the four combinations of feature descriptors (DMF and DWT) and model classifiers [24] (random forest (RFC) and multi-layer perceptron neural network classifiers (NNC)), denoted as DMRF (DMF + RFC), DMNN (DMF + NNC), DTRF (DWT + RFC) and DTNN (DWT + NNC). For each of the combinations, we repeated the two-fold cross validation to investigate the optimum parameter settings by varying the number of training pixels (500, 1000, 2000 pixels randomly selected from each of the foreground and background regions for each image), the number of trees (100, 200 and 500 trees) in RFC and the number of hidden neurons (20, 50 and 100) in the three-layer NNC.

For performance evaluation, as in [21], the response images (before denoising) were thresholded and thinned to one-pixel wide lines. These lines were then compared pixel by pixel to the manually generated skeletons acting as ground-truth, a true positive (TP) being scored if the detected pixel is within a three-pixel tolerance of ground truth and a false positive (FP) if it is outside this tolerance. True negative (TN) and false negative (FN) pixels are recorded if the pixel in the detected image is zero while the ground-truth is zero and one respectively. The three pixel tolerance deals with the imprecision in placing hand-drawn centrelines. By varying the threshold of the response images, ROC curves can be generated for each of the parameter settings. Optimum performance, in terms of specificity and sensitivity at the equal error rate point and computational time, was achieved by using 1000 foreground and background pixels from each image for training, and 200 trees for RFC and 50 hidden neurons for NNC.

In the testing stage, we applied the two optimised models (one from each of the two-fold cross validation runs on dataset 1) to the independent dataset 2. ROC performances were compared between models trained using DMNN, DMRF, DTNN and DTRF. The values of sensitivity and specificity at the equal error rate point for the two models were as follows. DMNN: 0.917 and 0.913, DMRF: 0.912 and 0.908, DTNN:

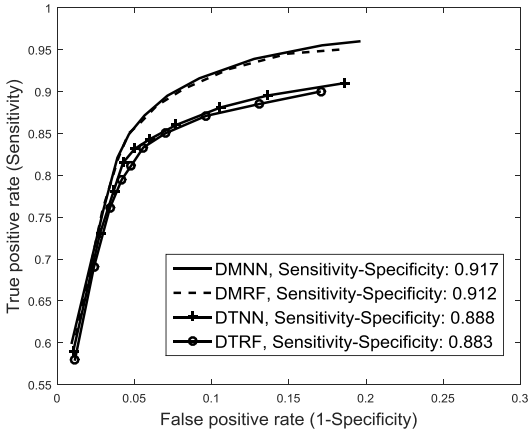


Fig. 4. ROC curves for nerve-fibre detection on dataset 2, using DMNN (Dual Model, Neural Network), DMRF (Dual Model, Random Forest), DTNN (Dual-Tree Wavelet, neural Network) and DTRF (Dual-Tree Wavelet, Random Forest) respectively.

0018-9294 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2016.2573642, IEEE Transactions on Biomedical Engineering

0.888 and 0.882, DTRF: 0.883 and 0.878. In Fig. 4, we show the ROC plots of the model that with the higher performance. These were obtained using the raw detections before denoising to obtain an insight into the underlying detector performance. From the ROC curves, it is clear that the combination based on the DMF outperforms the DTW feature descriptor, having higher sensitivity at any value of specificity.

Although the specificity is a fair measurement for the detection of both background and foreground pixels, the value is dominated by the very high TN count. The absolute value of specificity is potentially misleading, as we have noted in section II-B-1 that the initial detection results in detection of a high number of background pixels that are removed by a subsequent denoising step. We therefore also calculated the False Discovery Rate (FDR=FP/ (FP+TP)) and the False Negative Rate (FNR=FN/ (FN+TP)). The smallest (best) FDR/FNR measures for the four methods before and after image denoising are listed in Table I. Since the two cross validation models produce very similar classifications, we only report the results from one of the cross validation models. All four detector/classifier combinations have similar FDR values before denoising and, significantly reduced, after denoising, consistent with the similar specificity values at most values of sensitivity in Fig. 4. The FNR values increase only slightly by denoising, the DM detector achieving better FNR figures. Following denoising there is no real difference between the RF and NN classifiers.

TABLE I
FDR AND FNR FOR THE FOUR COMPARED DETECTORS BEFORE AND AFTER IMAGE DENOISING.

Method	FDR/FNR Before denoising	FDR/FNR After denoising
DMNN	0.4810/0.2700	0.2013/0.2934
DMRF	0.4828/0.2877	0.2014/0.2983
DTNN	0.4890/0.2961	0.2012/0.3141
DTRF	0.4881/0.3221	0.2013/0.3261

B. Evaluation of Nerve-Fibre Quantification

As observed in section III-A the two cross-validation models produced very similar performance on the independent dataset 2. We chose the detector model with the slightly higher performance as the basis for automated measurements of nerve-fibre parameters, denoted ACNFD, ACNFL and ANCBD. Additionally, total nerve-fibre area, orientation histogram and width histogram were calculated (CNFA, CNFOH ad CHWH). For the multidimensional features CNFOH and CNFWH, we investigated the use of the maximum, standard deviation, skewness, kurtosis and logistic regression combing all elements of the histogram feature vectors to represent the feature. The standard deviation of the histogram proved to be the most effective; these are denoted as ASDOH and ASDWH.

For each of the subjects, the average feature values obtained from their CCM images were used. Fig. 5 and Fig. 6 show the box plots of each of the manual and automated CCM features respectively. In these figures, the central red lines are the median, the edges of the box are the 25th and 75th percentiles (q1 and q3), and the whiskers extend to the most extreme data points that are not identified as being outliers (within the range q1-1.5(q3-q1) to q3+1.5(q3-q1)). The outliers are plotted individually as red dots. A common decreasing trend from control group to neuropathy group can be observed on all manual and automated CCM features. The values of the manually generated measurements are higher than those generated automatically. One reason for this is that the manual tracing process deviates from the exact fibre path (Fig 1(b)), resulting in a larger CNFL value. Additionally, the automated method is less effective than human annotators at connecting weak branches, resulting in generally higher CNFD and CNBD values for the manual analysis. However, the important point is the relative correlation between manual and automated measures across the control and patient groups. The Pearson correlation coefficients between automatically and manually derived CNFL, CNFD and CNBD measurements were 0.861, 0.859 and 0.701 respectively. The lower correlation in the case of CNBD measurement is due to poor reproducibility in the manual measurement of this feature. This has been reported in [15] and arises from the subjective judgement required for identifying branch points.

We used both the ANOVA test [27] and ROC analysis to demonstrate the capability of using the CCM image features to discriminate between control and non-neuropathic groups, and between non-neuropathic and neuropathic patients, as defined by the Toronto Criteria.

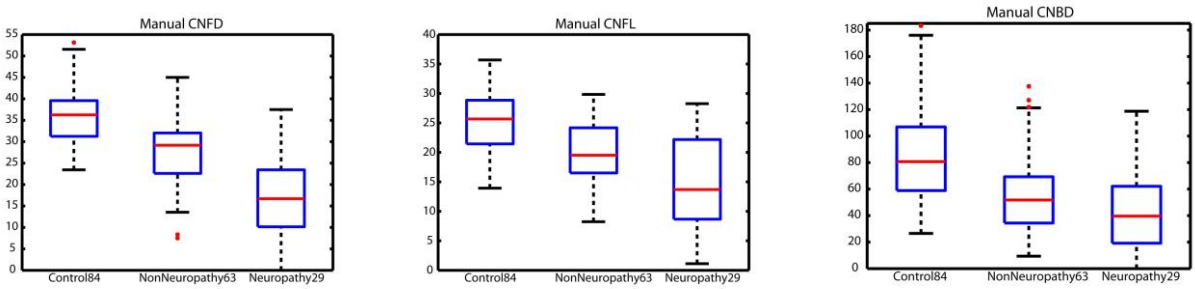
Tables II and III show the respective ANOVA p-values, the area under the ROC curve (AUC) measures and sensitivity and specificity values calculated at the equal error point (EEP) of the ROC curves. We also experimented with different combinations of features, from both manual and automated analysis, using logistic regression in a leave-one-out manner. In these experiments each subject was predicted by the logistic regression model built from the remaining n-1 subjects, where n is the total number of subjects in both groups. ROC measures for the combinations of all manual features or all automated features are listed in Table II and Table III along with the single-feature measures. The confidence intervals for the combined methods indicate that the combination results in a discriminating power indistinguishable from the best manual or automatic methods respectively. This would indicate that the features are accessing the same underlying information about each of the groups. It is unsurprising that there should be dependency between total fibre length and fibre density or fibre area.

IV. DISCUSSION

A number of studies have shown the features extracted from Corneal Confocal Microscopy images are associated with the severity of diabetic peripheral neuropathy [7, 12, 13] and the potential of CCM to quantify severity of neuropathy and assess therapeutic benefit has been demonstrated [28]. In this paper, we have described the details of a complete system for measurement of CCM images to enable discrimination between control and diabetic subjects and between diabetic subjects with and without neuropathy.

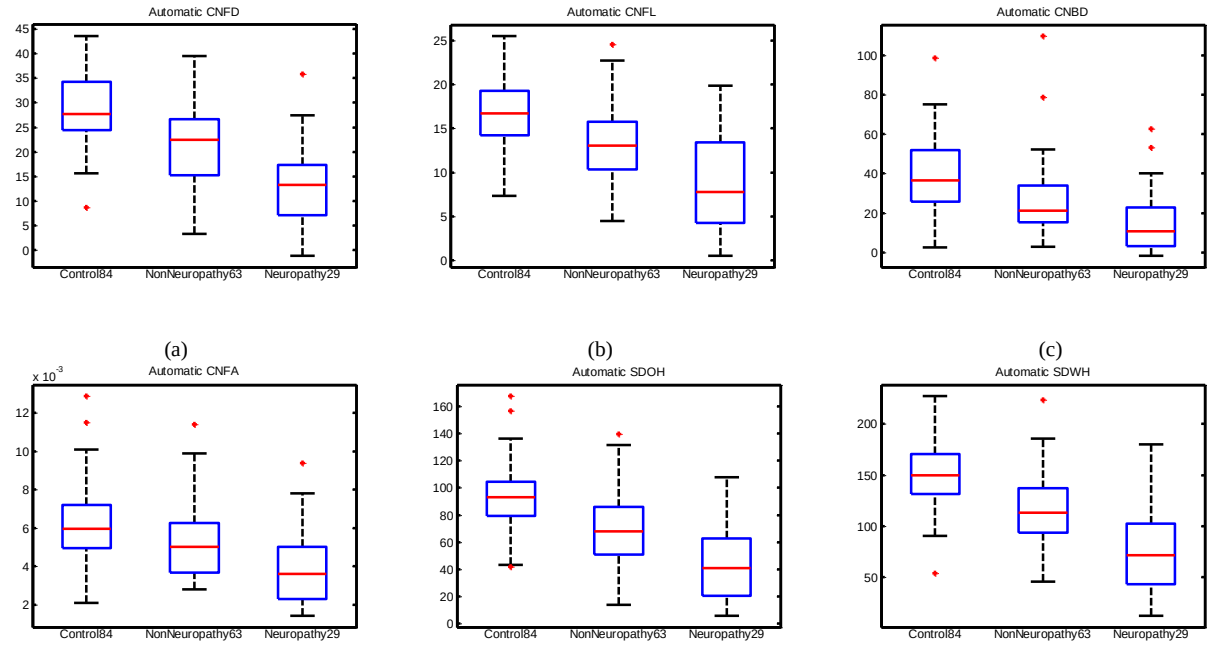
0018-9294 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2016.2573642, IEEE Transactions on Biomedical Engineering



(a) (b) (c)

Fig. 5. Boxplots of manually measured features for control, non-neuropathy and neuropathy groups (a) MCNFD (b) MCNFL (c) MCNBD.



(a) (b) (c)

(d) (e) (f)

Fig. 6. Boxplots of automatically measured features for control, non-neuropathy and neuropathy groups in dataset 2 (a) ACNFD (b) ACNFL (c) ACNBD (d) ACNFA (e) ASDOH (f) ASDWH.

Petropoulos et al. [22] reported a clinical evaluation study comparing the system described in this paper with manual analysis of CCM images and a broader range of subjective and objective clinical assessment methods, including the Neuropathy Symptom Profile, vibration perception thresholds, cool and warm thermal thresholds, and cold and heat induced pain. CCM features, measured both automatically and manually, were found to be significantly correlated with these methods. They noted that the automatic analysis of CCM images was significantly faster than manual analysis, taking 10-22s per image, depending on the density of fibres, as opposed to 2-7 minutes.

Based on the well-established Toronto Criteria, we show that both manual and automated CCM features discriminate diabetic patients with and without neuropathy. Manual and automatic measurement regimes result in broadly similar results: about 0.77 AUC value and 73% sensitivity-specificity at the equal error rate point. There were no significant differences between the ROCs of manual (MCNFD) and automated measurements (e.g. p=0.44 and 0.55 for ACNFD and SDWH respectively).

Corneal confocal microscopy has shown considerable success in translation to the assessment of other neuropathies including Fabry disease [29], ISFN [30], CMT1A [31], sarcoidosis [32]. Automated quantification of corneal nerves provides a major step forward in the early diagnosis and assessment of progression, but in particular for use in clinical trials to establish therapeutic benefit in diabetic and other peripheral neuropathies.

The automatic quantification software can be requested freely from [33] for research purposes. It is currently being used by over 40 research groups worldwide to investigate potential relationships between CCM features and different types of neuropathy [34].

V. Conclusion

We have presented a technical evaluation of a complete system that is able to automatically quantify six different types of nerve-fibre features in CCM images. We have proposed an optimum configuration for detection of nerve fibres based on a

0018-9294 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2016.2573642, IEEE Transactions on Biomedical Engineering

TABLE II
AUC, 95% CONFIDENCE INTERVAL VALUES AND SENSITIVITY-SPECIFICITY AT THE EQUAL-ERROR POINT (EEP) FOR MANUAL AND AUTOMATED CCM FEATURES FOR DISCRIMINATION BETWEEN **CONTROL SUBJECTS** AND **DIABETIC PATIENTS** WITHOUT DSPN.

CCM FEATURES	AUC	95% CI	SENSITIVITY SPECIFICITY AT EEP	P-VALUE OF ANOVA
MCNFD	**0.8063**	**[0.73 0.88]**	**0.7460**	**<0.0001**
MCNFL	0.7627	[0.68 0.84]	0.6825	<0.0001
MCNBD	0.7492	[0.67 0.83]	0.6984	<0.0001
ACNFD	0.7401	[0.66 0.82]	0.7305	<0.0001
ACNFL	**0.7766**	**[0.70 0.85]**	**0.7613**	**<0.0001**
ACNBD	0.7103	[0.63 0.80]	0.6414	<0.0001
CNFA	0.6837	[0.59 0.77]	0.6601	0.0002
SDOH	0.7671	[0.69 0.85]	0.7002	<0.0001
SDWH	0.7798	[0.71 0.86]	0.7402	<0.0001
COMBINED MANUAL	0.7940	[0.72 0.87]	0.7143	-
COMBINED AUTOMATED	0.7373	[0.67 0.83]	0.7009	-

TABLE III
AUC, 95% CONFIDENCE INTERVAL VALUES AND SENSITIVITY-SPECIFICITY AT THE EQUAL-ERROR POINT (EEP) FOR MANUAL AND AUTOMATED CCM FEATURES FOR DISCRIMINATION BETWEEN **NON-NEUROPATHIC** AND **NEUROPATHIC** GROUPS OF DIABETIC PATIENTS.

CCM FEATURES	AUC	95% CI	SENSITIVITY SPECIFICITY AT EEP	P-VALUE OF ANOVA
MCNFD	**0.7890**	**[0.68 0.90]**	**0.7241**	**<0.0001**
MCNFL	0.7137	[0.59 0.83]	0.6552	0.001
MCNBD	0.6136	[0.49 0.74]	0.5862	0.081
ACNFD	0.7600	[0.65 0.87]	0.6482	<0.0001
ACNFL	0.7576	[0.65 0.88]	0.6186	<0.0001
ACNBD	0.6801	[0.56 0.80]	0.5798	0.002
CNFA	0.7601	[0.64 0.87]	0.7301	<0.0001
SDOH	0.7799	[0.68 0.90]	0.6907	<0.0001
SDWH	**0.7709**	**[0.67 0.88]**	**0.7219**	**<0.0001**
COMBINED MANUAL	0.7843	[0.68 0.89]	0.7100	-
COMBINED AUTOMATED	0.7419	[0.63 0.86]	0.6779	-

previously reported foreground and background model trained with a neural network. The automatic quantification results show a high correlation with manually measured CCM features (CNFL, CNFD and CNBD). The results also show significant differences (p-values of ANOVA test in table II) between the control and non-neuropathic group, indicating the system's ability to detect early signs of change from a healthy to a diabetic condition. The automated system is able to produce additional CCM features that measure the area, width and orientation of the nerve fibres (CNFA, CNFWH and CNFOH). All these new measures show significant differences between the non-neuropathic and neuropathic groups (p-values of ANOVA test in table III), with some features achieving 72% sensitivity-specificity at the equal error rate point, indicating the capacity to identify individuals suffering from neuropathy. The advantages in time labour and reproducibility suggest that automatically measured features may be used as a new, non-invasive method for diagnosing diabetic peripheral neuropathy, providing information on small nerve-fibre damage that is not accessible by most currently used methods. The only method in current clinical use that addresses small fibre damage is the intra-epidermal nerve-fibre density (IENFD) measure, which is invasive, requiring a skin biopsy, and currently cannot be evaluated automatically. We have recently shown [35] that analysis of CCM features has favourable diagnostic efficacy to IENFD (AUC of 0.66)

ACKNOWLEDGMENT

This research was funded by awards from: National Institutes of Health (R105991) and Juvenile Diabetes Research Foundation International (27-2008-362).

REFERENCES

[1] A. J. Boulton, "Management of Diabetic Peripheral Neuropathy," *Clinical Diabetes*, vol. 23, no. 1, pp. 9-15, 2005.

[2] C. Daousi, I. A. MacFarlane, A. Woodward, T. J. Nurmikko, P. E. Bundred, and S. J. Benbow, "Chronic painful peripheral neuropathy in an urban community: a controlled comparison of people with and without diabetes," *Diabetic Medicine*, vol. 21, no. 9, pp. 976-982, 2004.

[3] P. J. Dyck, C. J. Overland, P. A. Low, W. J. Litchy, J. L. Davies, P. C. O'Brien, C. v. N. T. Investigators, J. W. Albers, H. Andersen, C. F. Bolton, J. D. England, C. J. Klein, J. G. Llewelyn, M. L. Mauermann, J. W. Russell, W. Singer, A. G. Smith, S. Tesfaye, and A. Vella, "Signs and symptoms versus nerve conduction studies to diagnose diabetic sensorimotor polyneuropathy: CI vs. NPhys trial," *Muscle Nerve*, vol. 42, no. 2, pp. 157-164, 2010.

[4] P. J. Dyck, J. W. Albers, J. Wolfe, C. F. Bolton, N. Walsh, C. J. Klein, A. J. Zafft, J. W. Russell, K. Thomas, J. L. Davies, R. E. Carter, L. J. Melton, W. J. Litchy, and C. v. N. T. Investigators, "A trial of proficiency of nerve conduction: greater standardization still needed," *Muscle nerve*, vol. 48, no. 3, pp. 369-374, 2013.

[5] P. J. Dyck, J. E. Norell, H. Tritschler, K. Schuette, R. Samigullin, D. Ziegler, E. J. Bastyr, W. J. Litchy, and P. C. O'Brien, "Challenges in Design of Multicenter Trials: Endpoints Assessed Longitudinally for Change and Monotonicity," *Diabetes Care*, vol. 30, pp. 2619-2625, 2007.

[6] S. Tesfaye, A. J. Boulton, P. J. Dyck, R. Freeman, M. Horowitz, P. Kempler, G. Lauria, R. A. Malik, V. Spallone, A. Vinik, L.

0018-9294 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2016.2573642, IEEE Transactions on Biomedical Engineering

Bernardi, and P. Valensi, "on behalf of the Toronto Diabetic Neuropathy Expert Group, Diabetic neuropathies: Update on definitions, diagnostic criteria, estimation of severity, and treatments," *Diabetes Care*, vol. 33, no. 10, pp. 2285-2293, 2010.

[7] P. Hossain, A. Sachdev, and R. A. Malik, "Early detection of diabetic peripheral neuropathy with corneal confocal microscopy," *The Lancet*, vol. 366, no. 94, pp. 1340-1343, 2005.

[8] N. Pritchard, K. Edwards, A. W. Russell, B. A. Perkins, R. A. Malik, and N. Efron, "Corneal confocal microscopy predicts 4-year incident peripheral neuropathy in type 1 diabetes," *Diabetes Care*, vol. 38, no. 4, pp. 671-675, 2015.

[9] M. Brines, A. N. Dunne, M. V. Velzen, P. L. Proto, C. G. Ostenson, R. I. Kirk, I. Petropoulos, S. Javed, R. A. Malik, A. Cerami, and A. Dahan, "ARA 290, a non-erythropoietic peptide engineered from erythropoiethin, improves metabolic control and neuropathic symptoms in patients with type 2 diabetes," *Molecular Medicine*, vol. 6, 2014.

[10] M. Tavakoli, M. Mitu-Pretorian, I. N. Petropoulos, H. Fadavi, O. Asghar, U. Alam, G. Ponirakis, M. Jeziorska, A. Marshall, N. Efron, A. J. Boulton, T. Augustine, and R. A. Malik, "Corneal confocal microscopy detects early nerve regeneration in diabetic neuropathy after simultaneous pancreas and kidney transplantation," *Diabetes Care*, vol. 62, no. 1, pp. 254-260, 2013.

[11] S. Azmi, M. Ferdousi, I. N. Petropoulos, G. Ponirakis, H. Fadavi, M. Tavakoli, U. Alam, and W. Jones, "Corneal confocal microscopy shows an improvement in small-fiber neuropathy in subjects with type 1 diabetes on continuous subcutaneous insulin infusion compared with multiple daily injection," *Diabetes Care*, vol. 38, no. 1, pp. e3-e4, 2015.

[12] M. Tavakoli, C. Quattrini, C. Abbott, P. Kallinikos, A. Marshall, J. Finnigan, P. Morgan, N. Efron, A. Boulton, and R. Malik, "Corneal Confocal Microscopy: A Novel Non-invasive Test to Diagnose and Stratify the Severity of Human Diabetic Neuropathy," *Diabetes Care*, vol. 33, no. 8, pp. 1792-1797, 2010.

[13] R. A. Malik, P. Kallinikos, C. A. Abbott, C. H. M. v. Schie, P. Morgan, N. Efron, and A. J. M. Boulton, "Corneal confocal microscopy: a noninvasive surrogate of nerve fibre damage and repair in diabetic patients," *Diabetologia*, vol. 46, no. 5, pp. 683-688, 2003.

[14] C. Dehghani, N. Pritchard, K. Edwards, D. Vagenas, A. W. Russell, R. A. Malik, and N. Efron, "Morphometric stability of the corneal subbasal nerve plexus in healthy individuals: 1 3-year longitudinal study using corneal confocal microscopy," *Invest Ophthalmol Visual Science*, vol. 55, no. 5, pp. 3195-3199, 2014.

[15] I. Petropoulos, T. Manzoor, P. Morgan, H. Fadavi, O. Asghar, U. Alam, G. Ponirakis, M. Dabbah, X. Chen, J. Graham, M. Tavakoli, and R. Malik, "Repeatability of In Vivo Corneal Confocal Microscopy to Quantify Corneal Nerve Morphology," *Cornea*, vol. 32, no. 5, pp. 83-89, 2013.

[16] M. Niemeijer, J. J. Staal, B. v. Ginneken, M. Loog, and M. D. Abramoff, "Comparative study of retinal vessel segmentation methods on a new publicly available database," *SPIE Medical Imaging*, vol. 5370, pp. 648-656, 2004.

[17] M. Berks, Z. Chen, S. Astley, and C. Taylor, "Detecting and classifying linear structures in mammograms using random forests," *IPMI 11 proceedings of the 22nd international conference on information processing in medical imaging*, pp. 510-524, 2011.

[18] F. Scarpa, E. Grisan, and A. Ruggeri, "Automatic Recognition of Corneal Nerve Structures in Images from Confocal Microscopy," *Investigative Ophthalmology and Visual Science*, vol. 49, no. 11, pp. 4801-4807, 2008.

[19] T. Holmes, M. Pellegrini, C. Miller, T. Epplin-Zapf, S. Larkin, S. Luccarelli, and G. Staurenghi, "Automated Software Analysis of Corneal Micrographs for Peripheral Neuropathy," *Investigative Ophthalmology and Visual Science*, vol. 51, no. 9, pp. 4480-4491, 2010.

[20] C. W. Sindt, B. Lay, H. Bouchard, and J. R. Kern, "Rapid Image Evaluation System for Corneal In Vivo Confocal Microscopy," *Cornea*, vol. 32, no. 4, pp. 460-465, 2013.

[21] M. A. Dabbah, J. Graham, I. N. Petropoulos, M. Tavakoli, and R. A. Malik, "Automatic analysis of diabetic peripheral neuropathy using multi-scale quantitative morphology of nerve fibres in corneal confocal microscopy imaging," *Medical Image Analysis*, vol. 15, pp. 738-747, 2011.

[22] I. N. Petropoulos, U. Alam, H. Fadavi, A. Marshal, O. Asghar, M. A. Dabbah, X. Chen, J. Graham, G. Ponikaris, A. J. M. Boulton, M. Tavakoli, and R. A. Malik, "Rapid automated diagnosis of diabetic peripheral neuropathy with in vivo corneal confocal microscopy," *Investigative Optics and Visual Science*, vol. 55, pp. 2071-2078, 2014.

[23] N. G. Kingsbury, "Complex wavelets for shift invariant analysis and filtering of signals," *Journal of Applied and Computational Harmonic Analysis*, vol. 10, no. 3, pp. 234-253, 2001.

[24] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, "Machine learning: a review of classification and combining techniques," *Artificial Intelligence Review*, vol. 26, no. 3, pp. 159-190, 2006.

[25] J. W. G. Meijer, A. J. Smit, E. V. Sonderen, J. W. Groothoff, W. H. Eisma, and T. P. Links, "Symptom scoring systems to diagnose distal polyneuropathy in diabetes: the Diabetic Neuropathy Symptom score," *Diabetic Medicine*, vol. 19, no. 11, pp. 962-965, 2002.

[26] M. J. Young, A. J. M. Boulton, A. F. Macleod, D. R. R. Williams, and P. H. Sonksen, "A multicentre study of the prevalence of diabetic peripheral neuropathy in the United Kingdom hospital clinic population," *Diabetologia* vol. 36, pp. 150-154, 1993.

[27] K. Wallis, "Use of ranks in on-criterion variance analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583-621, 1952.

[28] T. Holmes, M. Pellegrini, C. Miller, T. Epplin-Zapf, S. Larkin, S. Luccarelli, and G. Staurengbi, "Automated Software Analysis of Corneal Micrographs for Peripheral Neuropathy," *Investigative Ophthalmology and Visual Science*, vol. 51, no. 9, pp. 4480-4491, 2010.

[29] M. Tavakoli, A. Marshall, L. Thompson, M. Kenny, S. Waldek, N. Efron, and R. A. Malik, "Corneal confocal microscopy: a novel noninvasive means to diagnose neuropathy in patients with Fabry disease," *Muscle Nerve*, vol. 40, no. 6, pp. 976-84, Dec, 2009.

[30] M. Tavakoli, A. Marshall, R. Pitceathly, H. Fadavi, D. Gow, M. E. Roberts, N. Efron, A. J. Boulton, and R. A. Malik, "Corneal confocal microscopy: a novel means to detect nerve fibre damage in idiopathic small fibre neuropathy," *Exp Neurol*, vol. 223, no. 1, pp. 245-50, May, 2010.

[31] M. Tavakoli, A. Marshall, S. Banka, I. N. Petropoulos, H. Fadavi, H. Kingston, and R. A. Malik, "Corneal confocal microscopy detects small-fiber neuropathy in Charcot-Marie-Tooth disease type 1A patients," *Muscle Nerve*, vol. 46, no. 5, pp. 698-704, Nov, 2012.

[32] M. van Velzen, L. Heij, M. Niesters, A. Cerami, A. Dunne, A. Dahan, and M. Brines, "ARA 290 for treatment of small fiber neuropathy in sarcoidosis," *Expert Opin Investig Drugs*, vol. 23, no. 4, pp. 541-50, Apr, 2014.

[33] Accmetrics. "http://www.medicine.manchester.ac.uk/ena/."

[34] ENAgroup. "http://www.human-development.manchester.ac.uk/ena/ACCMetricsuserinstructions/."

[35] X. Chen, J. Graham, M. Dabbah, I. Petropoulos, G. Ponirakis, O. Asghar, U. Alam, A. Marshall, H. Fadavi, M. Ferdousi, S. Azmi, M. Tavakoli, N. Efron, M. Jeziorska, and R. Malik, "Small nerve fiber quantification in the diagnosis of diabetic sensorimotor polyneuropathy: comparing corneal confocal microscopy with intraepidermal nerve fiber density," *Diabetes Care*, vol. 38, no. 6, pp. 1138-1144, 2015.

0018-9294 (c) 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

55. **Rapid automated diagnosis of diabetic peripheral neuropathy with in vivo corneal confocal microscopy**. I.N. Petropoulos, U. Alam, H. Fadavi, A. Marshall, O. Asghar, M.A. Dabbah, X. Chen, J. Graham, G. Ponikaris, A.J.M. Boulton, M. Tavakol1, R,A, Malik, *Investigational Ophthalmology and Visual Science, **55**, 2071-2078, 2014. doi: 10.1167/iovs.13-13787*

Rapid Automated Diagnosis of Diabetic Peripheral Neuropathy With In Vivo Corneal Confocal Microscopy

Ioannis N. Petropoulos,[1] Uazman Alam,[2] Hassan Fadavi,[1] Andrew Marshall,[2] Omar Asghar,[1] Mohammad A. Dabbah,[2] Xin Chen,[2] James Graham,[2] Georgios Ponirakis,[1] Andrew J. M. Boulton,[1] Mitra Tavakoli,[1] and Rayaz A. Malik[1]

[1]School of Medicine, Institute of Human Development, Centre for Endocrinology and Diabetes, Manchester Academic Health Science Centre, Manchester, United Kingdom
[2]Department of Clinical Neurophysiology, Central Manchester NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, United Kingdom
[3]Institute of Population Health, Imaging Science, University of Manchester, Manchester Academic Health Science Centre, Manchester, United Kingdom

Correspondence: Rayaz A. Malik, University of Manchester School of Medicine, Institute of Human Development, Centre for Endocrinology and Diabetes, 46 Grafton Street, Core Technology Facility, Manchester M13 9NT; rayaz.a.malik@manchester.ac.uk.

Submitted: December 17, 2013
Accepted: February 7, 2014

Citation: Petropoulos IN, Alam U, Fadavi H, et al. Rapid automated diagnosis of diabetic peripheral neuropathy with in vivo corneal confocal microscopy. *Invest Ophthalmol Vis Sci*. 2014;55:2071–2078. DOI:10.1167/iovs.13-13787

Purpose. To assess the diagnostic validity of a fully automated image analysis algorithm of in vivo confocal microscopy images in quantifying corneal subbasal nerves to diagnose diabetic neuropathy.

Methods. One hundred eighty-six patients with type 1 and type 2 diabetes mellitus (T1/T2DM) and 55 age-matched controls underwent assessment of neuropathy and bilateral in vivo corneal confocal microscopy (IVCCM). Corneal nerve fiber density (CNFD), branch density (CNBD), and length (CNFL) were quantified with expert, manual, and fully-automated analysis. The areas under the curve (AUC), odds ratios (OR), and optimal thresholds to rule out neuropathy were estimated for both analysis methods.

Results. Neuropathy was detected in 53% of patients with diabetes. A significant reduction in manual and automated CNBD ($P < 0.001$) and CNFD ($P < 0.0001$), and CNFL ($P < 0.0001$) occurred with increasing neuropathic severity. Manual and automated analysis methods were highly correlated for CNFD ($r = 0.9$, $P < 0.0001$), CNFL ($r = 0.89$, $P < 0.0001$), and CNBD ($r = 0.75$, $P < 0.0001$). Manual CNFD and automated CNFL were associated with the highest AUC, sensitivity/specificity and OR to rule out neuropathy.

Conclusions. Diabetic peripheral neuropathy is associated with significant corneal nerve loss detected with IVCCM. Fully automated corneal nerve quantification provides an objective and reproducible means to detect human diabetic neuropathy.

Keywords: corneal confocal microscopy, diabetic neuropathy, diabetes

Diabetic sensorimotor polyneuropathy (DSPN) is a frequent complication of diabetes affecting up to 53% of people with diabetes.[1] Diagnosis of the condition is important to define at-risk patients, anticipate deterioration, and assess new therapies. Neuropathic symptoms and signs, together with electrodiagnostic studies are the endpoints of choice to diagnose DSPN and assess therapeutic outcomes.[2] Although these tests offer a robust means of assessing neuropathy, they predominantly focus on large fiber deficits, yet the earliest alterations occur in the small unmyelinated C- and thinly myelinated Aδ-nerve fibers.[3] Small fiber neuropathy can be evaluated using quantitative sensory testing of thermal thresholds or skin biopsy to quantify intra-epidermal nerve fiber density (IENFD). However, the assessment of thermal thresholds is subjective and therefore liable to variability,[4] while skin biopsy is an invasive and costly technique, which is not routinely available across healthcare systems.[5]

We have pioneered the use of IVCCM and shown that this rapid, noninvasive ophthalmic technique can accurately quantify changes in the human subbasal nerve plexus of patients with diabetes.[6] Alterations in the subbasal corneal nerves occur early, increase with neuropathic severity,[7] and are paralleled by significant IENF loss.[8] Recent studies have shown that chronic glycemic exposure,[9] even in subjects without overt diabetes,[10] hypertension,[9] and elevated serum triglycerides,[11] are strong risk factors for corneal subbasal nerve loss. Furthermore, early reinnervation of the cornea has been shown in recipients of simultaneous pancreas and kidney transplantation (SPK).[12,13] It is important to note that other ocular diseases, such as dry eyes,[14] atopic keratoconjunctivitis,[15] epithelial membrane basement dystrophies,[16] cystic corneal disorders,[17] and other conditions[18] may also affect corneal innervation, and should therefore be excluded in any study using IVCCM in DSPN.

Concerns regarding the use of IVCCM have focused on the reproducibility[19,20] of the technique, its ability to prospectively assess neuropathy, and the absence of an automated image analysis system to allow objective corneal nerve quantification. The latter is essential to eliminate inconsistencies, produce comparable outcomes across centers, and enable the deployment of IVCCM for diagnosis, and as a surrogate endpoint in clinical trials of diabetic neuropathy. Previous studies[21–23] have proposed a variety of quantification algorithms, which differ by methodology and detection properties. In our recent work,[23]

Copyright 2014 The Association for Research in Vision and Ophthalmology, Inc.

Table 1. Medical and Peripheral Neuropathy Status

Variable	Controls, $n = 55$, NDS = 0	DSPN (−), $n = 86$, NDS ≤ 2	DSPN (+), $n = 100$, NDS > 2
Duration of diabetes	N/A	24.2 ± 21.2	34.4 ± 17.3
HbA_{1c}, %/mmol/mol‡	5.5 ± 0.3/34 ± 3.3	7.7 ± 1.6/61 ± 17.5§	7.9 ± 1.6/63 ± 17.5§
BMI, Kg/m*	25.6 ± 4.6	27.2 ± 5.2	27.6 ± 5.8‖
TC, mM‡	5.1 ± 0.9	4.3 ± 1.2§	4.4 ± 0.9§
Triglycerides, mM	1.5 ± 0.8	1.5 ± 0.9	1.4 ± 0.9
eGFR, mL/min/L‡	85.8 ± 7.8	81.8 ± 18.2	70.0 ± 24.5§‖
ACR, mg/mmol‡	1.0 ± 1.4	2.9 ± 1.3	18.8 ± 11.3§‖
BP, systolic†/diastolic, mm Hg	122 ± 16/70 ± 8.8	130 ± 18§/71 ± 9	138 ± 23§‖/72 ± 8
NSP	0	1.9 ± 3.0	5.6 ± 6.2
VPT, V‡	5.8 ± 4.6	9.2 ± 6.5§	22.3 ± 12.6§‖
WT†/CT†, °C	37.0 ± 3.0/28.2 ± 2.2	39.6 ± 3.9§/27.0 ± 9.2§	42.7 ± 4.6‖/20.8 ± 9.2§‖
HIP/CIP‡, °C	44.8 ± 2.9/11.9 ± 9.2	45.5 ± 6.6/9.8 ± 10.7	46.9 ± 7.3/4.1 ± 6.2§‖
PMNCV, m/s‡	48.8 ± 3.3	43.7 ± 4.7§	39.2 ± 6.1§‖
SSNCV, m/s‡	51.0 ± 4.8	46.4 ± 5.8§	42.2 ± 6.4§‖
PMNamp, μV‡	5.2 ± 1.8	4.5 ± 3.2	2.4 ± 2.1§‖
SSNamp, μV‡	20.0 ± 9.7	12.5 ± 7.8§	6.5 ± 6.6§‖

Results are expressed as mean ± SD, statistically significant differences using ANOVA/Kruskal-Wallis. N/A, not applicable for this group.
* $P < 0.05$.
† $P < 0.001$.
‡ $P < 0.0001$; post hoc results for DSPN (+) significantly different from § control subjects and ‖ DSPN (−).

we described an algorithm that concurrently uses a dual-model feature descriptor and a neural network classifier to distinguish nerve fibers from the background and presented an evaluation of its performance against other available detection methods. The aim of the present study was to assess the diagnostic validity of a fully automated image analysis algorithm of in vivo confocal microscopy images in quantifying corneal subbasal nerves to diagnose diabetic neuropathy.

Methods

Study Subjects

One hundred eighty-six patients with diabetes mellitus (108 male/78 female) and 55 age-matched control subjects (28 male/27 female) (50.4 ± 14.1 vs. 51.7 ± 11.4 years) were assessed for the presence and severity of DSPN between 2010 and 2011 based on the updated Toronto consensus criteria.[2] Informed written consent was obtained from all participants prior to their enrolment to the study. This research adhered to the tenets of the Declaration of Helsinki and was approved by the North Manchester Research Ethics Committee. Subjects were excluded if they had a positive history of malignancy, connective tissue or infectious disease, deficiency of vitamin B_{12} or folate, chronic renal failure, liver failure, active diabetic foot ulceration, and/or family history of peripheral neuropathy. Control subjects were excluded from the study if they had evidence of neuropathy or risk factors likely to cause neuropathy. All subjects were also assessed for the presence of corneal lesions by means of relevant history and slit-lamp biomicroscopy. Subjects were excluded if they had active ocular disease (e.g., severe dryness), systemic disease known to affect the corneal subbasal innervation, other than diabetes or chronic corneal pathologies (cystic corneal disorders, epithelial basement membrane dystrophies).

Medical Status Assessment

All participants underwent assessment of their cardiometabolic [glycated hemoglobin (HbA1c), total cholesterol (TC), triglycerides and body mass index (BMI)] and renal status [estimated glomerular filtration rate (eGFR) and albumin to creatinine ratio (ACR)].

Peripheral Neuropathy Assessment

The neuropathy disability score (NDS), a scale of 0 to 10, was used to stratify the neuropathic severity of the study participants into none (0–2), mild (3–5), moderate (6–8), and severe (9–10) as described elsewhere[21] (Tables 1, 2). The neuropathy symptom profile (NSP) was employed to assess symptoms of neuropathy. Vibration perception threshold (VPT) was evaluated on the hallux of both feet with a Neurothesiometer (Horwell Scientific Laboratory Suppliers, Wilford, UK). Cool and warm thermal (CT/WT) thresholds and cold- and heat-induced pain (CIP/HIP) were established on the dorsolateral aspect of the left foot (S1) with a TSA-II

Table 2. IVCCM Assessment of DSPN Status

Variable	Controls, NDS = 0	DSPN (−), NDS ≤ 2	DSPN (+), NDS > 2
Manual IVCCM quantification			
$CNFD_M$, no./mm^2§	37.2 ± 6.7	26.7 ± 8.5‖	20.5 ± 9.5‖¶
$CNBD_M$, no./mm^2‡	92.7 ± 38.6	54.9 ± 35.7‖	48.7 ± 33.2‖
$CNFL_M$, mm/mm^2‡	26.4 ± 5.6	20.3 ± 6.7‖	16.7 ± 7.6‖¶
Automated IVCCM quantification			
$CNFD_A$, no./mm^2§	30.0 ± 6.9	20.1 ± 8.7‖	14.4 ± 8.9‖¶
$CNBD_A$, no./mm^2‡	50.4 ± 24.7	31.4 ± 25.6‖	20.1 ± 18.7‖¶
$CNFL_A$, mm/mm^2§	21.2 ± 3.5	17.1 ± 4.5‖	13.7 ± 5.2 ‖¶
Corneal sensation			
NCCA, mbar†	0.7 ± 0.5	0.9 ± 0.8‖	1.5 ± 2.1‖

Results are expressed as mean ± SD, statistically significant differences using ANOVA/Kruskal-Wallis. no., number; mbar, millibar.
* $P < 0.05$.
† $P < 0.01$.
‡ $P < 0.00$.
§ $P < 0.0001$; post hoc results for diabetes DSPN (+) significantly different from ‖ control subjects and ¶ DSPN (−).

Investigative Ophthalmology & Visual Science

NeuroSensory Analyser (Medoc Ltd., Ramat-Yishai, Israel) using the method of limits.

Nerve conduction studies (NCS) were undertaken by a consultant neurophysiologist (AM) as previously described.[24] Peroneal motor nerve amplitude (PMNamp) and conduction velocity (PMNCV) and sural sensory nerve amplitude (SSNmap) and conduction velocity (SSNCV) were assessed. The diabetes cohort included 11 patients that did not agree or were unable to undergo NCS. These patients were not excluded from the study, but were not considered when NCS results were assessed.

Study Definition of Peripheral Neuropathy

The Toronto Diabetic Neuropathy Expert Group[2] recommendation was followed to define "Confirmed DSPN: the presence of an abnormality of NCS and a symptom or symptoms or a sign or signs of neuropathy. In the absence of an abnormal NCS, a validated measure of small fiber neuropathy should be used" and "Subclinical DSPN: the presence of no signs or symptoms of neuropathy confirmed with an abnormal NCS or a validated measure of small fiber neuropathy." To define an abnormal result for NCS and QST we have used a mean ± 2 SD cutoff based on our control population.

In Vivo Corneal Confocal Microscopy

All study subjects were scanned with a laser IVCCM (Heidelberg Retinal Tomograph III Rostock Cornea Module [HRT III RCM]; Heidelberg Engineering GmbH, Heidelberg, Germany) as described elsewhere.[20] The overall examination took approximately 5 minutes for both eyes of each subject, and in this study two experienced optometrists performed all IVCCM scans. All images were captured using the "section" mode and prior to scanning corneal sensation was assessed using noncontact corneal aesthesiometry (NCCA) as described elsewhere.[25]

Manual Image Analysis

During a bilateral IVCCM scan more than 100 images per patient were typically captured from all corneal layers. Six subbasal images from right and left eyes were selected for analysis. Criteria for image selection were depth, focus position, and contrast. A single experienced examiner (INP), masked from the outcome of the medical and peripheral neuropathy assessment, quantified 1506 images of all study participants using purpose-written, proprietary software (CCMetrics, MA Dabbah; Imaging Science and Biomedical Engineering, University of Manchester, Manchester, UK). The specific parameters measured per frame were: CNFD (no./mm^2), CNFL (mm/mm^2), and CNBD (no./mm^2) in accord with our previously published protocol.[20]

Automated Image Analysis

Automated corneal nerve fiber quantification consists of two steps: (1) IVCCM image enhancement and nerve fiber detection, and (2) quantification of the three morphometric parameters. As described in our earlier work,[22,23] a dual-model feature descriptor combined with a neural network classifier was used to train the computer to distinguish nerve fibers from the background (noise and underlying connective tissue). In the nerve fiber quantification process, all the end points and branch points of the detected nerve fibers are extracted and used to construct a connectivity map. Each segment in the connectivity map was then connected and classified as main nerve fibers or branches.

Statistical Analysis

Statistical analysis was performed using StatsDirect for Windows (version 2.7.9; StatsDirect Ltd., Cheshire, UK) and STATA 12 for Windows (Stata Corporation, College Station, TX, USA) was used to generate the receiver operating characteristic curves (ROC). Correlation analysis was performed to assess the strength of the relationship between automated and manually generated variables. Linear regression analysis was used to assess the consistency of the responses from the fully automated algorithm for a given manual estimate. The intraclass correlation coefficient (ICC) was calculated as a measure of reliability of the automated image analysis algorithm over repeated assessment of the dataset. One-way ANOVA (nonparametric Kruskal-Wallis) were used to evaluate within and between group differences. P value was maintained at 0.05 for multiple comparisons (Bonferroni adjustment or Conover-Inman pairwise comparisons) and a P less than 0.05 was considered significant.

Receiver operating characteristic curves analysis was performed for all corneal nerve parameters to identify the point closest to the upper left corner of the ROC graph, which concurrently optimized sensitivity and specificity and the AUC, OR, and positive (+LR) and negative likelihood ratios (–LR) associated with the point were calculated. The diagnostic validity of IVCCM was assessed in relation to four established measures of DSPN (PMNamp, SSNamp, PMNCV, and WT). A χ^2 test was used to compare the AUCs generated for all IVCCM parameters.

Results

Medical Status and DSPN Assessment

Detailed medical and DSPN assessment results for subjects with diabetes and controls are presented in Table 1. Diabetic sensorimotor polyneuropathy(+) compared with DSPN(–) and controls had a lower eGFR ($P < 0.0001$), higher ACR ($P < 0.0001$), systolic blood pressure (BP) ($P = 0.0003$), VPT ($P < 0.0001$), WT ($P = 0.0005$), and lower CT ($P = 0.0004$), CIP ($P < 0.0001$), PMNCV ($P < 0.0001$), SSNCV ($P < 0.0001$), PMNamp ($P < 0.0001$), and SSNamp ($P < 0.0001$). Diabetic sensorimotor polyneuropathy(+) subjects had a longer duration of diabetes (34.4 ± 17.3 vs. 24.2 ± 21.2, $P = 0.01$) and were older compared with DSPN(–) (55.3 ± 12.4 vs. 47.3 ± 15.6, $P = 0.001$). Metabolic control and BMI were significantly different between controls (HbA_{1c}, $P < 0.0001$; BMI, $P < 0.05$) and patients with diabetes, but comparable between DSPN(+) and DSPN(–). Total cholesterol (TC) was similar between the two groups with diabetes, and lower compared with controls ($P < 0.0001$), which is likely due to statin used in the diabetes cohort.

Manual and Automated Assessment of DSPN With IVCCM

Diabetic sensorimotor polyneuropathy(+) compared with DSPN(–) and controls had significantly lower manually quantified $CNFD_M$ ($P < 0.0001$), $CNBD_M$ ($P = 0.0005$), $CNFL_M$ ($P = 0.0002$), and automatically quantified $CNFD_A$ ($P < 0.0001$), $CNBD_A$ ($P = 0.0002$), and $CNFL_A$ ($P < 0.0001$) parameters. A significant reduction was also detectable between DSPN(–) and controls in $CNFD_M$ ($P < 0.0001$), $CNBD_M$ ($P = 0.0006$), $CNFL_M$ ($P = 0.0003$), and $CNFD_A$ ($P < 0.0001$), $CNBD_A$ ($P = 0.0003$), and $CNFL_A$ ($P < 0.0001$). Changes detected using automated image quantification were associated with a stronger significance level. Noncontact corneal aesthesiometry showed a significant elevation in the

Investigative Ophthalmology & Visual Science

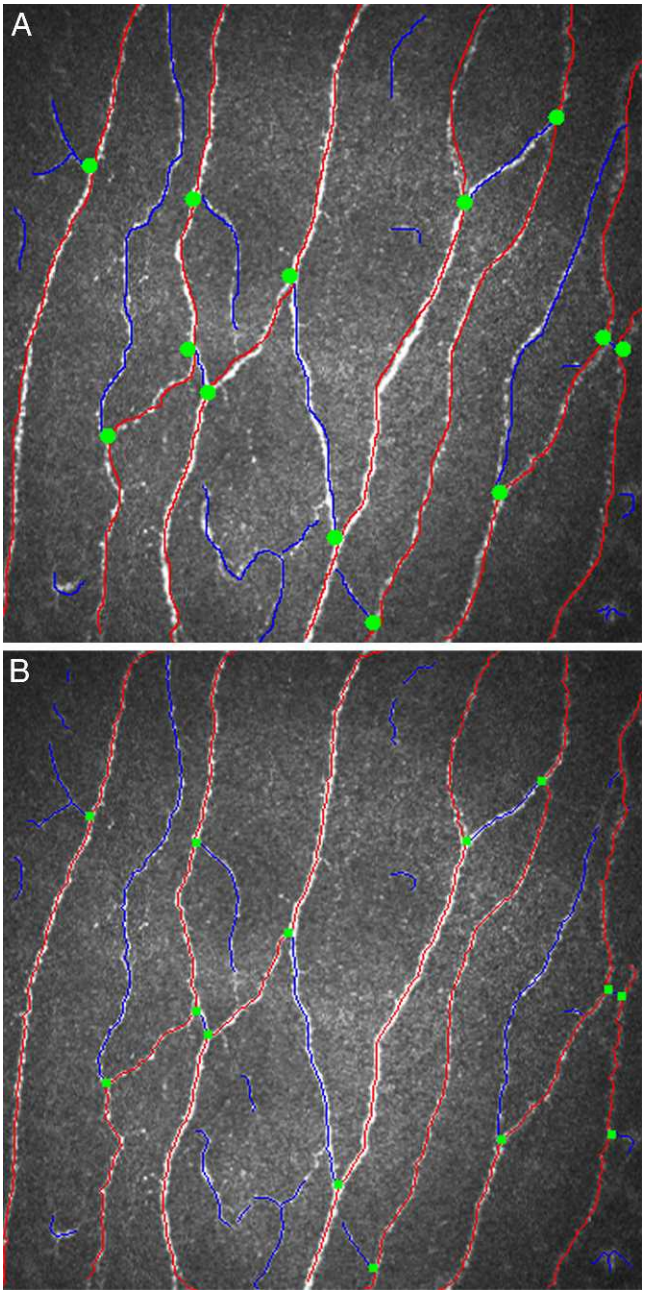


FIGURE 1. An IVCCM image of a control subject analyzed using (**A**) manual expert and (**B**) fully-automated image analysis to quantify corneal subbasal nerve morphology in DSPN. Use of either quantification method results in the detection of comparable structures in the image.

corneal sensation threshold in diabetic subjects and control subjects ($P = 0.004$). All results are presented in Table 2.

Manual Versus Automated Image Analysis

Manual and automated results were strongly correlated for CNFD (adjusted $R^2 = 0.81$, $r = 0.90$, $P < 0.0001$), CNBD (adjusted $R^2 = 0.58$, $r = 0.75$, $P < 0.0001$), and CNFL (adjusted $R^2 = 0.79$, $r = 0.89$. $P < 0.0001$) (Figs. 1A–C). Upon revaluation of the same dataset the reproducibility of the automated algorithm was excellent (ICC = 1.0) across all IVCCM parameters. Automated quantification significantly reduced image analysis time. Each image required 10 to 22 seconds to be processed automatically, while manual analysis took 2 to 7 minutes per image depending on the density of the nerves. Examples of analyzed images using the two methods are presented in Figure 1.

Validity of IVCCM Image Quantification for Diagnosis of DSPN. Receiver operating characteristic curves were inspected for concurrent optimization of sensitivity and specificity and the associated AUCs were calculated for manual and automated IVCCM parameters with respect to the study definition of "neuropathy" (Table 3).

PMNamp Less Than 1.4 μv. There were 53 (30%) diabetic patients who had neuropathy based on abnormal PMNamp. A $CNFD_M$ less than 18.7 no./mm^2 was the point where sensitivity (0.79) and specificity (0.78) were concurrently optimized and associated with the highest AUC = 0.84, OR = 16.5, +LR = 4.6 (95% confidence interval [CI] 3.0–6.9), and –LR = 0.3 (95% CI 0.2–0.4). The corresponding point for automated analysis was $CNFD_A$ less than 14.7 no./mm^2 with sensitivity (0.76) and specificity (0.72) and AUC = 0.80, OR = 11.0, +LR = 3.4 (95% CI 2.4–4.9), and –LR = 0.3 (95% CI 0.2–0.5) (Fig. 2A). Similarly, $CNFL_M$ and $CNFL_A$ were associated with an AUC of 0.82 and 0.84 respectively, +LR = 3.23 (95% CI 2.3–4.6) and –LR = 0.33 (95% CI 0.2–0.5) (Fig. 2).

SSNamp Less Than 5.5 μv. When an abnormal SSNamp result was used as an indicator of neuropathy, the number of abnormal cases increased to 72 (40%). Automatically quantified $CNFL_A$ was associated with the highest AUC (0.77) and the highest OR = 5.1. A $CNFL_A$ less than 16.1 mm/mm^2 optimized sensitivity (0.72) and specificity (0.66) with +LR = 2.1 (95% CI 1.6–2.9) and –LR = 0.4 (95% CI 0.3–0.6). A $CNFL_M$ less than 19.1 mm/mm^2 optimized sensitivity (0.68) and specificity (0.67), but was associated with a lower AUC (0.70) and OR = 4.6 and comparable +LR = 2.1 (95% CI 1.5–3.0) and –LR = 0.5 (95% CI 0.3–0.7). Both $CNFD_M$ and $CNFD_A$ were equally capable in ruling out neuropathy. Both $CNBD_A$ and $CNBD_M$ showed limited ability to differentiate between cases with and without neuropathy.

PMNCV Less Than 42 M/S. There were 96 (54%) diabetic patients who had an abnormal PMNCV result. Automatically quantified $CNFL_A$ was associated with the highest AUC (0.79) and a $CNFL_A$ less than 16.0 mm/mm^2 optimized sensitivity (0.74) and specificity (0.71) with OR = 7.2, +LR = 2.6 (95% CI 1.9–3.8), and –LR = 0.3 (95% CI 0.2–0.5). A $CNFL_M$ less than 19.7 mm/mm^2 was associated with 0.74 sensitivity and 0.63 sensitivity, AUC = 0.73, OR = 4.8, +LR = 2.0 (95% CI 1.6–2.6), and –LR = 0.4 (95% CI 0.3–0.6). Both $CNFD_A$ and $CNFD_M$ had comparable AUC, OR, LR, and sensitivity/specificity to rule out neuropathy.

WT Greater Than 42°C. There were 95 (51%) patients with diabetes who had an abnormal WT greater than 42°C. Both $CNFD_M$ and $CNFD_A$ were associated with the highest AUC and modest OR. Specifically, a $CNFD_M$ less than 24.0/mm^2 optimized sensitivity (0.63) and specificity (0.62) and was associated with AUC 0.69, OR 2.9, +LR 1.6 (95% CI 1.2–2.1) and –LR 0.7 (95% CI 0.5–0.8). The number of patients with an abnormal $CNFD_M$ and a WT was 61 (64%), while 35 (37%) had reduced $CNFD_M$ with a normal WT result. All $CNFD_A$, $CNFL_M$, and $CNFL_A$ values were comparable, but were associated with slightly lower AUC and OR while sensitivity and specificity remained modest (Table 3).

DISCUSSION

Diabetic peripheral neuropathy is the main initiating factor for foot ulceration and amputation and is associated with heavy morbidity, reduced quality of life, and poor healthcare

Table 3. Validity and Associated Probabilities of DSPN Detection Using Manual and Automated IVCCM Parameters Quantification

Definition of DSPN	IVCCM Value (Sensitivity/Specificity)	AUC	Odds Ratio (95% CI)	+LR (95% CI)	−LR (95% CI)
PMNamp, <1.4 μV					
$CNFD_M$	18.7 (0.79/0.78)	0.84	16.5 (7.0–39.9)	4.6 (3.0–7.0)	0.3 (0.2–0.4)
$CNFD_A$	14.7 (0.76/0.72)	0.80	11.0 (4.8–24.8)	3.4 (2.4–4.9)	0.3 (0.2–0.5)
$CNBD_M$	41.7 (0.73/0.68)	0.75	5.9 (2.7–13.1)	2.3 (1.7–3.1)	0.4 (0.2–0.6)
$CNBD_A$	14.9 (0.74/0.73)	0.79	9.2 (4.1–21.4)	2.9 (2.1–4.7)	0.3 (0.2–0.5)
$CNFL_M$	15.8 (0.77/0.76)	0.82	9.8 (4.4–22.0)	3.2 (2.3–4.6)	0.3 (0.2–0.5)
$CNFL_A$	14.6 (0.77/0.74)	0.84	12.9 (5.5–31.8)	3.3 (2.4–4.6)	0.2 (0.1–0.4)
SSNamp, <5.5 μV					
$CNFD_M$	23.1 (0.72/0.67)	0.74	4.7 (2.3–10.0)	1.9 (1.5–2.6)	0.4 (0.3–0.6)
$CNFD_A$	18.9 (0.73/0.56)	0.72	5.1 (2.4–11.1)	1.9 (1.5–2.5)	0.4 (0.2–0.6)
$CNBD_M$	47.1 (0.61/0.56)	0.65	2.1 (1.1–4.9)	1.4 (1.0–1.9)	0.7 (0.5–1.0)
$CNBD_A$	23.4 (0.63/0.54)	0.70	2.1 (1.1–4.2)	1.4 (1.0–1.9)	0.7 (0.5–0.9)
$CNFL_M$	19.4 (0.68/0.67)	0.70	4.6 (2.3–9.3)	2.1 (1.5–3.0)	0.5 (0.3–0.7)
$CNFL_A$	16.1 (0.72/0.66)	0.77	5.1 (2.5–10.4)	2.1 (1.6–2.9)	0.4 (0.3–0.6)
PMNCV, <42.0 m/s					
$CNFD_M$	25.4 (0.78/0.70)	0.74	8.2 (4.1–17.3)	2.6 (1.9–3.7)	0.3 (0.2–0.5)
$CNFD_A$	19.7 (0.80/0.61)	0.74	7.8 (3.7–16.7)	2.2 (1.7–3.0)	0.3 (0.2–0.4)
$CNBD_M$	49.0 (0.69/0.61)	0.68	3.7 (1.9–7.2)	1.8 (1.3–2.5)	0.5 (0.4–0.7)
$CNBD_A$	24.9 (0.68/0.52)	0.67	2.4 (1.2–4.6)	1.4 (1.1–1.9)	0.6 (0.4–0.9)
$CNFL_M$	19.7 (0.74/0.63)	0.73	4.9 (2.4–9.7)	2.0 (1.5–2.8)	0.4 (0.3–0.6)
$CNFL_A$	16.0 (0.74/0.71)	0.79	7.2 (3.5–14.7)	2.6 (1.8–3.8)	0.4 (0.3–0.5)
WT, >41°C					
$CNFD_M$	24.0 (0.63/0.62)	0.69	2.9 (1.5–5.3)	1.7 (1.3–2.3)	0.6 (0.4–0.8)
$CNFD_A$	17.3 (0.63/0.60)	0.67	2.5 (1.4–4.6)	1.5 (1.2–2.1)	0.6 (0.5–0.8)
$CNBD_M$	47.2 (0.65/0.55)	0.65	2.1 (1.2–3.8)	1.4 (1.1–1.9)	0.7 (0.5–0.9)
$CNBD_A$	22.9 (0.60/0.58)	0.64	2.1 (1.1–3.9)	1.4 (1.1–2.0)	0.7 (0.5–0.9)
$CNFL_M$	19.2 (0.63/0.61)	0.67	2.7 (1.5–5.0)	1.6 (1.2–2.2)	0.6 (0.4–0.8)
$CNFL_A$	15.9 (0.61/0.61)	0.68	2.3 (1.3–4.2)	1.5 (1.1–2.1)	0.7 (0.5–0.9)

outcomes.[26] The prevalence of DSPN, in the diabetic population varies from 10% to 53%.[1,27–29] However, only a few studies have used objective endpoints to estimate the rates of neuropathy and this may explain the reported variability. Dyck and colleagues[30] found that when NCS was used in combination with a functional abnormality to diagnose DSPN as opposed to conventional clinical examination, twice as many patients were detected. Electrodiagnostic studies are the gold standard to diagnose neuropathy, but they are limited to large fibers and previous research has shown that small nerve fibers are affected first.[3] An objective, noninvasive surrogate of small fiber damage, such as IVCCM,[7] is therefore desirable to diagnose neuropathy early and define patients at risk.

Previous studies have identified age, duration of diabetes, renal status, BP, cardiometabolic control, and anthropometric parameters as risk factors for the onset and severity of DSPN.[29,31–33] Recent studies using IVCCM, have reported an association between levels of HbA_{1c}, BP, and triglycerides with the density of corneal innervation.[9–11] This study assessed 188 subjects with diabetes, but no other identifiable cause of neuropathy, and found that a significant decline in eGFR, increased ACR, and systolic BP were associated with neuropathy. Both diabetes groups [DSPN (+), DSPN (−)] had modest to poor metabolic control.

Corneal confocal microscopy provides the unique opportunity to repeatedly and reliably visualize the corneal nerves adjacent to Bowman's membrane. An increasing body of literature supports the use of IVCCM in the diagnosis and severity stratification of DSPN.[6,7,9,34] At present, a major drawback is the absence of an automated analysis system, which would eliminate inconsistencies and make the technique suitable to a clinical setting. This study assessed, for the first time, the performance and validity of a novel fully-automated image analysis algorithm compared with manual human expert analysis in relation to multiple gold standard clinical endpoints used to define neuropathy.

We found that both methods of image quantification were highly correlated primarily for CNFD and CNFL but also CNBD. We detected a slight underestimation of corneal nerve density and length when automated analysis was used, which was however consistent. The detection of nerve structures in IVCCM images is a challenging task: Nerve fibers often show poor contrast on a relatively noisy background due to microscope properties and underlying structures. As described in our earlier work,[23] the algorithm operates through a combination of detection methods and predefined criteria, mainly nerve-specific characteristics such as orientation and axon reflectivity, to construct a connectivity map and distinguish a nerve structure from noise. In contrast, manual image analysis is a labor-intensive task, where a human investigator applies subjective criteria to define a nerve and an overestimation with less experience has been described.[20] In this study, we found a significant and progressive reduction in nerve density, branching, length between diabetic patients with and even without DSPN, and controls using either quantification method.

Corneal nerve branch density showed a significant positive correlation between manual and automated assessment, but this was not as high as for CNFD and CNFL. Corneal nerve branch density, a measurement of nerve branches directly connected to nerve fibers, has been reported to be highly variable and appears to have modest validity in diagnosing

Investigative Ophthalmology & Visual Science

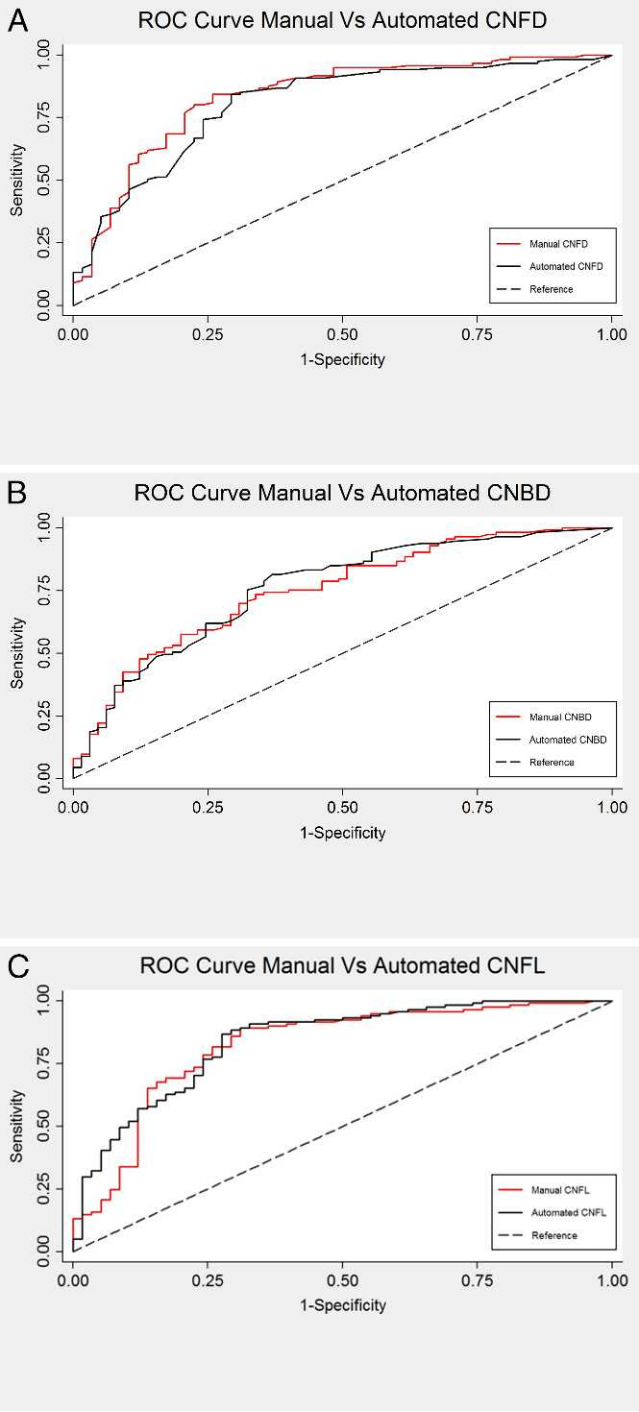


Figure 2. Receiver operating characteristic curves for manual (*solid black*) and automated (*red*) CNFD (**A**), CNBD (**B**), and CNFL (**C**). Corneal nerve fiber density and CNFL showed the highest validity to diagnose DSPN with comparable AUCs (no significant difference). Manual CNFD and automated CNFL were associated with the highest OR.

neuropathy in this and other studies.[13,34] Moreover, inter- and intraobserver estimation of the parameter in highly innervated corneas has shown moderate reproducibility.[20] The relevance of corneal nerve branching to DSPN is not clear. In our recent study,[35] of the 1-year effects of SPK transplantation in type 1 DM recipients, we found a significant and stable increase before an improvement in any other measure of regeneration. In this study, automated analysis of CNBD was more capable in staging neuropathy than manually quantified CNBD, likely due to less variability compared with manual human analysis.

Recently, two studies have assessed the validity of IVCCM in diagnosing DSPN. Tavakoli et al.[7] has reported a CNFD less than or equal to 27.8 no./mm^2 and less than or equal to 20.8 no./mm^2 as the values with the highest validity to define disease status among patients with mild and more severe neuropathy respectively. Ahmed et al.[34] found that a CNFL less than or equal to 14.0 mm/mm^2 was the value with the highest validity to rule in DSPN. We assessed the performance of manual and automated IVCCM quantification to identify patients "with" or "without" neuropathy based on gold standard measures of peripheral nerve damage. We found that $CNFD_M$, $CNFD_A$, $CNFL_M$, and $CNFL_A$ were associated with the highest sensitivity and specificity to diagnose DSPN when PMNamp was used as the primary measure of neuropathy. Corneal nerve branch density showed less but acceptable validity in diagnosing DSPN and $CNBD_A$ had a significantly higher AUC and OR compared with $CNBD_M$. When other endpoints of DSPN were used, such as SSNamp and PMNCV, the diagnostic validity of IVCCM remained high and $CNFL_A$ was consistently associated with the highest AUC and OR among all parameters. We observed a significant decline in sensitivity and specificity when an abnormality in WT was used as the primary marker of neuropathy. One would expect the opposite since warm detection is mainly mediated by small nerve fibers, and previously we have shown an association between IENFD and corneal nerve morphology.[8] More recently CNFL has been related to three different measures of small fiber neuropathy.[36] This is likely for two main reasons: NCS offer a robust and objective means of assessing neuropathy, while WT is a subjective measurement of small fiber function. Cassanova et al.[37] in their study found that even patients with no IENFs had consistent responses in WT and concluded that it is possible for partially damaged nerve endings to still generate a propagated action potential. We speculate that a similar association may exist for the corneal subbasal nerves.

The validity of fully automated corneal nerve quantification was comparable and in several cases exceeded the performance of human expert assessment in ruling out DSPN. A $CNFL_A$ between 14.6 mm/mm^2 and 16.1 mm/mm^2 was the value consistently associated with the highest AUC and OR given the case definition employed. Both $CNFD_M$ (18.7–25.4 no./mm^2) and $CNFD_A$ (14.7–19.7 no./mm^2) also showed excellent performance with high OR, but were slightly more variable.

This study has several strengths and limitations. The strengths of this study are the detailed clinical assessment by gold standard clinical techniques of a relatively large number of participants with diabetes, representing a wide range of disease duration and neuropathic severity. Moreover, the same highly trained individuals performed all examinations for the 241 participants of this study ensuring consistency of the results. Our findings and cutoff points selected for the diagnosis of DSPN by IVCCM are comparable with the previous studies of Ahmed et al.[34] and Tavakoli et al.[7]; slight differences could be due to the case definition of neuropathy employed in each study, the number of patients investigated, and the disease severity in each group. We have compared IVCCM with several objective and subjective markers of DSPN with significant findings for the validity of the technique. There are no directly comparable published results for the fully automated algorithm employed in this study, therefore we cannot exclude the possibility that another system may be superior to the one presented here. This is to date the only available purpose-built, automated corneal nerve quantification system that has been validated in a large cohort of patients

with diabetes and varying degrees of DSPN. Our results are cross-sectional and ongoing longitudinal studies[38] will determine the ability of IVCCM to predict the development and progression or regression of DSPN. Recent data generated from wide-field assessment of the subbasal plexus have suggested that both central and inferior whorl nerve density may be reduced early and therefore future studies should explore this further.[39]

In conclusion, we show that diabetic peripheral neuropathy is paralleled by a significant and progressive reduction in central CNFD and CNFL. We have validated a rapid fully automated analysis system to quantify alterations to replace human manual quantification. The use of this system will clearly enhance reproducibility, eliminate inconsistencies, and make the technique suitable to clinical practice and research centers worldwide.

Acknowledgments

The authors thank the Manchester Biomedical Research Centre and the Greater Manchester Comprehensive Local Research Network, who facilitated this research.

Supported by grants from the National Institutes of Health (R105991) and the Juvenile Diabetes Research Foundation International (27-2008-362).

Disclosure: **I.N. Petropoulos**, None; **U. Alam**, None; **H. Fadavi**, None; **A. Marshall**, None; **O. Asghar**, None; **M.A. Dabbah**, None; **X. Chen**, None; **J. Graham**, None; **G. Ponirakis**, None; **A.J.M. Boulton**, None; **M. Tavakoli**, None; **R.A. Malik**, None

References

1. Dyck PJ, Kratz KM, Karnes JL, et al. The prevalence by staged severity of various types of diabetic neuropathy, retinopathy, and nephropathy in a population based cohort. *Neurology*. 1993;43:817–824.
2. Tesfaye S, Boulton AJM, Dyck PJ, et al. Diabetic neuropathies: update on definitions, diagnostic criteria, estimation of severity, and treatments. *Diabetes Care*. 2010;33:2285–2293.
3. Dyck PJ, Giannini C. Pathologic alterations in the diabetic neuropathies of humans: a review. *J Neuropathol Exp Neurol*. 1996;55:1181–1193.
4. Freeman R, Chase KP, Risk MR. Quantitative sensory testing cannot differentiate simulated sensory loss from sensory neuropathy. *Neurology*. 2003;60:465–470.
5. Lauria G, Lombardi R, Camozzi F, Devigili G. Skin biopsy for the diagnosis of peripheral neuropathy. *Histopathology*. 2009; 54:273–285.
6. Malik RA, Kallinikos P, Abbott CA, et al. Corneal confocal microscopy: a non-invasive surrogate of nerve fibre damage and repair in diabetic patients. *Diabetologia*. 2003;46:683–688.
7. Tavakoli M, Quattrini C, Abbott C, et al. Corneal confocal microscopy: a novel non invasive test to diagnose and stratify the severity of human diabetic neuropathy. *Diabetes Care*. 2010;33:1792–1797.
8. Quattrini C, Tavakoli M, Jeziorska M, et al. Surrogate markers of small fiber damage in human diabetic neuropathy. *Diabetes*. 2007;56:2148–2154.
9. Ishibashi F, Okino M, Ishibashi M, et al. Corneal nerve fiber pathology in Japanese type 1 diabetic patients and its correlation with antecedent glycemic control and blood pressure. *J Diabetes Investig*. 2012;3:191–198.
10. Wu T, Ahmed A, Bril V, et al. Variables associated with corneal confocal microscopy parameters in healthy volunteers: implications for diabetic neuropathy screening. *Diabet Med*. 2012;29:e297–e303.
11. Tavakoli M, Marshall A, Pitceathly R, et al. Corneal confocal microscopy: a novel means to detect nerve fibre damage in idiopathic small fibre neuropathy. *Exp Neurol*. 2010;223:245–250.
12. Mehra S, Tavakoli M, Kallinikos PA, et al. Corneal confocal microscopy detects early nerve regeneration after pancreas transplantation in patients with type 1 diabetes. *Diabetes Care*. 2007;30:2608–2612.
13. Hertz P, Bril V, Orszag A, et al. Reproducibility of in vivo corneal confocal microscopy as a novel screening test for early diabetic sensorimotor polyneuropathy. *Diabet Med*. 2011;28: 1253–1260.
14. Benítez-del-Castillo JM, Acosta MC, Wassfi MA, et al. Relation between corneal innervation with confocal microscopy and corneal sensitivity with noncontact esthesiometry in patients with dry eye. *Invest Ophthalmol Vis Sci*. 2007;48:173–181.
15. Hu Y, Matsumoto Y, Adan ES, et al. Corneal in vivo confocal scanning laser microscopy in patients with atopic keratoconjunctivitis. *Ophthalmology*. 2008;115:2004–2012.
16. Rosenberg ME, Tervo TMT, Petroll WM, Vesaluoma MH. In vivo confocal microscopy of patients with corneal recurrent erosion syndrome or epithelial basement membrane dystrophy. *Ophthalmology*. 2000;107:565–573.
17. Chiou AG-Y, Kaufman SC, Beuerman RW, Ohta T, Soliman H, Kaufman HE. Confocal microscopy in cornea guttata and Fuchs' endothelial dystrophy. *Br J Ophthalmol*. 1999;83:185–189.
18. Kaufman SC, Musch DC, Belin MW, et al. Confocal microscopy: a report by the American Academy of Ophthalmology. *Ophthalmology*. 2004;111:396–406.
19. Efron N, Edwards K, Roper N, et al. Repeatability of measuring corneal subbasal nerve fiber length in individuals with type 2 diabetes. *Eye Contact Lens*. 2010;36:245–248.
20. Petropoulos IN, Manzoor T, Morgan P, et al. Repeatability of in vivo corneal confocal microscopy to quantify corneal nerve morphology. *Cornea*. 2013;32:e83–e89.
21. Scarpa F, Grisan E, Ruggeri A. Automatic recognition of corneal nerve structures in images from confocal microscopy. *Invest Ophthalmol Vis Sci*. 2008;49:4801–4807.
22. Dabbah M, Graham J, Petropoulos I, Tavakoli M, Malik R. Dual-model automatic detection of nerve-fibres in corneal confocal microscopy images. In: *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2010*. Berlin: Springer Berlin Heidelberg; 2010. Vol. 6361:300–307.
23. Dabbah MA, Graham J, Petropoulos IN, Tavakoli M, Malik RA. Automatic analysis of diabetic peripheral neuropathy using multi-scale quantitative morphology of nerve fibres in corneal confocal microscopy imaging. *Med Image Anal*. 2011;15:738–747.
24. Petropoulos IN, Alam U, Fadavi H, et al. Corneal nerve loss detected with corneal confocal microscopy is symmetrical and related to the severity of diabetic polyneuropathy. *Diabetes Care*. 2013;36:3646–3651.
25. Tavakoli M, Kallinikos PA, Efron N, Boulton AJ, Malik RA. Corneal sensitivity is reduced and relates to the severity of neuropathy in patients with diabetes. *Diabetes Care*. 2007;30: 1895–1897.
26. Boulton AJM, Vileikyte L, Ragnarson-Tennvall G, Apelqvist J. The global burden of diabetic foot disease. *Lancet*. 2005;366: 1719–1724.
27. Tesfaye S, Stevens LK, Stephenson JM, et al. Prevalence of diabetic peripheral neuropathy and its relation to glycaemic control and potential risk factors: the EURODIAB IDDM Complications Study. *Diabetologia*. 1996;39:1377–1384.
28. Young M, Boulton A, Macleod A, Williams D, Sonksen P. A multicentre study of the prevalence of diabetic peripheral neuropathy in the United Kingdom hospital clinic population. *Diabetologia*. 1993;36:150–154.

29. Adler AI, Boyko EJ, Ahroni JH, Stensel V, Forsberg RC, Smith DG. Risk factors for diabetic peripheral sensory neuropathy: results of the Seattle Prospective Diabetic Foot Study. *Diabetes Care*. 1997;20:1162–1167.
30. Dyck PJ, Davies JL, Litchy WJ, O'Brien PC. Longitudinal assessment of diabetic polyneuropathy using a composite score in the Rochester Diabetic Neuropathy Study cohort. *Neurology*. 1997;49:229–239.
31. Turner R, Holman R, Cull C, et al. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). *Lancet*. 1998;352:837–853.
32. Tesfaye S, Chaturvedi N, Eaton SE, et al. Vascular risk factors and diabetic neuropathy. *N Engl J Med*. 2005;352:341–350.
33. Reichard P, Nilsson B-Y, Rosenqvist U. The effect of long-term intensified insulin treatment on the development of microvascular complications of diabetes mellitus. *N Engl J Med*. 1993; 329:304–309.
34. Ahmed A, Bril V, Orszag A, et al. Detection of diabetic sensorimotor polyneuropathy by corneal confocal microscopy in type 1 diabetes: a concurrent validity study. *Diabetes Care*. 2012;35:821–828.
35. Tavakoli M, Mitu-Pretorian M, Petropoulos IN, et al. Corneal confocal microscopy detects early nerve regeneration in diabetic neuropathy after simultaneous pancreas and kidney transplantation. *Diabetes*. 2013;62:254–260.
36. Sivaskandarajah GA, Halpern EM, Lovblom LE, et al. Structure-function relationship between corneal nerves and conventional small-fiber tests in type 1 diabetes. *Diabetes Care*. 2013; 36:2748–2755.
37. Casanova-Molla J, Grau-Junyent JM, Morales M, Valls-Sole J. On the relationship between nociceptive evoked potentials and intraepidermal nerve fiber density in painful sensory polyneuropathies. *PAIN*. 2011;152:410–418.
38. Edwards K, Pritchard N, Vagenas D, Russell A, Malik RA, Efron N. Utility of corneal confocal microscopy for assessing mild diabetic neuropathy: baseline findings of the LANDMark study. *Clin Exp Opt*. 2012;95:348–354.
39. Edwards K, Pritchard N, Gosschalk K, et al. Wide-field assessment of the human corneal subbasal nerve plexus in diabetic neuropathy using a novel mapping technique. *Cornea*. 2012;31:1078–1082.

Investigative Ophthalmology & Visual Science

56. **Small nerve fiber quantification in the diagnosis of sensorimotor polyneuropathy: comparing corneal confocal microscopy with intraepidermal nerve fiber density.** X. Chen, J Graham, M.A. Dabbah, I.N. Petropoulos, G. Ponikaris, O. Ashgar, U. Alam. A. Marshall, H. Favadi, M. Ferdousi, S. Azmi, M. Tavokoli, N. Efron, M. Jeziorka and R.A. Malik, *Diabetes Care, **38(6),** 1138-1144, 2015. doi: 10.2337/dc14-2422*

Small Nerve Fiber Quantification in the Diagnosis of Diabetic Sensorimotor Polyneuropathy: Comparing Corneal Confocal Microscopy With Intraepidermal Nerve Fiber Density

Xin Chen,[1] Jim Graham,[1] Mohammad A. Dabbah,[1] Ioannis N. Petropoulos,[2,3] Georgios Ponirakis,[2,3] Omar Asghar,[2,3] Uazman Alam,[2,3] Andrew Marshall,[4] Hassan Fadavi,[2,3] Maryam Ferdousi,[2,3] Shazli Azmi,[2,3] Mitra Tavakoli,[2,3] Nathan Efron,[5] Maria Jeziorska,[2,3] and Rayaz A. Malik[2,3]

Diabetes Care 2015;38:1138–1144 | DOI: 10.2337/dc14-2422

OBJECTIVE

Quantitative assessment of small fiber damage is key to the early diagnosis and assessment of progression or regression of diabetic sensorimotor polyneuropathy (DSPN). Intraepidermal nerve fiber density (IENFD) is the current gold standard, but corneal confocal microscopy (CCM), an in vivo ophthalmic imaging modality, has the potential to be a noninvasive and objective image biomarker for identifying small fiber damage. The purpose of this study was to determine the diagnostic performance of CCM and IENFD by using the current guidelines as the reference standard.

RESEARCH DESIGN AND METHODS

Eighty-nine subjects (26 control subjects and 63 patients with type 1 diabetes), with and without DSPN, underwent a detailed assessment of neuropathy, including CCM and skin biopsy.

RESULTS

Manual and automated corneal nerve fiber density (CNFD) ($P < 0.0001$), branch density (CNBD) ($P < 0.0001$) and length (CNFL) ($P < 0.0001$), and IENFD ($P < 0.001$) were significantly reduced in patients with diabetes with DSPN compared with control subjects. The area under the receiver operating characteristic curve for identifying DSPN was 0.82 for manual CNFD, 0.80 for automated CNFD, and 0.66 for IENFD, which did not differ significantly ($P = 0.14$).

CONCLUSIONS

This study shows comparable diagnostic efficiency between CCM and IENFD, providing further support for the clinical utility of CCM as a surrogate end point for DSPN.

Diabetic sensorimotor polyneuropathy (DSPN) is one of most common long-term complications of diabetes. Up to 50% of patients with diabetes suffer from DSPN, and an estimated one in five patients with diabetes have chronic painful neuropathy (1). Accurate detection and assessment of neuropathy would have a major medical, social, and economic effect in relation to earlier diagnosis and timely intervention to prevent progression and the difficulties with end points used in clinical trials of DSPN (2) to address the major unmet need of a treatment for this condition (3,4).

[1]Centre for Imaging Sciences, Institute of Population Health, University of Manchester, Manchester, U.K.
[2]Centre for Endocrinology and Diabetes, Institute of Human Development, Manchester Academic Health Science Centre, Manchester, U.K.
[3]Division of Medicine, Weill Cornell Medical College in Qatar, Doha, Qatar
[4]Department of Clinical Neurophysiology, Central Manchester NHS Foundation Trust, Manchester, U.K.
[5]Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, Queensland, Australia

Corresponding author: Rayaz A. Malik, ram2045@qatar-med.cornell.edu or rayaz.a.malik@manchester.ac.uk.

Received 13 October 2014 and accepted 26 February 2015.

© 2015 by the American Diabetes Association. Readers may use this article as long as the work is properly cited, the use is educational and not for profit, and the work is not altered.

Methods to quantify neuropathy include clinical scores based on symptoms and neurological tests, quantitative sensory testing (QST), electrophysiological measurements, in the form of nerve conduction studies (NCS), and intraepidermal nerve fiber density (IENFD) in skin biopsy specimens (5). The neurological examination involves an assessment, such as the modified Neuropathy Disability Score (NDS) (6), a composite score that assesses touch, temperature, and vibration perception and reflexes, which requires expert clinical judgment, a strong element of subjectivity, and hence, poor reproducibility (7). Neurophysiology is objective and reproducible and is currently considered to be the most reliable measurement for confirming the diagnosis of diabetic neuropathy and indeed represents an essential part of the Toronto Criteria (TC) to identify those with "confirmed DSPN: the presence of an abnormality of NC[S] and a symptom or symptoms or a sign or signs of neuropathy" (8). However, these measures mainly assess large nerve fibers, making them less sensitive to early DSPN, which is more likely to involve small fibers (9,10).

Small fibers can be assessed by quantifying thermal thresholds (11) and IENFD in skin biopsy specimens (12). Although QST assessment has been shown to have good repeatability (11), IENFD is considered to be the most objective and quantitative for the diagnosis of small fiber neuropathy (13,14). However, its invasive nature makes it unsuitable for repeated investigations (12). Furthermore, the reliability of IENFD for the diagnosis of DSPN has never been thoroughly validated in a large cohort of patients with diabetes (15). Thus diabetic neuropathy currently lacks a noninvasive surrogate for accurately detecting small nerve fiber damage and repair.

Several studies (16–20) have shown that corneal confocal microscopy (CCM) is capable of making a quantitative assessment of small fiber damage and has the potential to be a surrogate end point for DSPN (9). Quantitative analysis using manual annotation of CCM images to identify fibers and branches is labor-intensive and subjective. However, a fully automated nerve fiber quantification method has been shown to have high correlation with the manually obtained measurements (21,22), and our recent study (23) compared manual and automated image analysis in a large cohort of patients with diabetes. We previously assessed CCM and IENFD in the same patients and showed that the measures were related (17). However, to date there has been no attempt to directly compare the ability of CCM and IENFD in the diagnosis of DSPN. In this report, we comprehensively evaluate manually and automatically quantified CCM-derived measures of nerve fiber morphology and compare their diagnostic performance with IENFD measurements according to the presence or absence of DSPN using the TC.

RESEARCH DESIGN AND METHODS

Study Subjects

The study recruited 63 patients with type 1 diabetes from clinics of the Manchester Diabetes Center, Manchester Royal Infirmary, and age-matched control subjects from the community. The updated TC was used to assess all subjects for the presence and severity of DSPN between 2010 and 2011 (8). This research adhered to the tenets of the Declaration of Helsinki and was approved by the North Manchester Research Ethics Committee. Informed written consent was obtained from all participants before their enrollment in the study. All assessments were performed by trained staff in a purpose-designed clinical research facility in central Manchester. Inclusion criteria were age between 14 and 85 years and a history of type 1 diabetes. Exclusion criteria were a positive history of malignancy, connective tissue or infectious disease, deficiency of vitamin B_{12} or folate, chronic renal failure, liver failure, active diabetic foot ulceration, family history of peripheral neuropathy, active ocular disease, systemic disease known to affect the cornea other than diabetes, or chronic corneal pathologies. All participants underwent assessment of glycated hemoglobin (HbA_{1c}), HDL and LDL cholesterol, triglycerides, BMI, and renal status (estimated glomerular filtration rate and albumin-to-creatinine ratio). Participants in this study represent a subcohort of participants with type 1 diabetes (n = 110) and control subjects (n = 97) who agreed to undergo skin biopsy in addition to routine neurological testing.

Peripheral Neuropathy Assessment

All study participants underwent an assessment of neurological deficits (NDS) (6) and symptoms (Diabetic Neuropathy Symptom [DNS] score) (24). Vibration perception threshold (VPT) was tested using a Horwell Neurothesiometer (Scientific Laboratory Supplies, Nottingham, U.K.). Cold thresholds (CT) and warm thresholds (WT) were established on the dorsolateral aspect of the left foot (S1) using the TSA-II NeuroSensory Analyzer (Medoc Ltd., Ramat-Yishai, Israel). Electrodiagnostic studies were undertaken using a Dantec Keypoint system (Dantec Dynamics Ltd., Bristol, U.K.) equipped with a DISA temperature regulator to keep limb temperature constantly between 32° and 35°C. Sural sensory nerve amplitude (SSNamp), sural sensory nerve conduction velocity (SSNCV), peroneal motor nerve amplitude (PMNamp), and peroneal motor nerve conduction velocity (PMNCV) were assessed by a consultant neurophysiologist.

The Toronto Diabetic Neuropathy Expert Group (8) recommendation was followed to define an individual as having neuropathy if he or she met both of the following criteria: *1*) abnormal nerve conduction—a PMNCV of $<$42 m/s; *2*) a symptom or sign of neuropathy, defined as one of the following: *a*) DNS of 1 or more of 4, or *b*) NDS of 3 or more of 10.

For the IENFD assessment, a 3-mm punch skin biopsy specimen was obtained from the dorsum of the foot, and a bright-field immunohistochemistry protocol was used according to published guidelines (12). Linear IENFD (number of fibers/mm) was established in at least four sections of 50-μm thickness according to published counting rules (IENFD have to cross or originate at the dermal–epidermal junction, and secondary branches and fragments are not counted) (14). The assessments were performed by two experts (M.J. and R.A.M.) who were masked to the neuropathic/diabetes status of participants and were cross-validated.

Manual and Automated Quantification of Corneal Nerves

CCM images (Fig. 1*A*) were captured from all participants using the Heidelberg Retina Tomograph Rostock Cornea Module (HRT-III), as described (23,25), by two purpose-trained optometrists (I.N.P. and M.T.). Their dimensions are 384 × 384 pixels with the pixel size of 1.0417 μm. During a bilateral CCM scan, more than 100 images per patient were

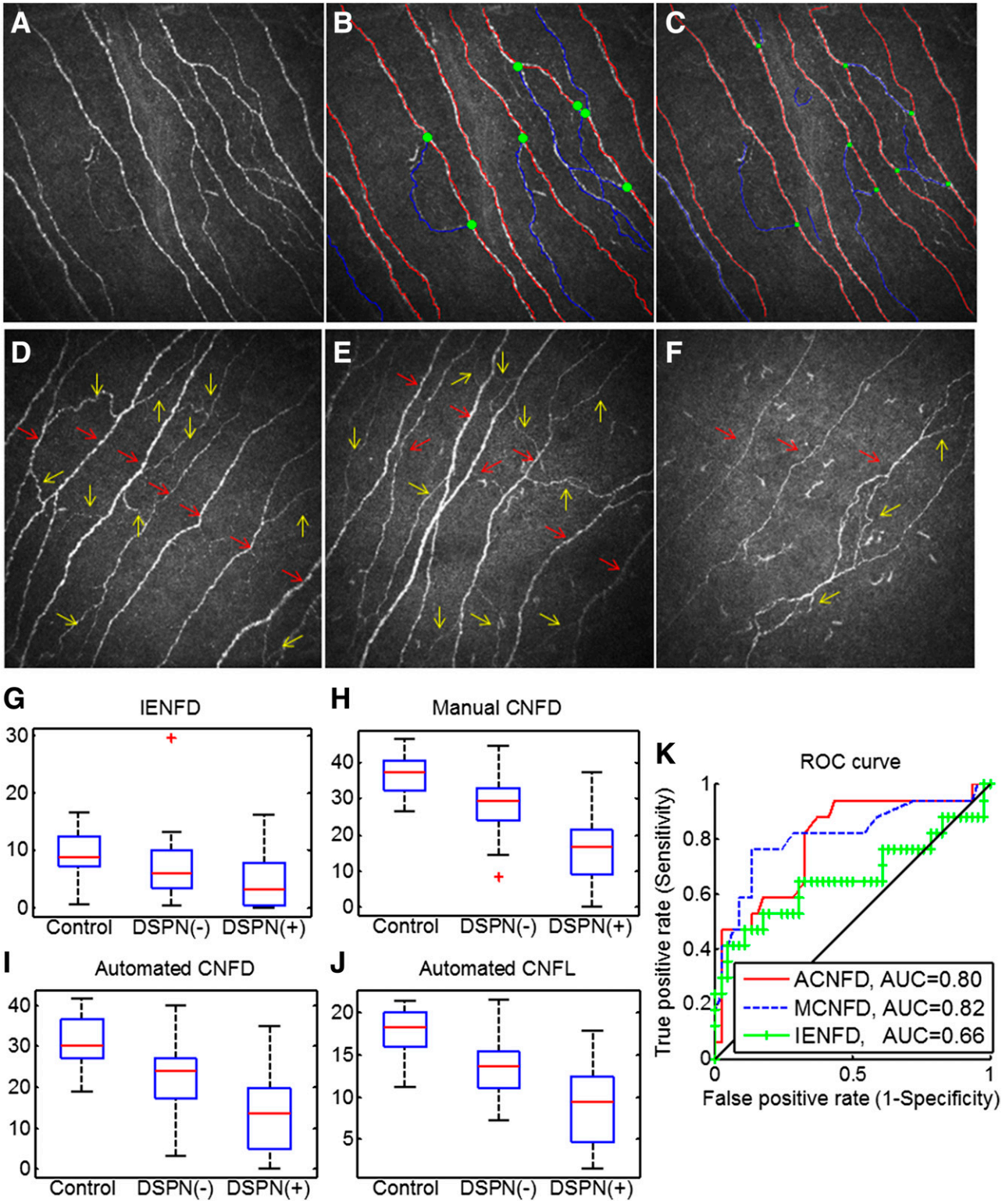


Figure 1—*A*: Original CCM image. *B*: Manually quantified CCM image. *C*: Automatically quantified CCM image. The red lines represent main nerve fibers, blue lines are branches, and green spots indicate branch points on the main nerve trunks. CCM images of the subbasal nerve plexus from a control subject (*D*), a DSPN(−) patient with type 1 diabetes (*E*), and a DSPN(+) patient with type 1 diabetes (*F*) show the reduction in corneal nerves in the DSPN(+) patient. The red arrows indicate main nerve fibers (to calculate CNFD), and yellow arrows indicate branch fibers (to calculate CNBD). Box plots of IENFD (*G*), manual CNFD values (*H*), automated CNFD (*I*), and automated CNFL (*J*) values in controls and in DSPN(−) and DSPN(+) patients with type 1 diabetes based on the TC. *K*: ROC curves for manual CNFD (MCNFD), automated CNFD (ACNFD), and IENFD to discriminate DSPN(+) and DSPN(−) patients with diabetes. *G*–*J*: Red lines represent median, the box borders 25th and 75th percentile. Whiskers represent the range of the data (without outliers). Red plus symbols represent outliers.

typically captured from all corneal layers, and 6 subbasal images from the right and left eyes were selected for analysis. Criteria for image selection were depth, focus position, and contrast. One experienced examiner (I.N.P.), masked from the outcome of the medical and peripheral neuropathy assessment, manually quantified 1,506 images of all study participants using purpose-written, proprietary software (CCMetrics, M.A. Dabbah, Imaging Science, University of Manchester) (Fig. 1*B*). The specific parameters measured per frame were corneal nerve fiber density (CNFD) (number of main fibers per mm^2), corneal nerve fiber length (CNFL) (total length of main fibers and branches per mm^2), and corneal nerve branch density (CNBD) (number of branches per mm^2) in accordance with our previously published protocol (23,25).

Automated corneal nerve fiber quantification consists of two steps: *1*) CCM image enhancement and nerve fiber detection and *2*) quantification of the three morphometric parameters. As described in our earlier work (21), a dual-model feature descriptor combined with a neural network classifier was used to train the detection software to distinguish nerve fibers from the background (noise and underlying connective tissue). In the nerve fiber quantification process, all of the end points and branch points of the detected nerve fibers are extracted and used to construct a connectivity map. Each segment in the connectivity map is then connected and classified as a main nerve fiber or branch (Fig. 1*C*). The software for automated CCM image quantification (ACCMetrics) is available via http://www.click2go.umip.com/i/software/Biomedical_Software/accmetrics_v2.html.

To evaluate the effectiveness of using IENFD and manually and automatically generated CCM features to diagnose DSPN, we used the TC as ground truth to categorize the subjects with diabetes into those with DSPN (DSPN[+]) and without DSPN (DSPN[−]).

Statistical Analysis

Statistical analysis and the receiver operating characteristic (ROC) curves were performed and generated using MATLAB R2012a software (The MathWorks, Inc.). One-way ANOVA (nonparametric Kruskal-Wallis) was used to evaluate within- and between-group differences (control group, the DSPN[+] group, and the DSPN[−] group). A $P < 0.05$ was considered significant. The area under the ROC curve (AUC) values, 95% CIs, and sensitivity and specificity at the equal error-rate point and at the threshold of 2 standard deviations below the mean of the control group were calculated for comparison. MedCalc 14.12.0 software (MedCalc Software bvba) was used to compare the difference between two ROC curves. The power analysis was performed using G*Power 3.1.9.2 software. The power analysis was performed based on the Wilcoxon-Mann-Whitney test comparing the group with type 1 diabetes and the control group. For PMNCV, the power was 0.999 (assuming an error rate α = 0.01), indicating that 26 control subjects and 63 patients with type 1 diabetes were sufficient to find a statistically significant difference. Then the power analysis was performed based on the Wilcoxon-Mann-Whitney test comparing DSPN(−) and DSPN(+) groups. For PMNCV, the power was 0.999 (assuming an error rate α = 0.01), indicating that a sample size of 46 DSPN(−) and 17 DSPN(+) was sufficient to find a statistically significant difference.

RESULTS

Demographics, Metabolic, and Anthropometric Assessment

The demographics and metabolic and anthropometric measurements in patients with diabetes and control subjects are summarized in Table 1. In the patients with type 1 diabetes, 57% were on a multiple daily insulin injection regimen, and 43% were on continuous subcutaneous insulin infusion. Other medications included an ACE inhibitor or angiotensin receptor blocker in 36%

Table 1—Clinical demographic results and neuropathy assessment in control subjects and in DSPN(−) and DSPN(+) patients with type 1 diabetes

Variable	Control subjects (*n* = 26)	DSPN(−) (*n* = 46)	DSPN(+) (*n* = 17)
Age, years	44 ± 15	44 ± 13	59 ± 11
Duration of diabetes, years	N/A	23 ± 15	39 ± 14
HbA_{1c} (%)‡	5.5 ± 0.3	8.2 ± 1.4	8.5 ± 1.3
HbA_{1c} (mmol/mol)‡	37.1 ± 3.5	62.2 ± 24.1¶	69.3 ± 14.3¶
BMI (kg/m^2)*	26.8 ± 4.0	26.4 ± 4.5	27.5 ± 3.5¶
Cholesterol (mmol/L)			
Total*	5.0 ± 0.8	4.4 ± 0.9¶	4.3 ± 0.9¶
HDL	1.5 ± 0.3	1.6 ± 0.5	1.6 ± 0.4
Triglycerides (mmol/L)	1.4 ± 0.7	1.2 ± 0.7	1.3 ± 0.6
Blood pressure (mmHg)			
Systolic†	126.7 ± 16.3	130.3 ± 17.8¶	141.1 ± 25.2¶§
Diastolic	70.2 ± 9.1	71.6 ± 9.6	73.0 ± 9.8
VPT (V)‡	6.0 ± 5.5	7.6 ± 5.5	25.2 ± 13.4¶§
WT (°C)†	36.4 ± 2.0	38.7 ± 3.6¶	43.5 ± 4.6¶§
CT (°C)†	28.8 ± 1.6	27.1 ± 2.7¶	16.8 ± 10.6¶§
PMNCV (m/s)‡	49.1 ± 3.4	43.9 ± 3.1¶	31.0 ± 9.5¶§
SSNCV (m/s)‡	50.9 ± 3.9	45.3 ± 5.2¶	37.8 ± 6.8¶§
PMNamp (μV)‡	6.0 ± 2.4	6.0 ± 8.3	1.6 ± 1.6¶§
SSNamp (μV)‡	19.7 ± 8.3	12.5 ± 6.9¶	4.3 ± 3.5¶§
IENFD¿	9.8 ± 3.7	7.0 ± 5.0¶	5.0 ± 5.5¶§
Manual			
CNFD‡	36.8 ± 5.3	28.3 ± 7.2¶	16.9 ± 10.1¶§
CNBD*	92.8 ± 36.4	56.1 ± 30.3¶	48.2 ± 32.9¶
CNFL‡	26.7 ± 3.7	20.2 ± 5.1¶	14.8 ± 8.3¶§
Automated			
CNFD‡	31.3 ± 6.5	22.6 ± 7.3¶	13.5 ± 9.1¶§
CNBD‡	44.6 ± 17.2	26.2 ± 15.1¶	15.4 ± 12.1¶§
CNFL‡	17.7 ± 2.8	13.4 ± 3.3¶	8.8 ± 4.7¶§

Results are expressed as mean ± SD. N/A, not applicable for this group. Statistically significant differences using ANOVA/Kruskal-Wallis: *$P < 0.05$; ¿$P < 0.01$; †$P < 0.001$; ‡$P < 0.0001$. Post hoc results for DSPN(+) significantly different from ¶control subjects and §DSPN(−).

of subjects and statins in 71%. Age was comparable between control subjects and patients with diabetes. HbA_{1c} was significantly higher in patients with diabetes than in control subjects, with no difference between DSPN(+) and DSPN(−) patients. BMI was significantly higher in DSPN(+) patients with diabetes compared with control subjects. Total cholesterol was significantly lower in DSPN(+) and DSPN(−) patients with diabetes, whereas HDL and triglycerides did not differ between the groups. Systolic blood pressure was significantly higher in DSPN(+) and DSPN(−) patients with diabetes compared with control subjects, whereas diastolic blood pressure did not differ between groups.

Neurological Assessment

The NDS differed significantly between DSPN(+) patients and control subjects (Table 1).

QST

VPT was significantly greater in DSPN(+) patients compared with control subjects and DSPN(−) patients (Table 1). CT and WT both differed significantly in DSPN(+) and DSPN(−) patients with diabetes compared with control subjects.

Electrophysiology

PMNCV, SSNCV, and SSNamp were significantly reduced in DSPN(−) patients with diabetes compared with control subjects (Table 1). PMNCV, SSNCV, PMNamp, and SSNamp were all reduced in DSPN(+) patients with diabetes compared with control subjects and DSPN(−) patients with diabetes.

IENFD

IENFD was significantly reduced in DSPN(+) patients (P = 0.002) and in DSPN(−) patients (P = 0.001), and was further reduced in DSPN(+) compared with DSPN(−) patients (P = 0.05) (Table 1 and Fig. 1*G* and Fig. 2). The median value of the control group was 9.35 and the 0.05 quantile was 4.31, which is consistent with previously published IENFD measurements (12).

CCM

Manual CNFD was significantly reduced in DSPN(+) patients ($P < 0.0001$) and in DSPN(−) patients ($P < 0.0001$) compared with control subjects and was further reduced in DSPN(+) patients compared with DSPN(−) patients ($P < 0.0001$) (Table 1 and Fig. 1*H*). Manual CNBD was significantly reduced in DSPN(+) patients ($P < 0.0001$) but not in DSPN(−) patients (P = 0.09) compared with control subjects. Manual CNFL was significantly reduced in DSPN(+) patients ($P < 0.0001$) and in DSPN(−) patients ($P < 0.0001$) compared with control subjects and was further reduced in DSPN(+) patients compared with DSPN(−) patients (P = 0.001). Automated CNFD was significantly reduced in DSPN(+) patients ($P < 0.0001$) and DSPN(−) patients ($P < 0.0001$) compared with control subjects and was further reduced in DSPN(+) patients compared with DSPN(−) patients ($P < 0.0001$) (Fig. 1*I*). Automated CNBD was significantly reduced in DSPN(+) patients ($P < 0.0001$) and DSPN(−) patients ($P < 0.0001$) compared with control subjects and was further reduced in DSPN(+) patients compared with DSPN(−) patients (P = 0.002). Automated CNFL was significantly reduced in DSPN(+) patients ($P < 0.0001$) and DSPN(−) patients ($P < 0.0001$) compared with control subjects and was further reduced in DSPN(+) patients compared with DSPN(−) patients ($P < 0.0001$) (Fig. 1*J*).

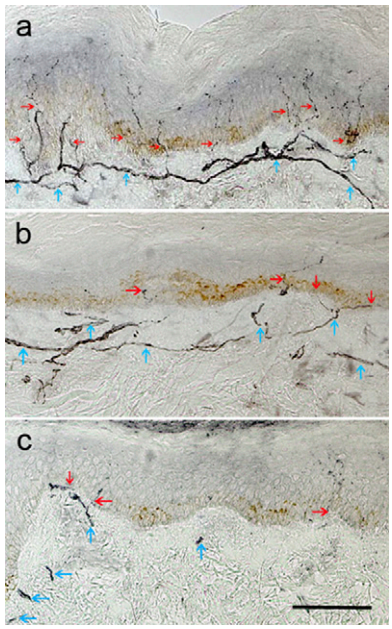


Figure 2—Skin biopsy specimens immunostained for neuronal marker PGP 9.5 from a healthy subject (*A*), a DSPN(−) patient with type 1 diabetes (*B*), and a DSPN(+) patient with type 1 diabetes (*C*). Note the depletion of IENFD (red arrows) and reduction of subepidermal nerve plexus (blue arrows) in *B* and *C*, with both features more severe in the DSPN(+) patient (*C*). Original magnification ×200, scale bar = 100 μm.

ROC Analysis

The patients with diabetes were categorized into DSPN(−) (n = 46) and DSPN(+) (n = 17). Table 2 reports the AUC values, 95% CIs, and sensitivity/specificity at the equal error-rate point on the ROC curve for manual and automated CCM features as well as IENFD values. The highest AUC values among the manual and automated CCM measures were obtained for CNFD, with AUC values of 0.82 and 0.80, respectively. Almost all individual CCM measurements resulted in higher AUC values than IENFD (0.66). Furthermore, sensitivity and specificity values were calculated at the equal error-rate point for the purpose of consistency. For this measure of diagnostic performance also, CNFD provided the best discrimination (76% for manual measurement and 70% for automated measurement), which exceeded the 65% achieved by IENFD.

In using IENFD to identify DSPN, a decision threshold for neuropathy is commonly set at 2 standard deviations below the mean of the control group. Table 2 also reports the sensitivity/specificity values obtained by applying this

Table 2—AUC, 95% CI values, and sensitivity-specificity for manual and automated CCM and IENFD for the diagnosis of DSPN

	AUC	95% CI	Sensitivity-specificity at equal-error rate	Sensitivity/specificity at mean ± 2 SD (threshold)
Manual				
CNFD	0.82	0.68–0.95	0.76	0.82/0.71 (24.0)
CNFL	0.70	0.54–0.85	0.71	0.59/0.74 (16.5)
CNBD	0.59	0.43–0.75	0.53	0.17/0.96 (15.0)
Automated				
CNFD	0.80	0.66–0.93	0.70	0.60/0.83 (15.5)
CNFL	0.77	0.63–0.91	0.70	0.59/0.80 (10.5)
CNBD	0.70	0.55–0.86	0.59	0.29/0.98 (4.0)
IENFD	0.66	0.50–0.82	0.65	0.53/0.76 (3.3)

threshold. When this threshold was used, manual CNFD and automated CNFD result in better sensitivity/specificity than IENFD: 0.82/0.71, 0.60/0.83, and 0.53/0.76, respectively. There were no statistically significant differences between the ROC curves for manual CNFD and IENFD (P = 0.14) and for automated CNFD and IENFD (P = 0.19) (26). However, CCM measurements show considerably less variability within the subject groups than IENFD measurements (Fig. 1*G*) and larger AUC values (Fig. 1*K*).

CONCLUSIONS

There is a need for surrogate end points of diabetic neuropathy that accurately detect early disease, quantify disease progression, and measure therapeutic response (2). The current gold standard for the diagnosis of neuropathy, neurophysiology, is a robust measure but has poor reproducibility (27). Other measures of neuropathy, such as symptoms and signs, are also poorly reproducible (7), and although QST is reproducible, it is subjective (11).

Small fiber neuropathy has direct pathophysiological relevance to the main outcomes of pain and foot ulceration. Skin biopsy assessment of IENFD has been proposed as a valid measure of diabetic neuropathy (15). Furthermore, skin biopsy detects early small nerve fiber damage even when results of electrophysiology and QST are still within normal ranges (28,29), suggesting that it could detect early neuropathy. It has been shown to be abnormal in subjects with IGT (19) and in recently diagnosed patients with type 2 diabetes (30). IENFD has also been shown to increase with an improvement in metabolic risk factors in subjects with IGT (31) but not after combined pancreas and kidney transplantation in patients with type 1 diabetes (20). Furthermore, the invasive nature of this technique limits its practical use as a diagnostic test and particularly when a repeat biopsy is required in longitudinal studies or clinical intervention trials.

CCM is a novel, rapid, and readily reiterative technique that quantifies small nerve fibers noninvasively and shows promise as a surrogate end point for neuropathy (9,18,30,32–34). A number of studies have shown the nerve fiber features extracted from CCM are associated with the severity of diabetic peripheral neuropathy (17,23,33).

Because IENFD represents a measure of the most distal nerve fibers, which are affected in DSPN, a natural assumption is that it should have a better diagnostic ability than CCM. However, a comparison between IENFD and CCM features for the individual diagnosis of DSPN has not been reported to date. In this report, we present a comparison of nerve fiber features, quantified manually or automatically from CCM images (CNFL, CNFD, and CNBD) with IENFD measurement in identifying DSPN in individuals. CCM and IENFD are comparable in their diagnostic performance for detecting patients with diabetic neuropathy. Neither technique appears to have an optimal diagnostic performance. However, there were relatively small numbers of patients in the study because a significant proportion were not willing to undergo biopsy. Furthermore, the diagnosis of DSPN does not incorporate a measure of small fiber damage, which limits the assessment of the diagnostic performance of these small fiber tests. The added advantage of CCM compared with IENFD assessment is the more rapid and noninvasive acquisition of images and automated corneal nerve image analysis allowing rapid and consistent quantification (22,23,35). The exception is the manually measured CNBD, which has been found previously (25) to be unreliable due to the subjective judgment required in identifying branches. The algorithmic definition of branches in the automated measurement results in greater consistency, although this is the least useful individual automated CCM measurement. CCM and IENFD both seek to measure small fibers, but IENFD showed a poorer discrimination between DSPN(+) and DSPN(−) patients. Furthermore, CCM measurements show considerably less variability within the subject groups than IENFD measurements. Interestingly, very low IENFD values were observed, even in control subjects.

This study has strengths and limitations. Strengths include the study design and techniques used to assess neuropathy. This is the first study to report the clinical utility of two highly sensitive techniques, CCM and skin biopsy, in the same group of patients with type 1 diabetes and control subjects. Thus, CCM appears to be an emerging surrogate end point of diabetic neuropathy that shows comparable performance to the current gold standard of IENFD.

The limitations of the current study are the relatively small number of patients with established neuropathy and the use of the more distal site for the biopsy, which makes comparison of the IENFD results with other studies difficult. Furthermore, these data are only applicable to Caucasian patients with type 1 diabetes and need to be confirmed in nondiabetic neuropathies.

In conclusion, we show that the diagnostic efficiency of CCM is comparable to IENFD. However, CCM may be preferred due to its rapid, noninvasive, automated and, hence, unbiased means of quantifying small nerve fiber damage and repair in DSPN(+) patients.

Acknowledgments. This research was facilitated by the Manchester Biomedical Research Centre and the Greater Manchester Comprehensive Local Research Network.

Funding. This research was funded by awards from the National Institutes of Health (R01-DK-077903-0101) and JDRF (27-2008-362).

Duality of Interest. No potential conflicts of interest relevant to this article were reported.

Author Contributions. X.C. developed the automated CCM software, performed statistical analysis, and wrote the manuscript. J.G. contributed to manual and automated software development and reviewed and edited the manuscript. M.A.D. developed the manual and automated CCM software. I.N.P. generated the CCM data and performed the CCM data analysis. G.P. researched data and coordinated patient assessment. O.A., A.M., H.F., and S.A. researched data. U.A. recruited patients and researched data. M.F. and M.T. generated the CCM data. N.E. designed the study and reviewed and edited the manuscript. M.J. generated IENFD data and reviewed and edited the manuscript. R.A.M. designed and oversaw the study, generated IENFD data, and reviewed and edited the manuscript. R.A.M. is the guarantor of this work and, as such, had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

References

1. Abbott CA, Malik RA, van Ross ER, Kulkarni J, Boulton AJ. Prevalence and characteristics of painful diabetic neuropathy in a large community-based diabetic population in the U.K. Diabetes Care 2011;34:2220–2224

2. Dyck PJ, Norell JE, Tritschler H, et al. Challenges in design of multicenter trials: end points assessed longitudinally for change and monotonicity. Diabetes Care 2007;30:2619–2625

3. Malik RA. Why are there no good treatments for diabetic neuropathy? Lancet Diabetes Endocrinol 2014;2:607–609

4. Boulton AJ, Kempler P, Ametov A, Ziegler D. Whither pathogenetic treatments for diabetic polyneuropathy? Diabetes Metab Res Rev 2013;29:327–333
5. Dyck PJ, Herrmann DN, Staff NP, Dyck PJ. Assessing decreased sensation and increased sensory phenomena in diabetic polyneuropathies. Diabetes 2013;62:3677–3686
6. Young MJ, Boulton AJ, MacLeod AF, Williams DR, Sonksen PH. A multicentre study of the prevalence of diabetic peripheral neuropathy in the United Kingdom hospital clinic population. Diabetologia 1993;36:150–154
7. Dyck PJ, Overland CJ, Low PA, et al.; Cl vs. NPhys Trial Investigators. Signs and symptoms versus nerve conduction studies to diagnose diabetic sensorimotor polyneuropathy: Cl vs. NPhys trial. Muscle Nerve 2010;42:157–164
8. Tesfaye S, Boulton AJ, Dyck PJ, et al.; Toronto Diabetic Neuropathy Expert Group. Diabetic neuropathies: update on definitions, diagnostic criteria, estimation of severity, and treatments. Diabetes Care 2010;33:2285–2293
9. Malik RA. Which test for diagnosing early human diabetic neuropathy? Diabetes 2014; 63:2206–2208
10. Breiner A, Lovblom LE, Perkins BA, Bril V. Does the prevailing hypothesis that small-fiber dysfunction precedes large-fiber dysfunction apply to type 1 diabetic patients? Diabetes Care 2014;37:1418–1424
11. Dyck PJ, Argyros B, Russell JW, et al.; Members of the Cl versus NPhys Trials. Multicenter trial of the proficiency of smart quantitative sensation tests. Muscle Nerve 2014;49:645–653
12. Lauria G, Lombardi R. Small fiber neuropathy: is skin biopsy the holy grail? Curr Diab Rep 2012;12:384–392
13. Hoeijmakers JG, Faber CG, Lauria G, Merkies IS, Waxman SG. Small-fibre neuropathies—advances in diagnosis, pathophysiology and management. Nat Rev Neurol 2012;8:369–379
14. Polydefkis M, Hauer P, Sheth S, Sirdofsky M, Griffin JW, McArthur JC. The time course of epidermal nerve fibre regeneration: studies in normal controls and in people with diabetes, with and without neuropathy. Brain 2004;127:1606–1615
15. Malik RA, Veves A, Tesfaye S, et al.; Toronto Consensus Panel on Diabetic Neuropathy. Small fibre neuropathy: role in the diagnosis of diabetic sensorimotor polyneuropathy. Diabetes Metab Res Rev 2011;27:678–684
16. Malik RA, Kallinikos P, Abbott CA, et al. Corneal confocal microscopy: a non-invasive surrogate of nerve fibre damage and repair in diabetic patients. Diabetologia 2003;46:683–688
17. Quattrini C, Tavakoli M, Jeziorska M, et al. Surrogate markers of small fiber damage in human diabetic neuropathy. Diabetes 2007;56: 2148–2154
18. Pritchard N, Edwards K, Dehghani C, et al. Longitudinal Assessment of Neuropathy in type 1 Diabetes using novel ophthalmic Markers (LANDMark): study design and baseline characteristics. Diabetes Res Clin Pract 2014;104:248–256
19. Asghar O, Petropoulos IN, Alam U, et al. Corneal confocal microscopy detects neuropathy in subjects with impaired glucose tolerance. Diabetes Care 2014;37:2643–2646
20. Tavakoli M, Mitu-Pretorian M, Petropoulos IN, et al. Corneal confocal microscopy detects early nerve regeneration in diabetic neuropathy after simultaneous pancreas and kidney transplantation. Diabetes 2013;62:254–260
21. Dabbah MA, Graham J, Petropoulos I, Tavakoli M, Malik RA. Dual-model automatic detection of nerve-fibres in corneal confocal microscopy images. Med Image Comput Comput Assist Interv 2010;13:300–307
22. Dabbah MA, Graham J, Petropoulos IN, Tavakoli M, Malik RA. Automatic analysis of diabetic peripheral neuropathy using multi-scale quantitative morphology of nerve fibres in corneal confocal microscopy imaging. Med Image Anal 2011;15:738–747
23. Petropoulos IN, Alam U, Fadavi H, et al. Rapid automated diagnosis of diabetic peripheral neuropathy with in vivo corneal confocal microscopy. Invest Ophthalmol Vis Sci 2014; 55:2071–2078
24. Meijer JW, Smit AJ, Sonderen EV, Groothoff JW, Eisma WH, Links TP. Symptom scoring systems to diagnose distal polyneuropathy in diabetes: the Diabetic Neuropathy Symptom score. Diabet Med 2002;19:962–965
25. Petropoulos IN, Manzoor T, Morgan P, et al. Repeatability of in vivo corneal confocal microscopy to quantify corneal nerve morphology. Cornea 2013;32:e83–e89
26. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36
27. Dyck PJ, Albers JW, Wolfe J, et al.; Clinical vs. Neurophysiology Trial 3 Investigators. A trial of proficiency of nerve conduction: greater standardization still needed. Muscle Nerve 2013;48:369–374
28. Sumner CJ, Sheth S, Griffin JW, Cornblath DR, Polydefkis M. The spectrum of neuropathy in diabetes and impaired glucose tolerance. Neurology 2003;60:108–111
29. Singleton JR, Smith AG, Bromberg MB. Increased prevalence of impaired glucose tolerance in patients with painful sensory neuropathy. Diabetes Care 2001;24:1448–1453
30. Ziegler D, Papanas N, Zhivov A, et al.; German Diabetes Study (GDS) Group. Early detection of nerve fiber loss by corneal confocal microscopy and skin biopsy in recently diagnosed type 2 diabetes. Diabetes 2014;63: 2454–2463
31. Smith AG, Russell J, Feldman EL, et al. Lifestyle intervention for pre-diabetic neuropathy. Diabetes Care 2006;29:1294–1299
32. Tavakoli M, Petropoulos IN, Malik RA. Corneal confocal microscopy to assess diabetic neuropathy: an eye on the foot. J Diabetes Sci Tech 2013;7:1179–1189
33. Sivaskandarajah GA, Halpern EM, Lovblom LE, et al. Structure-function relationship between corneal nerves and conventional small-fiber tests in type 1 diabetes. Diabetes Care 2013;36:2748–2755
34. Halpern EM, Lovblom LE, Orlov S, Ahmed A, Bril V, Perkins BA. The impact of common variation in the definition of diabetic sensorimotor polyneuropathy on the validity of corneal in vivo confocal microscopy in patients with type 1 diabetes: a brief report. J Diabetes Complications 2013;27:240–242
35. Dehghani C, Pritchard N, Edwards K, Russell AW, Malik RA, Efron N. Fully automated, semi-automated, and manual morphometric analysis of corneal subbasal nerve plexus in individuals with and without diabetes. Cornea 2014;33: 696–702

Applications of Image Analysis: Carpal Kinematics

57. **Inferring 3D kinematics of carpal bones from single-view fluoroscopic sequences.** X. Chen, J. Graham, C.E. Hutchinson and L. Muir. *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI) 2011 Toronto, Canada. Part 2 Lecture Notes in Computer Science 6892 (Springer) G Fichtinger, A. Martel, T. Peters Eds, pp 680-687.* doi: 10.1007/978-3-642-23629-7_83

Inferring 3D Kinematics of Carpal Bones from Single View Fluoroscopic Sequences

Xin Chen[1], Jim Graham[1], Charles Hutchinson[2], and Lindsay Muir[3]

[1] ISBE, School of Cancer and Enabling Sciences, University of Manchester, Oxford Road, Manchester, M13 9PT, UK
{xin.chen,jim.graham}@manchester.ac.uk
[2] Clinical Sciences Research Institute, Clinical Sciences Building, University Hospital -Walsgrave Campus, Clifford Bridge Road, Coventry, CV2 2DX, UK
[3] Consultant Orthopaedic Surgeon, Salford Royal Hospital NHS Foundation Trust, Stott Lane, Salford, M6 8HD, UK

Abstract. We present a novel framework for inferring 3D carpal bone kinematics and bone shapes from a single view fluoroscopic sequence. A hybrid statistical model representing both the kinematics and shape variation of the carpal bones is built, based on a number of 3D CT data sets obtained from different subjects at different poses. Given a fluoroscopic sequence, the wrist pose, carpal bone kinematics and bone shapes are estimated iteratively by matching the statistical model with the 2D images. A specially designed cost function enables smoothed parameter estimation across frames. We have evaluated the proposed method on both simulated data and real fluoroscopic sequences. It was found that the relative positions between carpal bones can be accurately estimated, which is potentially useful for detection of conditions such as scapholunate dissociation.

Keywords: Carpal bones kinematics, 2D 3D registration, Statistical model.

1 Introduction

Chronic pain in the wrist arises due to a number of conditions, such as instability patterns, nonunion or malunion of fractures, primary osteoarthritis and inflammatory arthritis. The result for patients is a severe reduction in quality of life due to impairment of everyday functions, lost work time, increased morbidity and loss of the capacity to live independently. The current method of distinguishing between these conditions is by examining 2D video fluoroscopy sequences showing movement of the hand from full ulnar to full radial deviation and from full flexion to extension in two orthogonal views. From these images clinicians can infer the three-dimensional translations and rotations of the carpal bones that take place during wrist movement, and arrive at a differential diagnosis on the basis of variations from normal bone kinematics. The interpretation is difficult and the accuracy of diagnosis depends wholly on the experience of the practitioner. Currently, accurate diagnosis requires referral to a specialist hand consultant and treatment is often delayed to the detriment of the patient.

The aim of the project is computer interpretation of the fluoroscopy sequences to attain a higher degree of objectivity and quantification in the diagnostic process. During

G. Fichtinger, A. Martel, and T. Peters (Eds.): MICCAI 2011, Part II, LNCS 6892, pp. 680–687.
© Springer-Verlag Berlin Heidelberg 2011

wrist movement, the eight carpal bones follow a complex, multi-dimensional trajectory, making interpretation of radiographs difficult. For this study we have trained a hybrid statistical model (SM) from a set of CT images from different subjects at different poses. Subsequently, the full 3D carpal bone motions can be recovered by matching the SM with the fluoroscopy sequences through 3D-2D image registration techniques. A number of studies have sought to represent the carpal kinematics using CT or MR data, mainly concentrating on representing 'average' kinematics over a small number of individuals (e.g. [1], [2]). More recently, Van deGiessen et al. [3] presented a 3D rigid registration method based on segmented meshes, which aims to build SM of carpal bones. A study of carpal bone kinematics based on a 4D imaging system was reported in [4]. 3D-2D registration has been the subject of many studies (e.g. [5]), mainly in the field of registration of pre-operative MR or CT images to intra-operative 2D images. Our work differs from the above in that we seek to achieve registration of a 2D image sequence to a 3D model (not derived from the same individual) to derive the kinematics of an individual wrist. Zheng [6] took a similar approach to estimate the orientation of pelvis from a single X-ray image.

The main contributions of this paper, distinguishing it from these studies, are: (1) A hybrid SM is developed representing both the complex kinematics and shape variation of the eight carpal bones plus radius and ulna. (2) The full 3D motion and bone shapes are recovered by matching the SM with a single view fluoroscopy sequence: a difficult ill-posed problem. (3) Our initial results show that the relative positions between the carpal bones can be estimated accurately through the proposed framework. We are not aware of any study which attempts to make a 2D to 3D inference in a system of this level of complexity.

The system consists of a training phase and a 3D-2D image registration phase. We currently have CT data from 10 subjects, each at five poses (neutral pose and two extreme poses in flexion-extension and radial-ulnar deviation). In the training phase, only the data from the neutral pose and two extreme poses in the radial-ulnar movement were used, as the radial-ulnar movement is the current concern of this paper. The segmentation of each bone and rigid registration parameters that align bones at different poses within and across the subjects were obtained using an iterative segmentation and registration algorithm [7]. A hybrid statistical model, representing both the kinematics and shape variation, was built efficiently from the results of the segmentation-registration framework. The kinematic model was built based on the transformation parameters, while the segmentation result was used to build the statistical shape model for each individual bone. In the 3D-2D image registration phase, the 3D rigid transformation, the kinematic motion and bone shapes were estimated in sequence from each frame of the fluoroscopy sequences. Detailed descriptions are given in the following sections.

2 Problem Parameterisation

We use a perspective projection model to represent the relationship between the 2D fluoroscopy image and the 3D configuration of bones. Almost all parameters necessary for this model (pixel size and optical centre) are known. The distance from the X-ray source and the detector needs to be measured for each patient. If this parameter is not

accurate, it will lead to a scale difference of the estimated 3D model. The resulting translation effects on the relative motion between carpal bones at pixels away from the centre of the field is very small.

Three sets of parameters need to be estimated during image registration in order to interpret the true 3D motion of the carpal bones: (1) Rigid transformation parameters of the wrist and a global scale factor, denoted by $\theta=\{t_x, t_y, t_z, \alpha, \beta, \gamma, s_{global}\}$. t_x, t_y and t_z denote the translations, and α, β and γ denote the rotation angles. s_{global} controls the distance between the centroid of each bone to the origin in the radius, and the global size of the bones. (2) Kinematic model parameters M representing the carpal bone poses during movement. (3) Shape model parameters Q_i and scale factor s_i for each bone (i).

3 Training of Kinematic Model and Shape Model

We use the six rigid transformation parameters for each bone to train the kinematic model. The common coordinate system for all pose and scale parameters has an origin at the centroid of the head of radius for one subject. The pose of one subject is described by $(tx_1, ty_1, tz_1, \alpha_1, \beta_1, \gamma_1, ..., tx_{10}, ty_{10}, tz_{10}, \alpha_{10}, \beta_{10}, \gamma_{10})^t$. (8 carpal bones, 1 radius and 1 ulna). The orientation parameters all occupy values distant from the angular discontinuity. Then the kinematic model can be parameterised as,

$$M = \mu^m + \phi^m b^m \tag{1}$$

where the mean pose μ^m (m is a notation indicating the model parameters) and the principal subspace matrix ϕ^m are computed from 3 (poses)$\times$ 10 (subjects) training samples using PCA. The vector b^m represents the kinematic parameters that describe the pose of M along each principal direction. In our experiments, the first 8 significant modes are used, which keeps 98% of variation.

The statistical shape model of each bone is a point distribution model, built using the segmented volume of the same training subjects. The 3D structure of each bone is described by a set of approximately 1000 points on the segmented surface. Correspondence between these points across subjects was established by the minimum description length algorithm [8]. The deformable shape model is then described as,

$$Q_i = \mu_i^q + \phi_i^q b_i^q \tag{2}$$

where μ_i^q and ϕ_i^q (q is a notation indicating the shape parameters) are the mean shape and the principal subspace matrix for the i^{th} bone. b_i^q is the shape model parameter to be estimated. In order to keep the complexity within limits, only the first 3 significant modes are used which keeps 84% of variation.

Based on the point distribution model of each bone and the kinematic model, a hybrid statistical mesh model can be built by using the Crust mesh construction algorithm [9]. Figure 1 shows the poses of the first mode of the kinematic model (represented by the mean shapes of each bone) and the shapes of the first mode of the scaphoid.

4 3D-2D Image Registration

The statistical mesh model from the training data is then used to match with the fluoroscopic sequence to infer the 3D motion and bone shapes. Figure 2(a) summarises

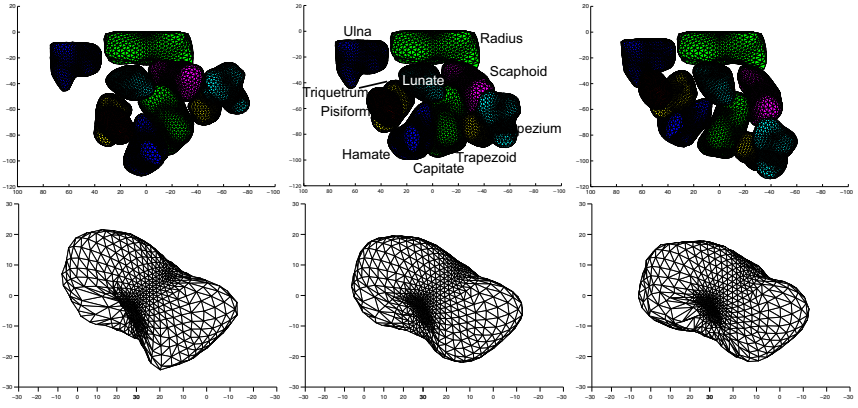


Fig. 1. Top row: The poses of the first mode of kinematic model. Bottom row: the first mode of the shape model of the scaphoid. In each case the mean +/-1.5s.d. are shown.

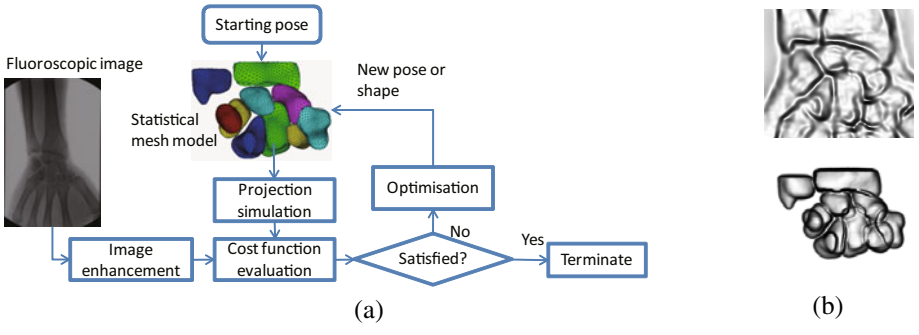


Fig. 2. Overview of the 3D-2D image registration process. (b) The gradient magnitude map of the fluoroscopic image after enhancement (cropped to show the region of interest) (top) and the simulated image from mesh model (bottom).

the registration process, in which the preprocessed fluoroscopic image is iteratively matched with a simulated projection generated from an updated pose of the mesh model. Detailed descriptions are given in the following subsections.

4.1 Fluoroscopic Image Enhancement and Projection Simulation

As the edges are strong features that can be used for image matching, the fluoroscopic image was firstly pre-processed to enhance the edges and reduce noise in homogenous regions. Local intensity normalisation was achieved by subtraction of the local mean intensity and division by the local standard deviation. The anisotropic diffusion [10] filter is then used to smooth the image while preserving the edges. Figure 2(b) shows an example of the gradient magnitude map of the fluoroscopic image after enhancement.

To optimise the pose parameters we iteratively generate projections from the mesh model with updated parameters, using the perspective projection described in section

2. The mesh model is considered to be a binary volume, and the projected intensity is negatively proportional to the sum of binary values along the ray from the source to each pixel in the image plane. Figure 2(b) shows an example.

4.2 Cost Function

To evaluate the similarity between the fluoroscopic image and the simulated image, we investigated several forms of the cost function, achieving best results from the one shown in Eqn. (3), based on the gradient along horizontal and vertical directions as well as the gradient magnitude of the two images. Additionally, the adjacent frames of the current fluoroscopic image were also taken into account in the cost function to make the estimated poses smooth across frames.

Taking C(A,B) as the Normalised Correlation Coefficient between two images A and B, we can write the cost function as:

$$E = C(Om_{k-1}, Om_k) + \sum_{p=k-1,k,k+1} w_p(C(Im_p, Dm_k) + C(Ix_p, Dx_k) + C(Iy_p, Dy_k)) \quad (3)$$

where k is the current frame number of the fluoroscopic sequence. Im_p, Ix_p and Iy_p are the gradient magnitude image, vertical gradient and horizontal gradient of the fluoroscopic image at the p^{th} frame respectively. Dm_k, Dx_k and Dy_k are the corresponding values of the simulated image. The second term calculates a cross-correlation between sets of three adjacent frames with weights w_{k-1}, w_k and w_{k+1}= 0.2, 0.6, 0.2 respectively, making the estimated pose smooth across frames. For the first term of the cost function, the vertices in the statistical mesh model are projected to the image plane, we assume the intensities at those projected points are similar across adjacent frames. Om_{k-1} and Om_k represent the gradient magnitude of the previous frame and the current frame at the projected correspondence positions. The first term makes the shape of the cost function sharper, leading to a faster and more accurate optimisation result. The $(k-1)^{th}$ frame and $(k+1)^{th}$ frame are not evaluated for the first and last frame respectively.

4.3 Optimisation

The optimisation method used is the best neighbour search combined with parabola fitting. The multi-dimensional search space (θ, M and Q) is explored by iteratively individual 1D line search. The cost function is evaluated at the current position, positive and negative neighbour positions (defined by a search range), then an optimum is found by fitting a parabola to the 3 evaluated positions. The optimum is iteratively refined by reducing the search range until convergence.

In our case, the true sizes of the bones are unknown; recovering the 3D pose from a single image is therefore a difficult, ill posed, problem. Any movement along the out-of-plane translation, could be compensated by scaling of the bone. In order to minimise this effect, the optimisation is carefully sequenced. We firstly assume that the wrist is not moving along the out-of-plane direction during radial-ulnar movement (t_y=0), as it is placed on a flat surface. The position of the model is firstly initialised by clicking the centre of the radius in the first frame of the fluoroscopic sequence. In the first step of the optimisation, only the first frame of the fluoroscopic sequence is used, and only the in-plane rigid transformation parameters (t_x, t_z, β) are estimated along with the global scale

factor (s_{global}) and the relative scale parameters of each bone (s_i). The first significant parameter of the kinematic model (b^m) is also estimated to provide an estimate of the overall pose. Other, less significant modes may include components of deviation along the out-of-plane direction that would affect the estimation of the global scale parameter. Inclusion of this first step resulted in significantly lower estimated error along the out-of-plane direction than optimisation without this step. Starting from this initial estimate of pose, the first frame is evaluated again, taking all the parameters into account (except t_y) in the following sequence: t_x, t_z, β, α, γ, b^m, s_{global}, s_i and b_i^q. After convergence, the estimated pose of the current frame is used as the starting pose for the next frame. The shape model parameters b_i^q are only estimated once in the first frame. From our initial experiments, the shape parameters are not improved significantly when we include more frames and the fitting is made significantly more complex and time consuming. At each stage, when t_x, t_z, β, α and γ are estimated, only the region immediately surrounding the radius and ulna are used for cost function evaluation, while the larger region that includes the carpal bones is used for estimating the other parameters. There are about 60-80 frames per sequence. The whole process was performed in a 3-level multi-scale framework at each frame to enhance the robustness of the registration.

5 Evaluation

The ground truth of the recovered 3D pose corresponding to real fluoroscopic sequences is almost impossible to obtain. It would require the synchronisation of 3D imaging with the fluoroscopy. Hence, we evaluated our framework based on a number of simulated fluoroscopic sequences generated from the 3D CT data. All CT volumes have been resampled to an isocubic volume with voxel dimension of 0.5 mm. We linearly interpolated a number of poses between the neutral pose and two extreme poses of radial-ulnar deviation in a full movement cycle containing 50 poses for each of 10 subjects. The ray-casting method was then used to generate a simulated fluoroscopic sequence from those interpolated 3D data. We evaluated the proposed framework in the leave-one-out manner. The 3D pose of the simulated test subject was then calculated as described in section 4, and registration error measured by the 3D Euclidian distance of each corresponding point of the mesh between the target pose and the estimated pose is presented in Table 1. The error of the registration is mainly caused by the ill posed problem (confusion between the scale and translation along the out-of-plane direction), whereas the errors along the in-plane directions are very small with average error of about 2 pixels and maximum error within 4 pixels.

It is important to mention that the relative positions of the carpal bones with respect to each other can be estimated much more accurately than the absolute positions of the individual bones. The registration error of the 3D distance between the centroid of Triquetrum and the centroid of Lunate (dTL), and the distance between the centroid of Lunate and the centroid of Scaphoid (dLS) were also measured. The errors are 1.18 ± 0.74 and 1.82 ± 0.99 pixels for dTL and dLS respectively, compared to a bone size of about 30 pixels. One of the conditions that may be assessed using this method is dissociations, where the distance between the bones is larger than normal. Scapholunate dissociation is one of the most common of these. We normalise the dLS by dividing it by the estimated global scale factor s_{global} and an average of the scale factor s_i for lunate and

Table 1. The average error, measured in 3D, between the target and estimated correspondence points of each carpal bone of 10 subjects: Triquetrum(Tri), Lunate(Lun), Scaphoid(Sca), Pisiform(Pis), Hamate(Ham), Capitate (Cap), Trapezoid (Trd) Trapezium (Trm). The measurement errors of dTL and dLS.

	eTri	eLun	eSca	ePis	eHam	eCap	eTrd	eTrm	Total	eTL	eLS
Err3D	5.4±2.6	5.1±2.5	6.5±3.6	6.8±3.7	6.5±3.8	6.6±4.0	6.5±4.6	7.6±4.3	6.3 ±3.7	1.18±0.74	1.82±0.99
ErrX	1.6±1.3	2.0±1.6	2.1±1.8	2.4±1.9	1.8±1.4	2.1±1.5	1.8±1.4	2.2±1.8	2.1 ± 1.7	/	/
ErrY	3.7±2.8	3.0±2.6	4.9±3.7	4.8±3.9	5.4±4.2	5.3±4.4	5.5±5.0	6.1±4.6	4.6±4.0	/	/
ErrZ	2.5±2.0	2.5±1.9	2.2±1.8	2.6±2.1	1.6±1.3	1.7±1.3	1.5±1.2	2.2±2.0	2.3±1.9	/	/

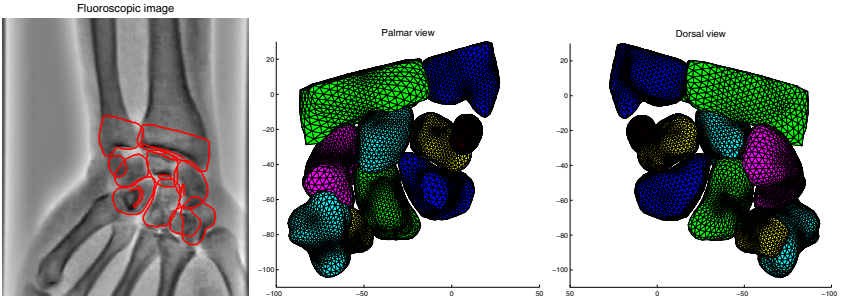


Fig. 3. Registration result of one frame from a real fluoroscopic sequence. The registration result for the whole sequence can be found in [11].

scaphoid. From the tested 10 subjects, we successfully identified the subjects suffering from scapholunate dissociation (dLS=38.78±1.53 pixels) from the normal subjects (dLS=34.49±0.83 pixels). Making this type of measurement without a 3D statistical model would be impossible.

We also tested our framework on real fluoroscopic sequences. Although the matching error cannot be quantified, the registration results show good visual correspondence and have been confirmed by a clinician. A sample frame of the matching result and the corresponding 3D pose are shown in Fig. 3 in which the projected contours from the 3D mesh model are superimposed on the preprocessed fluoroscopy image. The estimated 3D mesh model in the palmar and dorsal views are shown in middle and right respectively. The registration result for the whole sequence can be found in [11].

6 Concluding Remarks

We have presented a complete framework that is able to infer the 3D motion of carpal bones from a single view fluoroscopic sequence. It uses a hybrid statistical model to estimate both the kinematics and bone shapes from the fluoroscopic sequences allowing the motion of carpal bones during radial-ulnar deviation to be estimated. Particularly, the relative positions between carpal bones can be estimated accurately. This is potentially useful for detection of dissociation conditions, such as scapholunate dissociation,

where the underlying pathology is a rupture of one or more ligaments, and the diagnosis rests on a judgement regarding the joint separation.

In further work we will extend the current statistical model with more training data (in progress) and test the framework for the flexion-extension movement.

References

1. Snel, J.G., Venema, H.W., Moojen, T.M., Ritt, M., Grimbergen, C.A., den Heeten, G.J.: Quantitative in vivo analysis of the kinematics of carpal bones from three-dimensional CT images using a deformable surface model and a three-dimensional matching technique. Medical Physics 27, 2037–2047 (2000)
2. Sonenblum, S.E., Crisco, J.J., Kang, L., Akelman, E.: In vivo motion of the scaphotrapeziotrapezoidal (STT) joint. Journal of Biomechanics 37, 645–652 (2004)
3. van de Giessen, M., Streekstra, G.J., Strackee, S.D., Maas, M., Grimbergen, K.A., van Vliet, L.J., Vos, F.M.: Constrained Registration of the Wrist Joint. IEEE Transactions on Medical Imaging 28(12), 1861–1869 (2009)
4. Foumani, M., Strackee, S.D., Jonges, R., Blankevoort, L., Zwinderman, A.H., Carelsen, B., Streekstra, G.J.: In-vivo three-dimensional carpal bone kinematics during flexion-extension and radio-ulnar deviation of the wrist: Dynamic motion versus step-wise static wrist positions. Journal of Biomechanics 42, 2664–2671 (2009)
5. Penney, G.P., Batchelor, P.G., Hill, D.L.G., Hawkes, D.J., Weese, J.: Validation of a two- to three-dimensional registration algorithm for aligning preoperative CT images and intraoperative fluoroscopy images. Medical Physics 28, 1024–1032 (2001)
6. Zheng, G.: Statistically deformable 2D/3D registration for accurate determination of postoperative cup orientation from single standard X-ray radiograph. In: Yang, G., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009. LNCS, vol. 5761, pp. 820–827. Springer, Heidelberg (2009)
7. Chen, X., Graham, J., Hutchinson, C.E.: Integrated framework for simultaneous segmentation and registration of carpal bones. In: Accepted by the 18th ICIP, Belgium (2011)
8. Davies, R.H., Twining, C., Cootes, T.F., Taylor, C.J.: Building 3-D Statistical Shape Models by Direct Optimisation. IEEE Transactions on Medical Imaging 29(4), 961–980 (2010)
9. Amenta, N.: The Crust Algorithm for 3D Surface Reconstruction. In: Proceeding of the Fifteenth Annual Symposium on Computational Geometry (1999)
10. Black, M.J., Sapiro, G.: Edges as Outliers: Anisotropic Smoothing Using Local Image Statistics. In: Nielsen, M., Johansen, P., Fogh Olsen, O., Weickert, J. (eds.) Scale-Space 1999. LNCS, vol. 1682, pp. 259–270. Springer, Heidelberg (1999)
11. `http://personalpages.manchester.ac.uk/staff/xin.chen/CarpalReg.htm`

58. **Integrated framework for simultaneous segmentation and registration of carpal bones.** X Chen, J. Graham and C.E. Hutchinson, *Proceedings of IEEE International Conference on Image Processing, Brussels, September 2011 pp 441-444*. doi: 10.1109/ICIP.2011.6116543

INTEGRATED FRAMEWORK FOR SIMULTANEOUS SEGMENTATION AND REGISTRATION OF CARPAL BONES

Xin Chen, Jim Graham

Imaging Science and Biomedical Engineering
School of Cancer and Enabling Sciences
University of Manchester
Oxford Road, Manchester, M13 9PT, UK

Charles Hutchinson

Clinical Sciences Research Institute
Clinical Sciences Building
University Hospital -Walsgrave Campus
Clifford Bridge Road, Coventry, CV2 2DX, UK

ABSTRACT

A novel framework is presented in this paper for simultaneous multi-label segmentation and registration of carpal bones which leads to efficient statistical model building. It combines the Grow Cut segmentation algorithm with rigid image registration for propagating the segmentation of bones to new poses or different individuals. The proposed framework compares favourably to the conventional segmentation and non-rigid registration methods, in terms of flexibility and computational time, for our CT data of carpal bones. The segmentation code was implemented in a GPU, running about 15 times faster than CPU code.

Index Terms— Carpal bones, Grow Cut, Interactive segmentation, Rigid registration, CUDA programming

1. INTRODUCTION

The wrist is one of the most complex and vulnerable joints in the body, consisting of eight carpal bones. Wrist pain is currently diagnosed by expert assessment of abnormal carpal bone movements in 2D (projection) fluoroscopy sequences. The overall aim of the current project is computer interpretation of these sequences to attain a higher degree of objectivity and quantification in the diagnostic process. One important step towards this aim will be the development of statistical models (SD) of the carpal bones and their spatial relationships. Statistical models of shape and appearance [1, 2] have been widely used in the description and analysis of objects in medical and other forms of images. For this study we will train this statistical model from a large set of CT images.

Current well established methods of building SD require finding a set of correspondences on segmented surfaces across a training data set. The wrist presents particular segmentation challenges as the carpal bones are small and in close contact with each other, and the density and shape may vary between individuals. Van de Giessen et.al. [3] have reported using a geodesic active contour [4] for carpal bone segmentation, but this has not proved successful in our hands with our CT images. Interactive segmentation tools (e.g. Graph Cut [5] and Grow Cut [6]) provide a more promising route forward. However, our initial experience with our 3D CT data shows that a significant burden of interaction is required to obtain a satisfactory result. This burden is particularly daunting when a large 3D data set needs to be segmented for model building. Another class of methods to establish correspondence is to use non-rigid image registration techniques. These methods suffer from long computation time, are sensitive to initial starting pose and any inaccurate registration result can not be easily corrected.

To address these problems we propose an integrated framework which combines the Grow Cut segmentation method with rigid image registration to simultaneously segment and align the carpal bones in CT data captured from different subjects or in different poses from the same subject. The kinematics of the carpal bones is complex and significant pose difference can be introduced as the joint adopts different positions. The framework significantly reduces the workload of segmentation, simultaneously providing a good alignment of the carpal bones which is used for building a kinematic model. In the proposed framework, the segmentation of one data set was obtained interactively using the multi-class Grow Cut segmentation method [6] forming a template to segment and align with other images. In Grow Cut the user specifies a number of segmentation seeds which have initial strength values of 1, other points being initialised to 0. Then the labels are propagated interactively by comparing the transmitted strength of each voxel to that of its neighbours. The registration stage automatically initialises a strength map for the Grow Cut, which speeds up the segmentation significantly, whilst the segmentation result provides an updated pose for the registration, preserving the topology of the carpal bones.

2. METHODOLOGY

The proposed framework is designed to use the template volume V_S with its segmented volume V_{Sseg} to simultaneously register and segment the target volume V_T. The outputs of the framework contain a set of transformation parameters ($\vec{p}$: 3 translations, 3 rotations, 1 scale) $\vec{p_i} = [tx_i, ty_i, tz_i, \alpha_i, \beta_i, \gamma_i, s_i]^T$ ($i = 1, 2...n$, where n is the number of bones) as well as the segmented volume V_{Tseg} of V_T. The framework consists of the following steps:

1. *Rough Alignment*: Manually adjust the transformation parameter $\vec{p_i}$ for the whole or individual bone to roughly align with the target volume V_T generating a labeled volume V_{Trans}.

Thanks to Medical Research Council for funding.

2. *Strength Map Generation*: Using the labeled volume V_{Trans} (initialised from step 1 or updated from step 4), V_T can be classified as foreground (bone) and background (non-bone) from which the respective intensity histograms, P_{fore} and P_{back}, are calculated, leading to a strength map (Eqn. 1 to 5).

3. *Multi-class Grow Cut Segmentation*: Based on the labeled volume V_{Trans} from step 1 (or updated from step 4) and the strength map from step 2, run the Grow Cut algorithm to refine the segmentation volume V_{Tseg} and calculate the centroid point of each bone in V_{Tseg}.

4. *Rigid Image Registration*: Run the rigid image registration algorithm for each bone with its updated centroid to produce a new transformation parameter $\vec{p}_i$ and new labeled volume V_{Trans}.

5. *Iteration and Termination*: Steps 2 to 4 are repeated; the segmentation volume, registration parameters and the intensity histograms coherently improve each other until the termination conditions are satisfied. V_{Tseg} from step 3 and $\vec{p}_i$ from step 4 are the estimated segmentation and registration results respectively. Details of each stage are described in following subsections.

2.1. Rough Alignment

The segmented volume V_{Sseg} of the first data set V_S is obtained interactively using the Grow Cut segmentation method, resulting in different bones having different integer labels. The mesh model of each bone from V_{Sseg} is generated, together with a mesh model representing a rough shape of V_T constructed using a simple threshold. A straightforward GUI is used to change the transformation parameters of all bones simultaneously or individually, until the meshes of V_{Sseg} and V_T are roughly aligned. The sensitivity and robustness of the proposed algorithm to this manual alignment are evaluated in section 3. A new label volume V_{Trans} is then created from all the transformed bones. In V_{Trans}, all overlapped bone areas are set to zero, as new labels shouldn't be introduced.

2.2. Strength Map Generation

Here we present a novel method for initialising the strength map, which is used in Grow Cut in conjunction with the object labels to update the segmentation (Section 2.3). The objective here is to initialise this map, V_{Stren} with values of 1 (high certainty) and 0 (low certainty) of being either bone or non-bone. To obtain the V_{Stren}, an initial binary volume, $V_{bwTrans}$ (bone=1, non-bone=0) is generated from V_{Trans}. The normalised foreground and background histograms (P_{fore} and P_{back}) are calculated from the overlap of $V_{bwTrans}$ and the target volume V_T. Using Eqn. 1, we calculate the likelihood (V_L) of classifying each voxel as bone (positive) or non-bone (negative), from which Eqn. 2 and 3 generate new binary volumes (V_{bwL1}, V_{bwL2}) representing high certainty regions of bone and non-bone respectively. The thresholds of 0.9 and -0.5 were determined empirically. V_{bwL3} (Eqn. 4) represents the region of V_T that is not classified as bone either in $V_{bwTrans}$ or V_{bwL1}. Equation 5 identifies the regions that are identified with certainty to be bone or non-bone, based on the histograms (P_{fore} and P_{back}), constrained to be within the respective bone and non-bone regions defined by $V_{bwTrans}$. Following Grow Cut relabeling, $V_{bwTrans}$ and V_{Stren} are reinitialised for each iteration step.

$$V_L = \frac{(P_{fore}(V_T) - P_{back}(V_T))}{max(P_{fore}(V_T), P_{back}(V_T)))} \quad (1)$$

$$V_{bwL1} = \begin{cases} 1 & if V_L > 0.9 \\ 0 & otherwise \end{cases} \quad (2)$$

$$V_{bwL2} = \begin{cases} 1 & if V_L < -0.5 \\ 0 & otherwise \end{cases} \quad (3)$$

$$V_{bwL3} = 1 - (V_{bwTrans} \cup V_{bwL1}) \quad (4)$$

$$V_{Stren} = (V_{bwL2} \cap V_{bwL3}) \cup (V_{bwTrans} \cap V_{bwL1}) \quad (5)$$

2.3. Multi-class Grow Cut Segmentation

An advantage of Grow Cut is its ability to obtain a multi-label solution in simultaneous iteration, and it allows fast parallel implementation. Hence, it is a natural choice for our application in which a number of bones need to be labeled. For efficiency the Grow Cut code was parallelised using NVidia Quadro FX 3800 Graphic Card via CUDA API [7], which achieved a run time about 15 times faster than the CPU based code and 5 times faster than the Graph Cut algorithm running in the CPU.

In our proposed framework, the strength map V_{Stren} was initialised automatically in step 2, and V_{Trans} (from step1 or step 4) was used as the labeled volume. Since, there is only a small number of uncertain voxels with $V_{Strenh} = 0$ at each iteration, it takes less than 2 seconds to complete the segmentation of a $141 \times 268 \times 169$ volume.

2.4. Rigid Image Registration

The cost function for image registration is the summed normalised correlation coefficient calculated from the normalised image and gradient images in X, Y and Z directions.

The optimisation method used is a simplified but efficient version of Powell's method [8], and has been described in detail in [9]. The seven-dimensional search space (three translation, three rotation and one scale parameter) is explored by local 1D search, conducted by fitting a parabola to neighbouring points at multiple scales.

2.5. Iteration and Termination

The framework was implemented to run in a 3-level multi-scale manner. For a $141 \times 268 \times 169$ CT volume, the original data (level 3) was smoothed and downsampled by a factor of 2 and 4 for levels 2 and 1 respectively. In each level, step 2 to 4 are repeated, the segmentation volume, registration parameters and the intensity histograms coherently improving each other. The system terminates or moves to higher level, if the difference of the segmented volume V_{Tseg} between adjacent iterations stops decreasing. For the optimisation of rigid image registration in step 4, the initial searching range for all 3 levels are set to 1 pixel for translation, 1 degree for rotation and 0.05 for scale. The registration terminates when all the search ranges are smaller than 0.1. Finally in level 3, V_{Tseg}

from step 3 and $\vec{p}_i$ from step 4 are the final estimated segmentation and registration results respectively.

Following convergence, the framework also offers the flexibility to interactively refine the result by specifying additional seeds to achieve accurate detail in segmentation. For example Fig. 1 shows the result of the proposed method. Some segmentation details in Fig. 1(c) are corrected interactively using Grow Cut (making use of the final strength map from step 3) to produce the final result (Fig. 1(d)).

3. EVALUATION AND RESULTS

We applied the algorithm to CT data from 14 subjects. Five CT images were captured for each subject from different poses (neutral pose and two extreme poses in flexion and extension and medio-lateral movement). The CT slice thicknesses varied for different subjects, therefore the CT images were resized by trilinear interpolation to a volume with isotropic voxels of $0.5mm \times 0.5mm \times 0.5mm$. Figure 1 shows a slice from coronal view of the segmentation and registration results of one subject by using data from another subject as the template.

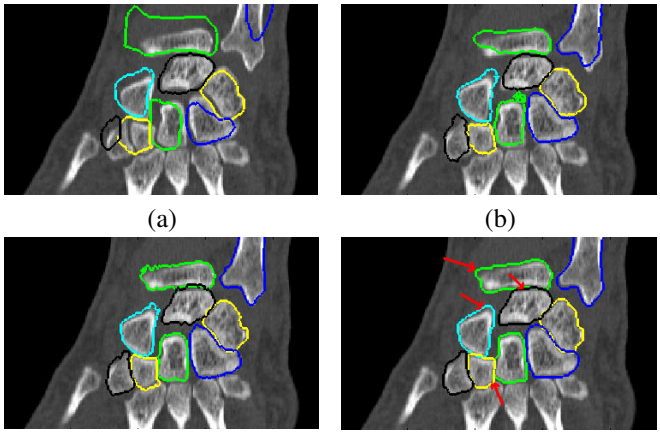


Fig. 1. A single slice from a volume registration, using the segmentation of the wrist of a different subject as a template. (a) Initial manual alignment. (b) Result from the registration step. (c) Result from the segmentation step. (d) Segmentation result after user interaction and smoothing. The arrows indicated locations where interactive correction have taken place.

One important advantage of the proposed framework is that the robustness to initial starting pose compares favorably to registration alone. We evaluated this by choosing one of the CT volumes as a template (with known 'ground truth' segmentation obtained by interactive Grow Cut, validated by a clinician). Copies of this segmented volume, translated, rotated and scaled to a number of starting poses, were used to register to and segment the original (untransformed) image.

The starting poses were grouped into 20 intervals based on the widely used mean Target Registration Error ($mTRE$) measurement [10], from 0-1 to 19-20 with 10 starting poses per interval. For each starting pose, the $mTRE$ was calculated by using randomly generate, uniformly distributed, transformation parameters $\vec{p}$, selecting those transformations for which the $mTRE$ is in the required interval. After registration, the $mTRE$ is again calculated to quantify the registration accuracy.

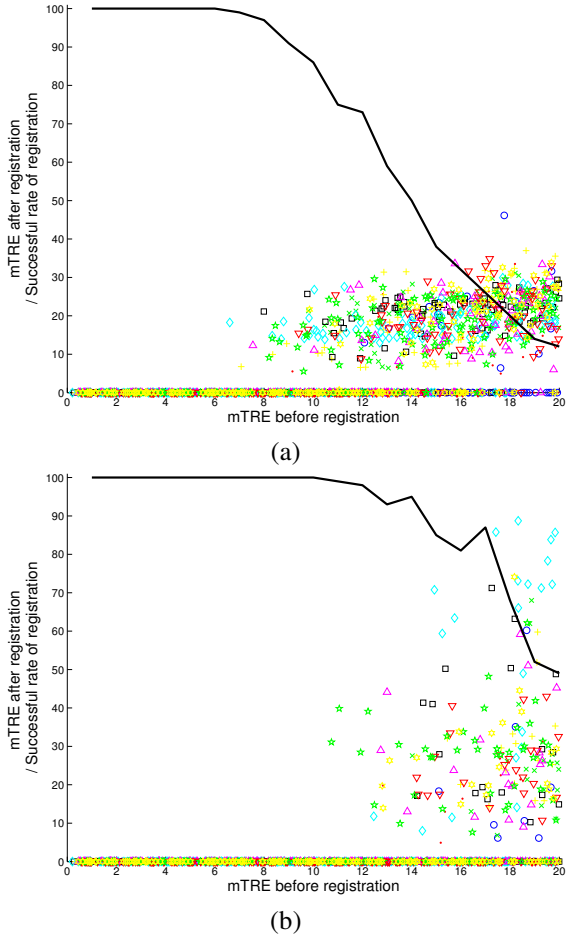


Fig. 2. (a) Registration results for registration only method (b) Registration results for the proposed method. The coloured symbols represent the registration errors for the individual carpal bones. The line shows the percentage successful registrations as a function of the initial pose displacement.

In our test, the registration accuracy of the heads of the radius and ulna and the 8 carpal bones were evaluated individually, each at its own rotation centre (centroid point). Hence, a total of 100 ($(8 + 2) \times 10$) registrations were calculated for each interval. We compared the proposed framework with propagation of segmentation using registration alone (using the method described in section 2.4) using the same set of starting poses. All parameters were set to be the same and the multi-resolution scheme was used for both methods. Figure 2 shows the registration results (measured by $mTRE$) against the initial $mTRE$ displacement. Each coloured symbol represents a single registration run on a particular bone. The graph represents the percentage of successful registrations as a function of the initial pose shift. The registration is defined as successful if the $mTRE$ value after registration is less than 1. Figure 2 shows that the proposed method results

in a smaller number of unsuccessful registrations. With registration alone, the success rate drops below 100% at an initial $mTRE$ of about 6, whereas the proposed method is more robust, retaining 100% success rate up to an initial $mTRE$ of about 10. The relatively high values of $mTRE$ in Fig. 2(b) are almost all due to the segmentation of a single bone (the trapezoid). Due to its small size and close articulation with neighouring bones its segmentation tends to be more sensitive to the initial pose, resulting in more failed cases and larger registration errors. We also compared the two methods by the following criteria (see table 1). Successful registration rate (SRR): the number of successful registration divided by the total tested number over all registrations (each individual bone counted as a registration); registration accuracy (RA): measured by $mTRE$ based on the final estimated transformation parameters for the successful registrations; capture range (CR): the maximum allowed initial displacement which can achieve 100% successful registration rate. Time: all experiments were tested on a 3.33 GHz PC with 24G memory, Grow Cut being implemented in the GPU, see section 2.3.

Table 1. Comparison of Registration Results

	SRR (%)	RA (pixel)	CR (pixel)	Time (s)
Reg. only	68.6	0.0087 ± 0.020	6	37±24
Proposed	90.4	0.0093 ± 0.007	10	223±135

Table 1 shows that the proposed method achieved higher successful registration rate, a larger capture range and more stable registration results, as the standard deviation is much smaller than the registration alone. The only drawback was that it required more computation time, where the generation of volume V_{Trans} of transformed bones in each iteration took the majority of the time. The unsuccessful registrations shown in Fig. 2 (mainly of the trapezoid) remain to be registered manually. This aspect was not part of the above evaluation. The mechanism for manual registration is to initiate the process with a closer initial alignment for the particular bones. The larger capture range for the majority of bones means that this interaction step needs to be carried out rather infrequently.

The segmentation results were further evaluated in the context of propagating the segmentation from one individual to images of others (as illustrated in Fig.1). We calculate an error rate between a set of 'ground truth' labeled volumes and the segmented volume by using the proposed method without any further refinement. The 'ground truth' data were obtained by applying the proposed method with further interactive refinement to 13 CT data from different subjects. All the segmentations were validated by a clinician. The segmentation error rate was calculated by counting the number of different labels between 'ground truth' and the final segmented volume without interaction, divided by the total 'ground-truth' bone volume. The segmentation error rate of our 13 tested images was $11.5 \pm 1.6\%$. Interactive segmentation using Grow Cut directly on the raw CT data normally took more than 20 minutes of interaction per volume (using the GPU code) to achieve similar error rate, and also requires reordering of the labels afterwards. Additionally, if different poses from the same subject are registered by the proposed method, almost no further interaction is required. We also tested rigid registration followed by local affine non-rigid registration without segmentation on our data. It suffers from long computation time and any incorrect registration result can not be easily corrected.

4. CONCLUSION AND DISCUSSION

In the context of segmentation for model-building, the ability to propagate a segmentation from one training example to others is particularly important. We have demonstrated a novel method for achieving this which is more robust to initial starting poses than a conventional registration method. The multi-class segmentation acts as a soft constraint which preserves the topology of the segmented volume, as it does not allow bone overlapping or jumps to random positions. Since different bones have the same intensity range, conventional segmentation methods tend to merge the bones that are in close contact. The registration stage acts to prevent the label floating to neighboring bone areas.

A hybrid statistical model, representing the motion and shape separately, can be built efficiently from the results of this procedure. The kinematic motion model is built based on the transformation parameters obtained from the rigid registration, and the segmentation result is used to build the shape model for each individual bone.

5. REFERENCES

[1] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models - their training and application," *Computer vision and Image Understanding*, vol. 61, pp. 38–59, 1995.

[2] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *European Conference on Computer Vision*, vol. 2, pp. 581–695, 1998.

[3] M. van de Giessen, G. J. Streekstra, S. D. strackee, M. Maas, K. A. Grimbergen, L. J. van Vliet, and F. M. Vos, "Constrained registration of the wrist joint," *IEEE Transactions on Medical Imaging*, vol. 28(12), pp. 1861–1869, December 2009.

[4] V. Caselles, R. Kimmel, and G. Saprio, "Geodesic active contours," *International Journal of Computer Vision*, vol. 22(1), pp. 61–79, 1997.

[5] Y. Boykov and M. P. Jolly, "Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images," *In Proc. of the International Conference on Computer Vision.*, vol. 1, pp. 105–112, 2001.

[6] V. Vezhnevets and V. Konouchine, "Grow-cut - interative multi-label n-d image segmentation," *Proc. Graphicon*, pp. 150–156, 2005.

[7] $www.nvidia.com/object/cuda\_home\_new.html$, ," .

[8] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, "Numerical recipes in c++," *U.K.: Cambridge University Press*, vol. 2nd ed., 1992.

[9] X. Chen, M. R. Varley, L. Shark, G. S. Shentall, and M. C. Kirby, "A computationally efficient method for automatic registration of orthogonal x-ray images with volumetric ct data," *Physics in Medicine and Biology*, vol. 53(4), pp. 967–983, Feb. 2008.

[10] E. B. van de Kraats, G. P. Penney, D. Tomazevic, T. van Walsum, and W. J. Niessen, "Standardized evaluation methodology for 2-d - 3-d registration," *IEEE Transaction on Medical Imaging*, vol. 24(9), pp. 1177–1189, August 2005.

59. **Automatic inference and measurement of 3D carpal bone kinematics from single view fluoroscopic sequences.** X Chen, J. Graham, C.E. Hutchinson and L. Muir, *IEEE Trans. Medical Imaging, **32**(2), 317-328, 2013.* doi: 10.1109/TMI.2012.2226740

Automatic Inference and Measurement of 3D Carpal Bone Kinematics From Single View Fluoroscopic Sequences

Xin Chen*, Jim Graham, *Member, IEEE*, Charles Hutchinson, and Lindsay Muir

***Abstract*—We present a novel framework for estimating the 3D poses and shapes of the carpal bones from single view fluoroscopic sequences. A hybrid statistical model representing both the pose and shape variation of the carpal bones is built, based on a number of 3D CT data sets obtained from different subjects at different poses. Given a fluoroscopic sequence, the wrist pose, carpal bone pose and bone shapes are estimated iteratively by matching the statistical model with the 2D images. A specially designed cost function enables smoothed parameter estimation across frames and constrains local bone pose with a penalty term. We have evaluated the proposed method on both simulated data and real fluoroscopic sequences and demonstrated that the relative poses of carpal bones can be accurately estimated. One condition that may be assessed using this measurement is dissociation, where the distance between the bones is larger than normal. Scaphoid–Lunate dissociation is one of the most common of these. The error of the measured 3D Scaphoid–Lunate distances were 0.75 ± 0.50 mm for simulated data (25 subjects) and 0.93 ± 0.47 mm for real data (15 subjects). We also propose a method for constructing a "standard" pathology measurement tool for automatically detecting Scaphoid–Lunate dissociation conditions, based on single-view fluoroscopic sequences. For the simulated data, it produced 100% sensitivity and specificity. For the real data, it achieved 83% sensitivity and 78% specificity.**

***Index Terms*—Carpal bone poses, fluoroscopic sequence, statistical pose model, statistical shape model, two-dimensional (2D) three-dimensional (3D) registration, wrist pathology.**

I. Introduction

WRIST pain, either acute or chronic, is a common presenting symptom in hand clinics. It may be due to a number of different pathologies, including acute trauma, arthritis (either osteo or inflammatory), vascular disorders, the sequelae of congenital abnormalities and the sequelae of trauma. These latter may include osteoarthritis secondary to fracture malunion or nonunion, and ligament instability. The standard assessment of a patient with pain of this nature will include history taking, clinical examination, and special investigations.

The wrist joint is complex, and the maintenance of the normal relationship of the carpal bones, both at rest and on movement is governed by intercarpal and extrinsic ligaments. Normal function and integrity of these ligaments is essential for the smooth movement of the wrist. No tendons insert onto the carpal bones themselves, and their movements are therefore dictated by the movements of the surrounding bones. Ligamentous injuries may lead to disordered movements of the bones. These disordered movements in turn lead to abnormal loading and hence to osteoarthritis. Standard assessment of these disordered movements includes plain radiography, MR scanning and cine radiography. The first two modalities give static images that may readily be examined and measurements taken, but are only static images of a dynamic problem. Cine radiography (e.g., fluoroscopic sequences) is more subjective and requires judgement and experience. If there is still doubt about the diagnosis, wrist arthroscopy may give further information, but this is an invasive procedure and therefore entails risk and expense. A method of determining carpal kinematics from fluoroscopic sequences that allowed more objective evaluation would be of value to the hand surgeon in accurate diagnosis. It would also contribute to treatment evaluation and to understanding an area of hand surgery that still remains challenging.

Here we present a method for computer interpretation of the fluoroscopic sequences to attain a higher degree of objectivity and quantification in the diagnostic process. The wrist is a complex joint (see Fig. 2); during wrist movement, the eight carpal bones follow a complex, multi-dimensional trajectory, making interpretation of radiographs difficult. One important step towards this aim is the development of statistical models (SM) of the carpal bones and their spatial relationships during movement, which is able to represent the pose and bone shape variation in much fewer dimensions. For this study we have trained this SM from a set of CT images from different subjects at different poses. Subsequently, the full 3D carpal bone motions can be recovered by matching the SM with the fluoroscopic sequences through 2D-3D image registration techniques.

A number of studies have sought to represent the carpal kinematics using CT or MR data, mainly concentrating on representing "average" kinematics over a small number of individuals (e.g., [1], [2]). Van de Giessen *et al.* [3] presented a 3D

Manuscript received August 29, 2012; revised October 19, 2012; accepted October 23, 2012. This work was supported by Medical Research Council, U.K., under Grant 87997. Date of publication October 26, 2012; date of current version January 30, 2013. *Asterisk indicates corresponding author.*

*X. Chen is with the Centre for Imaging Science, Faculty of Medicine, Manchester Academic Health Sciences Centre, The University of Manchester, M13 9PL Manchester, U.K. (e-mail: xin.chen@manchester.ac.uk).

J. Graham is with the Centre for Imaging Science, Faculty of Medicine, Manchester Academic Health Sciences Centre, The University of Manchester, M13 9PL Manchester, U.K.

C. Hutchinson is with the Division of Healthsciences, University of Warwick, CV4 7ES Coventry, U.K.

L. Muir is with the Department of Hand Surgery, Salford Royal NHS Foundation Trust, M6 8HD Salford, U.K.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2012.2226740

0278-0062/$31.00 © 2012 IEEE

rigid registration method based on segmented meshes, which aims to build a SM of carpal bones. More recently, they introduced a 4D statistical model that locally describes the relative positions of the carpal bones [4] in predefined poses, with the aim of detecting abnormal bone spaces. A comparison of wrist poses captured statically and dynamically was reported in [5]. They concluded that negligible differences were observed between the dynamic motion and the step-wise static motion of the carpal bones from "healthy" subjects. Some authors have focussed on building hierarchical statistical shape models ([6], [7]) or an articulated shape model [8]. Davatzikos *et al.* [6] presented a method of using the wavelet transform to capture different levels of shape detail in a coarse to fine structure, which enables the statistical shape model to cover a larger range of variability with a small number of training samples. Cerrolaza *et al.* [7] further extended the idea to deal with multiple objects for 2D brain image segmentation, where the objects to be included for model building at each level have to be carefully selected. Boisvert *et al.* [8] studied spine variation using 3D articulated pose models. The relative rigid transformation parameters of each vertebra with respect to the vertebra of the upper level were used to construct the articulated pose model. The spine variations between the same set of patients before and after treatment were compared using the model. Point-based statistical models, such as [7] do not retain the rigidity of each of the multiple objects. In our proposed framework we build a statistical pose model (SPM), based on geometrical transformation parameters and a separate point-based statistical shape model (SSM) to deal with the issue of shape variation and articulation of the carpal bones. We use the combined model to fit to image sequences for quantifying 3D movement.

Many studies have investigated 2D-3D image registration (e.g., [9]–[11]), mainly in the field of registration of pre-operative MR or CT images to intra-operative 2D images. Our work differs from these in that we seek to achieve registration of a 2D image sequence to a 3D model (not derived from the same individual) to infer the poses and shapes of an individual wrist. Zheng [12] took a similar approach to estimate the orientation of the pelvis from a single X-ray image. Whitmarsh *et al.* [13] presented a method to reconstruct both the 3D bone shape and 3D areal bone mineral density distribution of the proximal femur from a single dual-energy X-ray absorptiometry image. More recently, Baka *et al.* [14] and Zheng *et al.* [15] similarly presented a statistical shape model based framework to estimate femur shapes from multiple X-ray images. In the case of [14], fluoroscopic sequences were used, similarly to the work reported here.

The main contributions of this paper, distinguishing it from these earlier studies, are as follows. 1) A hybrid SM is developed representing both the complex pose and shape variation of the eight carpal bones plus radius and ulna. 2) The full 3D motion and bone shapes are recovered by matching the SM with a single view fluoroscopic sequence: a difficult ill-posed problem. 3) Our initial results show that the relative positions between the carpal bones can be estimated accurately through the proposed framework. 4) We have constructed a pathology detection tool that takes advantage of the inherent ability of the SPM to align wrist poses. In [4], they also detect abnormal bone spaces based on 3D input data sets for a limited number of predefined flexion–extension poses. We are not aware of any study which attempts to make a 2D to 3D inference and measurement in a system of this level of complexity. An early version of this work was published in [16]. In this paper, we describe the framework in greater detail and report the following further developments. 1) The SPM presented here is generated based on both the radial–ulnar poses and flexion–extension poses, where the SPM used in [16] is only based on radial–ulnar poses. 2) Faster optimization and more robust registration, arising from the use of a more constrained model. 3) Additional registration accuracy is achieved by the use of local pose refinement, controlled by a new cost function term. 4) Rather than building a SSM for each individual bone, all bones are modelled simultaneously to represent the shape variations of the ensemble of bones. This helps to maintain the nature of the relationships between adjacent bone shapes and reduces the number of shape parameters. 5) We include more comprehensive experimental results based on real fluoroscopic sequences using extended training datasets. 6) A method of constructing the pathology detection tool, based on the SPM, is introduced for the first time. The evaluation results demonstrate the feasibility of using the proposed system for clinical diagnosis.

The overview of the proposed framework is illustrated in Fig. 1(a). The system consists of a training phase and a 2D-3D image registration phase. We currently have CT data from 25 subjects, each at five poses (neutral pose and two extreme poses in flexion–extension and radial–ulnar deviation). The segmentation of each bone and rigid registration parameters that align bones at different poses within and across the subjects in the training set were obtained using an iterative segmentation and registration algorithm [17]. Segmentation results were confirmed by an experienced radiologist. A hybrid statistical model, representing both the pose and shape variations, was built from the results of the segmentation-registration framework. The SPM was built based on the transformation parameters, while the segmentation result was used to build the SSM. In the 2D-3D image registration phase, the global 3D rigid transformation, the poses of carpal bones, the local 3D rigid transformation of each bone and the bone shapes were estimated iteratively in sequence from each frame of the fluoroscopic video. The registration is performed sequentially, frame-by-frame, the estimated poses at each frame acting as the starting positions for the next [see Fig. 1(a)]. Detailed descriptions are given in Sections II–VI.

II. Coordinate System and Problem Parameterisation

In Fig. 1(b), the coordinate axes X^S, Y^S, and Z^S define the source coordinate system with the origin at the radiation source, whereas X^M, Y^M, and Z^M define the machine coordinate system with the origin at the isocenter. u and v define the image coordinates, normal to the direction of the radiation beam. The origin of the image plane is at the projection of the optical center.

In order to interpret the true 3D motion of the carpal bones, four sets of parameters are estimated iteratively in sequence during image registration. 1) Rigid transformation parameters of the wrist and a global scale factor, denoted by

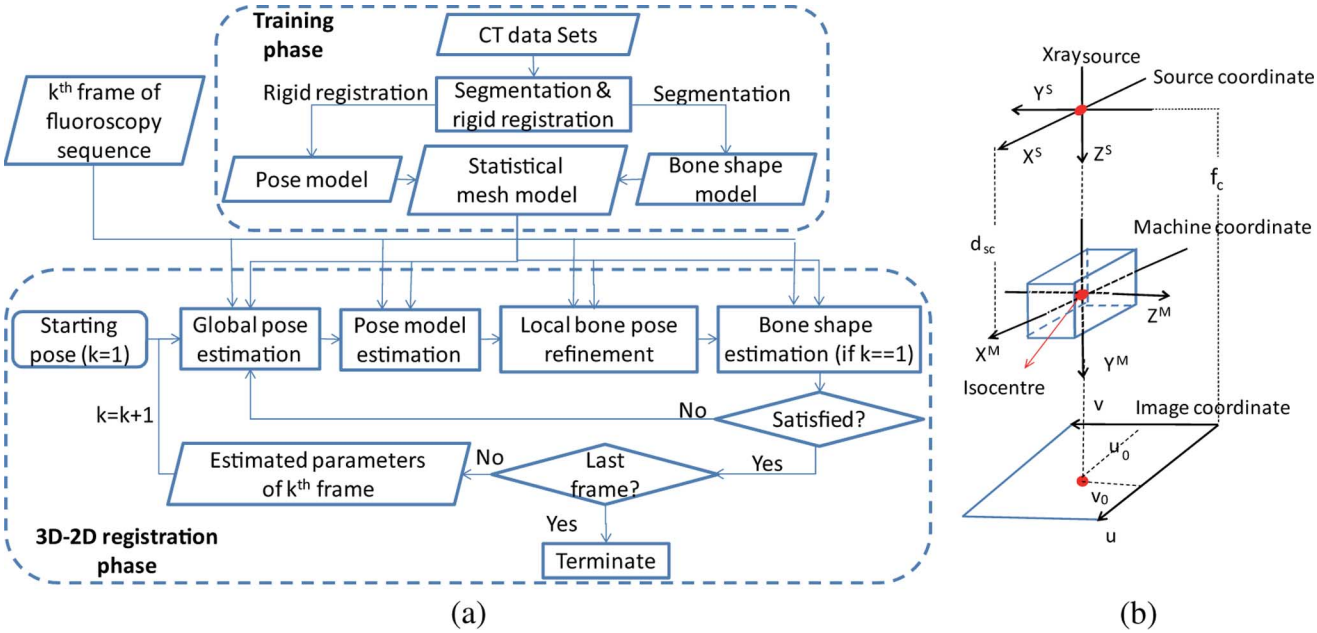


Fig. 1. (a) Overview of the proposed system. (b) Perspective projection geometry for the fluoroscopic imaging system.

$\theta = \{tx, ty, tz, r1, r2, r3, s\}$ in the machine coordinates. $t = [tx, ty, tz]^T$ denotes the translations along X^M, Y^M and Z^M axes. $r = [r1, r2, r3]^T$ is the set of Rodrigues parameters [18] representing the global orientations. The magnitude of vector r is the rotation angle around the axis represented by the normalized unit vector of r. s controls the distance between the centroid of each bone and the origin in the radius, and the global size of the bones. 2) SPM parameters b^m represent the carpal bone poses during movement. By using the pose model parameters, the transformation parameters of each bone can be obtained, denoted as $m_i = (tx_i^m, ty_i^m, tz_i^m, r1_i^m, r2_i^m, r3_i^m)$ (i is an index identifying each bone). 3) Transformation parameters $l_i = (tx_i^l, ty_i^l, tz_i^l, r1_i^l, r2_i^l, r3_i^l, s_i^l)$ of each bone used to refine the poses estimated from the pose model. 4) SSM parameters b^q for bone shape estimation. Using homogenous coordinates, the constructed 3D statistical mesh model can be projected to the image plane by

$$A_i = KTPD_i \begin{bmatrix} ss_i^l Q_i \\ 1 \end{bmatrix} \tag{1}$$

where Q_i indicates the mesh points of the estimated shape for the ith bone. s and s_i^l are the global and local scale factors respectively that control the size of the carpal bones. D_i is the pose matrix of the ith bone estimated using the pose model and the local pose refinement. P is the global rigid transformation matrix. T is the transformation matrix from the machine coordinate system to the source coordinate system, and K is the intrinsic projection matrix of the X-ray imaging system. In detail, P is denoted as

$$P = \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix} \tag{2}$$

where t is the translation vector $[tx, ty, tz]^T$. R is the 3×3 rotation matrix represented by Rodrigues parameters [4], [18], calculated as

$$R = I + B \sin|r| + B^2(1 - \cos|r|) \tag{3}$$

where $|r|$ is the magnitude of the orientation vector $[r1, r2, r3]^T$. I is the identity matrix, and B is the skew-symmetric matrix normalized by $|r|$, expressed as

$$B = \frac{\begin{bmatrix} 0 & r3 & -r2 \\ -r3 & 0 & r1 \\ r2 & -r1 & 0 \end{bmatrix}}{|r|}. \tag{4}$$

In (1), D_i is calculated as

$$D_i = \begin{bmatrix} R_i^g & t_i^g \\ 0 & 1 \end{bmatrix} \tag{5}$$

where $t_i^g = s[tx_i^m, ty_i^m, tz_i^m]^T + [tx_i^l, ty_i^l, tz_i^l]^T$ is the summation of translation vectors estimated from the pose model and local bone refinement. $R_i^g = R_i^m R_i^l$ is the 3×3 rotation matrix that combines the rotations estimated from pose model and local bone refinement, respectively. R_i^m and R_i^l can be calculated individually by (3) using their corresponding Rodrigues parameters.

Furthermore, T in (1) is given by

$$T = \begin{bmatrix} C_{az} & S_{az} & 0 & 0 \\ -S_{el}S_{az} & S_{el}C_{az} & C_{el} & 0 \\ C_{el}S_{az} & -C_{el}C_{az} & S_{el} & d_{SC} \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{6}$$

where C and S denote cosine and sine functions, subscripts az and el denote the view angles for 3D-2D projection, with $az = 0°$ and $el = 180°$ producing the anteroposterior view, $az = 90°$ and $el = 180°$ producing the left lateral view, and $az = -90°$ and $el = 180°$ producing the right lateral view. d_{SC} indicates the distance between the isocenter and the X-ray source.

K in (1) is given by

$$K = \begin{bmatrix} \frac{f_c}{pix_u} & 0 & u_0 & 0 \\ 0 & \frac{f_c}{pix_v} & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{7}$$

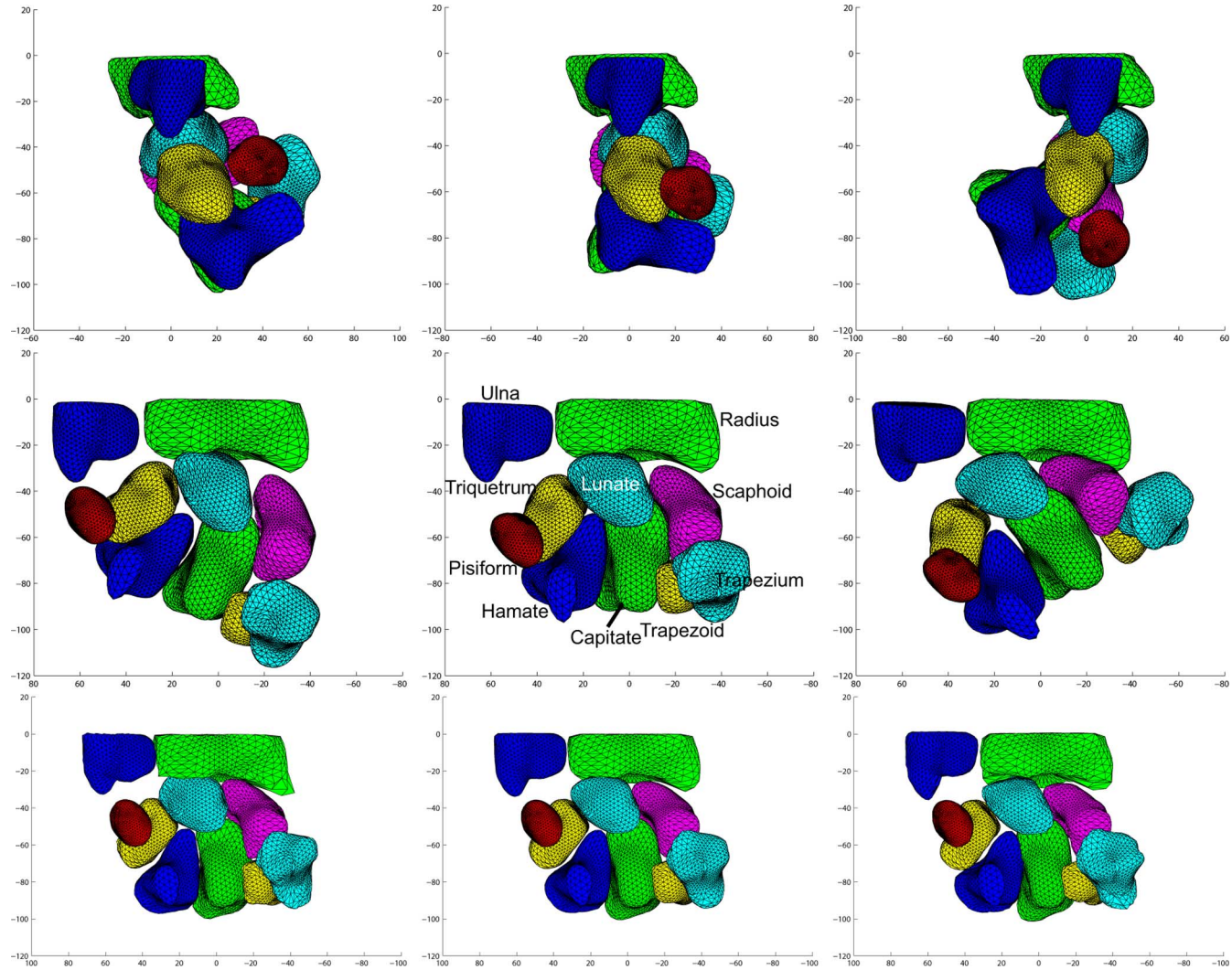


Fig. 2. Top row: The poses of the first component of the pose model (lateral view) that mainly describes the flexion–extension movement. Middle row: The poses of the second component of the pose model (AP view) that mainly represents the radial–ulnar movement. Bottom row: the first component of the shape model. (Major shape variations occur in the ulna, radius, and lunate.) In each case the mean ± 2 s.d. are shown.

where f_c is the distance between the X-ray source and detector plane. pix_u and pix_v are the physical pixel sizes along the horizontal and vertical directions of the detector, and (u_0, v_0) are the coordinates of the optical center on the image. In our data, (u_0, v_0) were always the center of the image. pix_u and pix_v are known from the detector specification. Therefore, only f_c needs to be estimated, which can be done by measuring the distance between the X-ray source and detector for each subject.

The use of Rodrigues parameters to represent bone orientations is convenient for pose model building and parameter optimization. More importantly, unlike the quaternion representation, it does not require vector normalization, nor does it suffer from the singularity problem that arises when using the Euler angle rotations.

III. Training of Pose Model and Shape Model

To generate training data it was necessary to achieve consistent segmentations and poses of the bones across subjects in the training set and across the five wrist positions within each subject. For this we developed an integrated framework [17] that combines the Grow Cut segmentation method with rigid image registration to simultaneously segment and align the carpal bones in the CT data sets. The kinematics of the carpal bones is complex and significant pose differences can be introduced as the joint adopts different positions. The framework significantly reduces the workload of segmentation, while simultaneously providing a good alignment of the carpal bones. Each bone segmentation was verified by an experienced radiologist.

As the shape of each bone may vary from individual to individual, we modelled this variation using a point distribution model (PDM) [19]. This was built using the segmented volume of the same set of training subjects. Correspondence between these surfaces of bones across subjects was established by the minimum description length (MDL) algorithm [19]. The 3D structure of each bone is described by a set of 1002 points on the segmented surface. In our earlier work [16] we modelled the shape of each bone independently. However here we maintain the nature of the relationships between adjacent bone shapes and reduce the number of shape parameters by representing the shape points of all bones in a single column

TABLE I
AVERAGE ERROR IN MM, MEASURED IN 3D, X^M, Y^M, AND Z^M AXES, BETWEEN THE TARGET AND ESTIMATED CORRESPONDENCE POINTS OF EACH CARPAL BONE OF 25 SUBJECTS: TRIQUETRUM(TRI), LUNATE(LUN), SCAPHOID(SCA), PISIFORM(PIS), HAMATE(HAM), CAPITATE (CAP), TRAPEZOID (TRD) TRAPEZIUM (TRM). THE "NO LOCAL SCALE" COLUMN LISTS THE REGISTRATION ERRORS WITHOUT ESTIMATION OF LOCAL SCALE OF EACH BONE (SEE TEXT)

	eTri	eLun	eSca	ePis	eHam	eCap	eTrd	eTrm	Average	No local scale
Err3D	2.51±1.42	2.23±1.24	2.18±1.50	2.63±1.72	2.31±1.44	2.34±1.50	2.51±1.72	2.86±1.85	2.45±1.07	2.86±1.08
ErrX	0.59±0.44	0.58±0.50	0.44±0.37	0.58±0.53	0.44±0.33	0.45±0.35	0.53±0.44	0.47±0.43	0.51±0.39	0.60±0.41
ErrY	2.15±1.56	1.38±1.35	1.43±1.60	2.24±1.84	2.10±1.55	2.11±1.61	2.23±1.82	2.64±1.94	2.16±1.14	2.53±1.20
ErrZ	0.68±0.55	0.59±0.49	0.48±0.39	0.68±0.61	0.44±0.35	0.45±0.35	0.55±0.46	0.57±0.52	0.55±0.43	0.64±0.46

vector in a consistent order. One training example is described by $(x_1, y_1, z_1, \ldots x_{10020}, y_{10020}, z_{10020})^t$ (10 bones $\times$ 1002 points each). The coordinates of the shape points of each bone are expressed with respect to its own centroid, eliminating any linkage between the shape model and the pose model. The deformable shape model is then described as

$$q = \mu^q + \phi^q b^q \tag{8}$$

where μ^q and ϕ^q (superscript q is a notation indicating the shape parameters) are the mean shape and the principal subspace matrix for the shapes. b^q is the shape model parameter to be estimated. We retain the first 15 significant components in the shape model, which keeps about 90% of variation.

The statistical pose model was trained using the six rigid transformation parameters. The common coordinate system for all pose parameters has an origin at a specified point in the radius for a reference subject. The sizes of all the wrists are normalized to a consistent scale. The pose of one subject is described by $(tx_1, ty_1, tz_1, r1_1, r2_1, r3_1, \ldots, tx_{10}, ty_{10}, tz_{10}, r1_{10}, r2_{10}, r3_{10})^t$ (eight carpal bones, one radius, and one ulna). The orientation parameters allow for a continuous description of the wrist movement (see Section II). Then the pose model can be parameterized as

$$m = \mu^m + \phi^m b^m \tag{9}$$

where the mean pose μ^m (superscript m is a notation indicating the pose model parameters) and the principal subspace matrix ϕ^m are computed from 5 (poses) $\times$ 25 (subjects) training samples using PCA. The vector b^m represents the pose parameters that describe the pose (m) along each principal direction. In our experiments, only the first two significant components are used, which keeps 90% of variation. The first component reflects the flexion–extension motion and the second component represents the radial–ulnar motion. By contrast, our earlier work [16] used eight significant modes representing 98% of the variation based on 10 training subjects for the radial–ulnar movement only. Experimentally we found that the use of fewer model components reduced computational time by 40%. The inclusion of flexion–extension poses for training also extends the motion range which helps to reduce the registration errors. This probably arises because, in capturing the training data, there was no constraint on the radial–ulnar movement in CT, so that the correspondence between the extreme positions of radial–ulnar movement in CT and fluoroscopy may not be exact. There are also potentially small differences in the directions of flexion–extension and radial–ulnar movement between the fluoroscopy and CT image capture processes. By further combining with the local bone refinement procedure, the more constrained model achieved smaller registration error in 3D by around 0.7 mm (values shown in Table I), compared with the results in [16]. Our experiments also showed that including significant components beyond two does not improve the registration accuracy, which indicates that the local bone refinement process (Section IV) dealt with the deviation from the linear pose model very well.

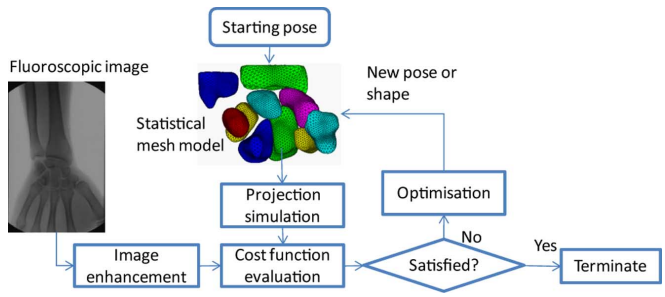


Fig. 3. Overview of the 3D-2D image registration process.

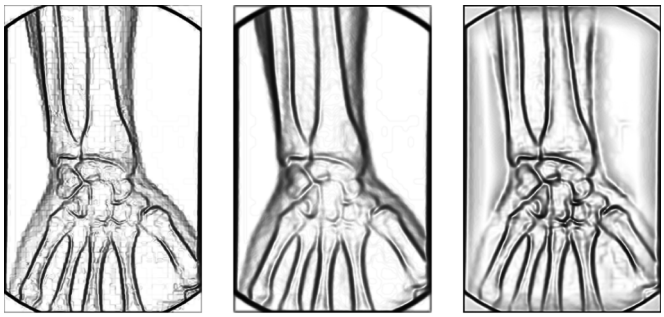
Based on the SSM and the SPM, a hybrid statistical mesh model can be built by using the Crust mesh construction algorithm [20]. Fig. 2 shows the poses of the first two components of the SPM (represented by the mean shapes of each bone) and the first mode of the shape variation.

IV. 3D-2D IMAGE REGISTRATION

The statistical mesh model is then used to match with each of the frames in the fluoroscopic sequence to infer the 3D motion and bone shapes [see Fig. 1(a)]. The position of the model is firstly initialised interactively by indicating a central point on the radius in the first frame of the fluoroscopic sequence. Then the poses of the bones in each frame are estimated in sequence, the poses from the current frame being used as the starting poses of the next. Fig. 3 summarizes the registration process, in which the preprocessed fluoroscopic image is iteratively matched with a simulated projection generated from an updated pose of the mesh model. This registration procedure is used specifically in the pose estimation and refinement steps illustrated in Fig. 1(a). For each iteration, the global pose parameter $\theta = \{tx, ty, tz, r1, r2, r3, s\}$, the SPM parameter b^m, the local transformation parameters $l_i = (tx_i^l, ty_i^l, tz_i^l, r1_i^l, r2_i^l, r3_i^l, s_i^l)$ of each bone, and the SSM parameters b^q are updated iteratively in sequence. Detailed descriptions are given in Sections IV-A–IV-C.

A. Fluoroscopic Image Enhancement and Projection Simulation

As there is considerable variation in the quality of fluoroscopic images, preprocessing is necessary to achieve consistent



(a) (b) (c)

Fig. 4. Left: Gradient map of the original flouroscopic image. Middle: Gradient map of the image after applying the diffusion filter. Right: Gradient map of the image after local normalization and diffusion.

results. Firstly the intensities are normalized to zero mean and unit standard deviation. This is followed by anisotropic diffusion [21] to smooth the image and preserve edges. Local gradients are used for image matching, and Fig. 4 shows the gradient maps of an original fluoroscopic image, and the image after anisotropic diffusion and after normalization and anisotropic diffusion respectively.

In order to optimize the pose parameters, we iteratively generate projections from the statistical mesh model with updated pose parameters. The mesh model is considered as a binary volume with background set to zero and bone set to unity. Based on the perspective projection model described in Section II, the simulated projection can be generated by ray casting. The projected intensity is in negative proportion to the sum of binary values along the ray from the source to each pixel in the image plane. The simulated image that represents the mean shape and mean pose of the model is shown in Fig. 5.

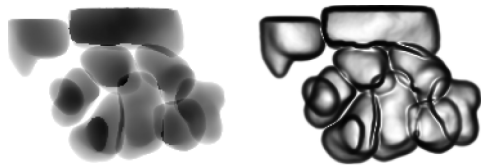
B. Cost Function

To evaluate the similarity between the fluoroscopic image and the simulated image, we investigated several forms of the cost function, achieving best results from the one shown in (10), based on the gradient along horizontal and vertical directions as well as the gradient magnitude of the two images. Additionally, the adjacent frames to the current fluoroscopic image were also taken into account in the cost function to make the estimated poses smooth across frames.

If we define the normalized correlation coefficient (NCC) between two images A and B as C(A,B), then the proposed cost function can be described as

$$\begin{aligned} E_1 = &-C(Om_{k-1}, Om_k) \\ &- \sum_{p=k-1,k,k+1} w_p(C(Im_p, Dm_k) \\ &+ C(Ix_p, Dx_k) + C(Iy_p, Dy_k)) \end{aligned} \tag{10}$$

where k is the current frame number of the fluoroscopic sequence. Im_p, Ix_p and Iy_p are the gradient magnitude, vertical gradient and horizontal gradient of the fluoroscopic image at the pth frame respectively. Dm_k, Dx_k, and Dy_k are the corresponding values of the simulated image. The use of the absolute gradient magnitude in the second term, in addition to the signed gradient, results in a smoother objective function, resulting in a reduced tendency to converge to local minima than is the case when using signed gradients alone. Calculating the cross-correlation between sets of three adjacent frames makes the estimated pose smooth across frames. The inter-frame weighting parameters, w_{k-1}, w_k, and w_{k+1} were set at 0.2, 0.6, and 0.2, respectively. For the first term of the cost function, the vertices in the statistical mesh model are projected to the image plane; we assume the intensities at those projected points are similar across adjacent frames. Om_{k-1} and Om_k represent the gradient magnitude of the previous frame and the current frame at the projected correspondence positions. The first term makes the shape of the cost function sharper, leading to a faster and more accurate optimization result. The $(k-1)$th frame and $(k+1)$th frame are not evaluated for the first and last frame, respectively.



(a) (b)

Fig. 5. Left: Simulated image that represents the mean shape and mean pose of the model. Right: Magnitude of gradient.

Equation (10) is used to estimate the global pose parameter θ and the SPM parameter b^m. The wrist motion can be described as approximately linear by the SPM parameters, where the deviations from linear positioning are accommodated by the local refinement of individual bone poses. In the local refinement procedure, a different cost function is used, where an additional term is added to E_1 as described in (11). The additional term makes the estimated local pose as close as possible to the pose model, weighted by a Gaussian distribution. This is able to preserve the topology of the carpal bones, when the intensity term E_1 is weak

$$E = E_1 + \omega \exp\left(-\frac{\frac{1}{p}\sum_{i=1}^{p} \|x_i^g - T^l(x_i^g)\|^2}{2\phi^2}\right). \tag{11}$$

In (11), x_i^g represents the ith 3D mesh points after the global pose and pose model estimation. p is the total number of mesh points of the currently evaluated bone (In our case, $p = 1002$ for each bone). T^l is the local transformation matrix for that bone. ω is the weighting parameter that balances the image intensity term E_1 and the added geometric penalty term. ϕ is the standard deviation of the Gaussian distribution. In our evaluation tests, $\omega = -0.2$ and $\phi = 10$ were experimentally determined and used.

C. Optimization

The coordinate origin for all motions is the centroid of the radius in the statistical mesh model. The global transformation parameters are estimated based on the regions surrounding all bones and iteratively refined by alternating with the SPM parameter and local transformation parameter estimations [see Fig. 1(a)]. By estimating the SPM parameter b^m based on all carpal bone and ulna regions, a set of transformation parameters $m_i = \{tx_i^m, ty_i^m, tz_i^m, r1_i^m, r2_i^m, r3_i^m\}$ representing the

kinematic pose of the ith bone can be generated by (9). The local transformation parameters l_i for each bone are calculated by evaluating the cost functions on the corresponding bone volumes. The set of mesh points Q_i that represent the ith bone shape are obtained by substituting the estimated SSM parameter b^q into (8). Subsequently, the 2D projection that represents the current estimated 3D pose of the carpal bones can be generated using (1).

The optimization method we have used is a simplified version of the Brent–Powell method [22], requiring a smaller number of optimization steps. We used parabola fitting to replace the Brent line search in the Brent–Powell method. The multi-dimensional search space (θ, b^m, l_i, and b^q) is explored by iterative individual 1D line searches. For each parameter search, the cost function is evaluated three times at the current position and its negative and positive neighbors, respectively, with the initial distance between the current position and its neighbors predefined by a search range. To fit a parabola to these three values, the following three criteria are applied to select the best parameter value for the next iteration.

1) A minimum is found by equating the first derivative of the fitted parabola to zero, with the second derivative being positive: In this case, the minimum is selected as the current best parameter value for the next iteration or for evaluation of the next parameter.
2) A maximum is found with the second derivative being negative: In this case, the parameter value corresponding to the smallest cost function value, evaluated at the current position and its neighbors, is selected.
3) A minimum is found, but it is too far away from the evaluated position (located outside twice the initial search range due to the cost function being too flat): In this case, the transformation parameter value corresponding to the smallest cost function value, evaluated at the current position and its neighbors, is again selected.

When a better value is found for one parameter, it will be used for evaluating the next one. When all the transformation parameters satisfy the first criterion, the search range is reduced by dividing it by a factor to refine the estimation results. The whole optimization is terminated when the changes in the evaluated cost function values are smaller than a preset threshold or the search range is small enough.

In our case, the true sizes of the bones are unknown; recovering the 3D pose from a single image is therefore a difficult, ill posed, problem. Any movement along the Y^M axis in the machine coordinates, could be compensated by scaling of the bone. In order to minimize this effect, the optimization is carefully sequenced. We firstly assume that the wrist is not moving along Y^M axis during radial–ulnar movement ($ty = 0$), as it is placed on a flat surface. Following the interactive initialization (Section IV), the first frame is evaluated, taking all the parameters into account (except ty) in the following sequence: tx, tz, $r1$, $r2$, $r3$, s, b^m, tx_i^l, ty_i^l, tz_i^l, $r1_i^l$, $r2_i^l$, $r3_i^l$, s_i^l, and b^q. After convergence, the estimated pose of the current frame is used as the starting pose for the next frame. The global scale factor s, local scale factors s_i^l and shape model parameters b^q are only estimated once in the first frame. From our initial experiments, the shape parameters are not improved significantly when we include more frames and the fitting is made significantly more complex and time consuming.

V. Evaluation

The true 3D poses corresponding to the recovered poses for real fluoroscopic sequences are not available: there is no ground-truth against which to judge the accuracy of the recovered poses. This would require the synchronization of 3D imaging with the fluoroscopic imaging devices. The proposed framework was therefore evaluated based on 25 simulated sequences in addition to 15 real fluoroscopic sequences. All evaluations were conducted using a leave-one-out strategy, based on the training data. In all of the evaluation tests, the input fluoroscopic sequences were preprocessed to construct a three-level multi-scale pyramid (down-sampled by a factor of 2 at each level). In the optimization procedure, the same set of fixed initial search ranges was used at each level (four voxels for translation, $4\pi/180$ for rotation, 0.2 for scale, one standard deviation for pose model parameters and shape parameters). The search ranges were divided by 2 each time the criteria were met (see Section IV-C), and the whole process was terminated when the maximum value of the search ranges was smaller than a preset threshold. The registration accuracy of the simulated data and real data are shown in Sections V-A–V-C. More importantly, a measurement model that represents the healthy pose of carpal bones at each kinematic pose is generated. This model can be used for pathology detection and quantification.

A. Evaluation Based on Simulated Data

We evaluated our framework quantitatively based on a number of simulated fluoroscopic sequences generated from the 3D CT data. All CT volumes have been resampled to an isocubic volume with voxel dimension of 0.5 mm. We interpolated (cubic spline) a number of poses between the neutral pose and two extreme poses of radial–ulnar deviation in a particular movement cycle (neutral–full radial–neutral–full ulnar), resulting in 39 poses for each of 25 subjects. While we assume a linear model for variation in pose, the cubic spline interpolation makes the trajectory smooth around the observed poses. The ray-casting method was then used to generate a simulated fluoroscopic sequence from those interpolated 3D data. The tested dataset was not included in the training datasets.

We conducted initial leave-one-out experiments to evaluate the number of PCA components required for the SSM. In these we altered the number of shape model components, leaving all other parameters unchanged. The final 2D-3D registration accuracy stopped improving when 15 components were selected. This may be due to the shape errors estimated using components greater than 15 being less significant than the pose errors. This suggests that using 25 subjects and 15 significant components are sufficient for this application.

To test the registration accuracy of the whole framework, the 3D pose of the simulated test subject was then calculated as described in Section IV. The registration error measured by the 3D Euclidean distance at each corresponding point of the mesh between the target pose and the estimated pose is presented in

Table I. The average 3D registration error is 2.45 ± 1.07 mm. The main contribution to this error is the ill-posed problem (confusion between the scale and translation along Y^M), whereas the errors along the in-plane directions, X^M and Z^M, are very small with average error about 1 voxel (0.5 mm).

As described in previous sections, the local scale factor of each bone is also estimated in the local refinement procedure to take account of the fact that the relative sizes of bones will vary between individuals. Based on the 25 independent tests, the mean value of this local scale factor varies between 0.98 and 1.13, depending on which bone is being considered. The standard deviations are around 0.05, indicating that the relative sizes of bones varies between individuals. While this complicates the optimization, the last column of Table I shows that the optimization without the local scale results in a larger registration error.

In clinical diagnosis, the absolute positions of the carpal bones in 3D space are not important; of greater significance is the relative movement of the bones. By using our method, the relative positions of the carpal bones with respect to each other can be estimated much more accurately than the absolute positions of the individual bones. In calculating the distance between bones we use the average distance between corresponding surface points. Each bone is represented by the same number of surface points (1002), determined when the shape model was constructed using the MDL method (Section III). Correspondences are determined using the index of each point, giving a consistent set of correspondences.

One condition that may be assessed using this measurement is dissociation, where the 3D distance between the bones is larger than normal. As an example of this, we investigate Scapholunate dissociation, which is one of the most common of these conditions. The registration error of the 3D distance between the Lunate and the Scaphoid (dLS) was measured. The error is 0.75 ± 0.50 mm, compared to an average surface to surface distance of 2 mm between the Scaphoid and Lunate. The surface to surface distance is measured by the average of the 20 shortest Euclidean distances between the surface points of the two bones. More importantly, using the statistical model, the measured 3D bone distances can be normalized to a consistent scale by dividing them by the estimated global scale factor s and an average of the two bones' local scale factors, calculated as $(s_i^l + s_j^l)/2$. This leads to automatic classification of the bone dissociation cases, which could not possibly achieved without the statistical model (Section V-C).

B. Evaluation Based on Real Data

We also tested our framework on 15 real fluoroscopic sequences. There were about 40–100 frames per sequence, covering the radial–ulnar movement. In the absence of ground truth, the absolute positions of bones cannot be used for evaluation. However, the key question is whether the estimated relative distances between bones are equivalent to the measurements from CT data, and the diagnostic conclusions unchanged. The registration accuracy of the real data can be validated by comparing the 3D distance between Scaphoid and Lunate (dSL) estimated from real fluoroscopic sequences and the original 3D volumes of the same subject.

One major advantage of using the SPM as one of the registration steps is that the kinematic pose of the wrist from different motion sequences can be aligned directly based on the SPM parameters. This provides an advantage compared with the method described in [4] where it is required to align the wrist to predefined discrete poses. The first two components of our SPM cover 90% of variation in the full range of flexion–extension and radial–ulnar movements. The combination of the two components is also able to generate interpolated poses within the motion range. To measure the error in the estimated 3D pose at each wrist position, we need to compare it with the pose of the 3D CT data at that position. To do this, we need to index the positions along the motion trajectory, which can be done using the first two components of the SPM. The values of these components define corresponding positions for the model and the CT data.

To produce ground truth corresponding to the original 3D data requires 3D-3D registration between each bone in the 3D statistical mesh model and the corresponding bone in the original 3D data set. This was done at a number of poses by estimating the global pose parameter, SPM parameter, and local pose parameter (θ, b^m, l_i-Section IV) at each pose. The CT volume was then set to the same pose location according to the first two components of the estimated SPM parameter. Poses of the original 3D data, other than the neutral and extreme poses were generated by cubic spline interpolation. Having matched the poses of the estimated and real 3D bone positions, the 3D distances between the Scaphoid and Lunate in the original and estimated volumes were measured and normalized using the estimated global scale factor (s). The 3D-3D registration was achieved using a method similar to that described in [17]. This is not the main focus of this paper, so we do not provide details of the implementation here.

Another important issue is the reliability of the 2D-3D registration, as it may give misaligned results due to low quality of the fluoroscopic sequence. Since the kinematic pose represents the "average" pose of the carpal bones, the local deviation from the kinematic pose should be relatively consistent across the sequence. A particular frame showing a larger deviation from the kinematic pose than other frames may indicate a failed registration at that frame. Hence, the 3D Euclidean distance between the local refined bone pose and the kinematic pose is used to indicate the reliability of the registration, which is calculated by

$$r = \frac{1}{p}\sum_{i=1}^{p} \|x_i^g - T^l(x_i^g)\|^2. \tag{12}$$

In (12), x_i^g represents the ith 3D mesh points after the global pose and SPM estimation. p is the total number of mesh points of the current evaluated bone (In our case, $p = 1002$ for each bone). T^l is the local transformation matrix for that bone. Then the value r is subtracted from the mean deviation $\bar{r}$ of the whole sequence. This is denoted as δr.

The registration was considered as successful if the deviation δr is smaller than 1 voxel (experimentally determined). Furthermore, if the smallest r is larger than a threshold, it indicates the registration of the whole sequence may not be accurate, which needs visual check by the user.

TABLE II
ERRORS (MM) OF ESTIMATED 3D SCAPHOID–LUNATE DISTANCE BETWEEN REAL FLUOROSCOPIC SEQUENCES AND THE CORRESPONDING 3D VOLUMES

Subjects	1	2	3	4	5	6	7	8
Error	0.77±0.60	1.35±0.46	0.65±0.43	0.45±0.32	0.80±0.30	0.30±0.12	1.02±0.05	1.20±0.68
Subjects	9	10	11	12	13	14	15	Average
Error	1.38±0.24	0.19±0.22	0.87±0.47	0.90±0.74	0.75±0.45	0.99±0.50	2.37 ± 1.40	0.93±0.47

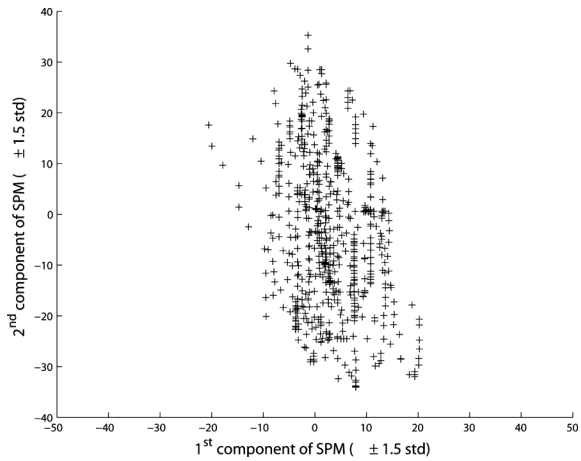


Fig. 6. Estimated first two SPM parameters for each frame of the 15 real fluoroscopic sequences.

The dSL of 15 subjects were calculated, each based on the original 3D volume and real fluoroscopic sequence of the same subject at their corresponding poses. Fig. 6 shows the estimated first two SPM parameters for each frame of the 15 real fluoroscopic sequences. As expected, the values of the second component, representing the major motion of the radial–ulnar fluoroscopic sequences, are distributed over the range of ± 1.5 standard deviation. The values of the first component (representing flexion–extension motion) are within a range of ± 0.5 standard deviation, making a small contribution to minimizing the out-of-plane transformation errors. Table II presents the mean and standard deviation of the absolute differences between the estimated dSL and the ground truth for each of the 15 subjects. Each estimated dSL was measured in the statistical model coordinate system by dividing each by their estimated scale factor, hence all the estimated dSL from different subjects can be compared at a consistent scale. 83.5% of the frames were considered as successful using the criterion based on (12), and these were used to generate the measurements shown in Table II. The average estimated error of successful registrations is 0.93 ± 0.47 mm, indicating good agreement of the dSL estimated from the real fluoroscopic sequences and the original 3D volume.

A sample frame of the matching result and the corresponding 3D poses are shown in Fig. 7 in which the projected contours from the 3D mesh model are superimposed on the preprocessed fluoroscopic image. The estimated 3D mesh model in the palmar and dorsal views is shown in middle and right, respectively. The registration result for the whole sequence can be found in [23].

C. Measurement Model for Pathology Detection

Our 3D CT and fluoroscopy datasets contain images of eight and six individuals, respectively, suffering from Scaphoid–Lunate dissociation, diagnosed radiologically on the basis of CT images. Here we demonstrate the potential to perform the diagnosis automatically from the fluoroscopic sequences.

The 3D CT volumes of 15 "healthy" subjects, assessed radiologically as not suffering from scaphoid-lunate dissociation, were used to determine a "standard" model, based on neutral and extreme radial–ulnar poses. The statistical mesh model was aligned with these volumes by estimating the global rigid transformation parameters, the SPM parameters and the local transformation parameters for each bone (see Section V-B). The kinematic poses at intermediate wrist positions were determined by cubic spline interpolation between the extreme and neutral positions, sampled at every two integer values of the second (radial–ulnar) component of the SPM, giving 36 wrist positions. In calculating the distance between bones we use the distance between corresponding surface points. As mentioned in Section V-A, correspondences can be established between surface points on different bones. Here we use a reduced number of surface points ($N = 100$, rather than 1002 used in building the model) for improved computational efficiency. Equation (13) and (14) show that we calculate the Mahalanobis distances (MD) using the means and covariances of individual pairs of corresponding points, rather than using the average distance, as in Section V-A. Letting $l^k_{\phi,j}$ and $s^k_{\phi,j}$ represent the jth surface point on the kth sample volume at pose ϕ on the lunate and scaphoid, respectively, the relative distance between the lunate and scaphoid at point j is

$$d^k_{\phi,j} = l^k_{\phi,j} - s^k_{\phi,j} \tag{13}$$

$d^k_{\phi,j}$ is a 3×1 vector, so the mean $m_{\phi,j}$ and covariance matrix $C_{\phi,j}$ of the jth point pair based on all k samples at pose ϕ can be calculated. The Mahalanobis distance between the new test data and the model at pose ϕ is calculated using

$$m_\phi = \frac{1}{N}\sum_{j=1}^{N}\sqrt{(d^{\text{new}}_{\phi,j} - m_{\phi,j})^T C^{-1}_{\phi,j}(d^{\text{new}}_{\phi,j} - m_{\phi,j})}. \tag{14}$$

To assess a new wrist, the 2D radial–ulnar fluoroscopic sequence can be registered with the statistical model using the methods described in Section IV, and the wrist poses determined by the second SPM component. The MD can then be calculated (14) at each pose to measure the deviation from the "standard" model. The results for the 25 (17 healthy and eight abnormal) simulated sequences and 15 (nine healthy and six abnormal) real fluoroscopic sequences are shown in Fig. 8. In this figure the triangles represent healthy subjects and the squares represent abnormal subjects. The lengths of the bars through the data points represent the reliability of each registration, as calculated in (12).

As shown in Fig. 8, for the simulated data, most of the abnormal subjects (squares) have larger MDs than the normal subjects (triangles). The distinction between the two groups is less

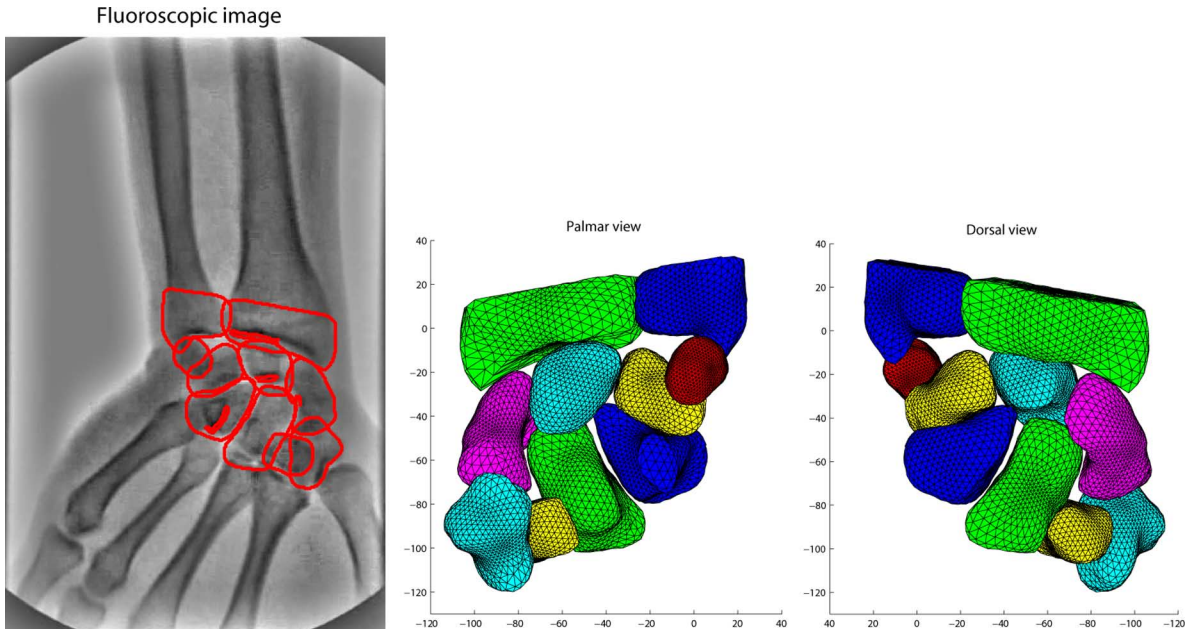


Fig. 7. Registration result of one frame from a real fluoroscopic sequence. The registration result for the whole sequence can be found in [23].

pronounced for the real fluoroscopic sequences. Additionally, the registration is less reliable compared with the simulated data, due to blurring effects generated by the wrist moving too fast.

By varying the threshold (the same threshold for all kinematic poses) of MD for classifying the normal and abnormal cases, the receiver operating characteristics (ROC) curve is generated and shown in Fig. 9. The ROC for both the simulated data and real data are presented, using only the successful registrations [Section V-B, (12)]. This resulted in using 89.3% of the frames for the simulated sequences and 83.5% of the frames for real sequences. The thresholds that produce the best error rate for simulated and real data are 2.75 and 2.86, respectively. These values result in 87.0% true positive rate (TPR) and 14.0% false positive rate (FPR) for simulated sequences, and 70.0% TPR and 30.0% FPR for real sequences.

The diagnostic conclusion for an individual can be obtained, by combining the classification results for all of the frames of the sequence. The test set for diagnosis is small, and the result rather dependent on a judicious choice of values for the MD threshold and the method used of combining the frames. We investigated two ways of deriving the classification result based on the MDs of frames. The first method is to use the weighted sum of the MD of each frame, which results in a single MD for each test sequence. The MD of each frame was weighted according to the reliability factor. The best operating point in the ROC evaluation, by varying the "averaged" MD threshold, is found at the threshold of 2.8 which resulted in sensitivity and specificity values of 68% and 90%, respectively. For the second method, we define the normal frame ratio (NFR) as the number of successful frames classified as "normal" divided by the total number of successful frames in the assessed fluoroscopic sequence. If the NFR is greater than a threshold, the particular subject is considered as "healthy," otherwise is diagnosed as having Scaphoid–Lunated dissociation. Fig. 10 shows the ROC curve obtained by varying the NFR, using a MD threshold of 2.5 (experimentally selected) for both the simulated and real data set. The highly quantized nature of the ROC curve reflects the size of the test set. The best operating point on this ROC curve is found at a NFR of 0.33 (requiring two thirds of the detected frames to be classed as abnormal before returning an abnormal diagnosis) resulting in sensitivity and specificity of 100% for simulated data and around 80% (83% TPR, 22% FPR) for real data. Other choices of MD threshold resulted in sensitivity-specificity combinations in the range (68%–90%) to (85%–70%).

VI. Concluding Remarks

We have presented a complete framework that is able to infer the 3D motion of carpal bones from a single view fluoroscopic sequence. It uses a hybrid statistical model to estimate both the pose and bone shapes from the fluoroscopic sequences allowing the motion of carpal bones during radial–ulnar deviation to be estimated. The positions and orientations in the image plane are estimated with high accuracy, and with slightly less accuracy in the out-of-plane direction. More importantly, the relative positions of the carpal bones can be estimated accurately. This is useful for detection of dissociation conditions. As an example of clinical application for this type of analysis, we have used Scaphoid–Lunate dissociation, where the underlying pathology is a rupture of one or more ligaments, and the diagnosis rests on a judgement regarding the bone separation.

The proposed framework was tested on both simulated (25 subjects) and real (15 subjects) fluoroscopic sequences in the leave-one-out manner. The average absolute 3D point to point registration error is 2.45 ± 1.07 mm, whereas the errors along the in-plane directions, X^M and Z^M, average about 0.5 mm. There have been no comparable studies reporting cross-subject 2D-3D registration of multiple objects based on a single view. For comparison, [14] and [15] estimated the shape of the femur

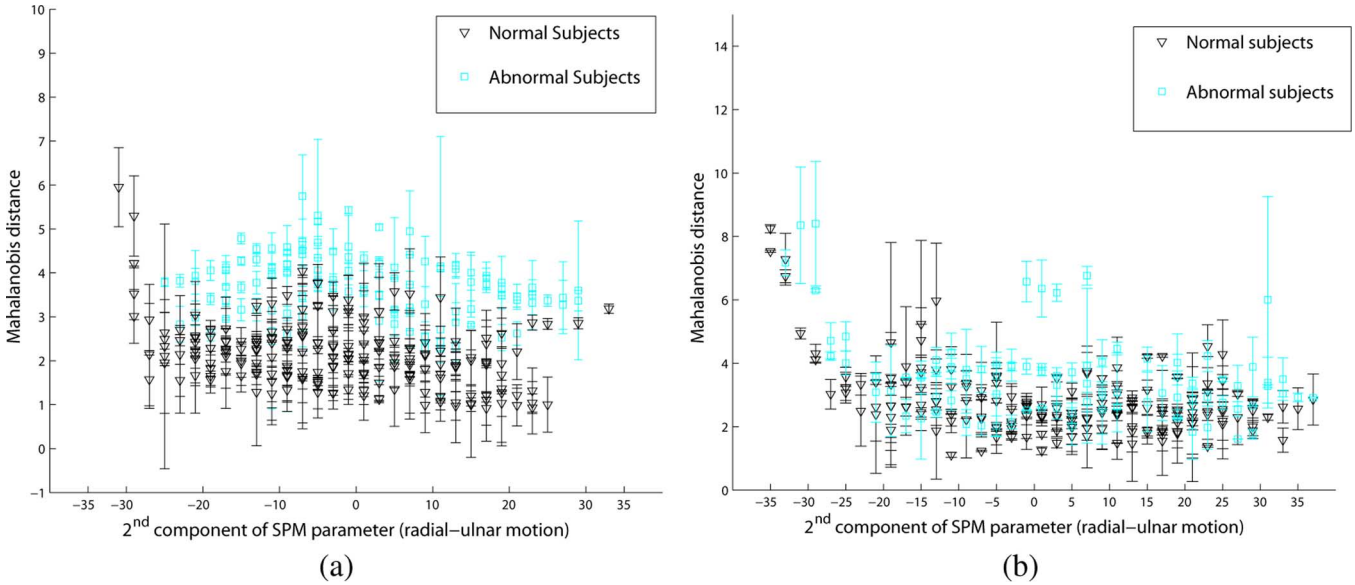


Fig. 8. (a) The Mahalanobis distances of 25 simulated sequences for Scaphoid–Lunate measurement. (b) The Mahalanobis distances of 15 real sequences for Scaphoid–Lunate measurement.

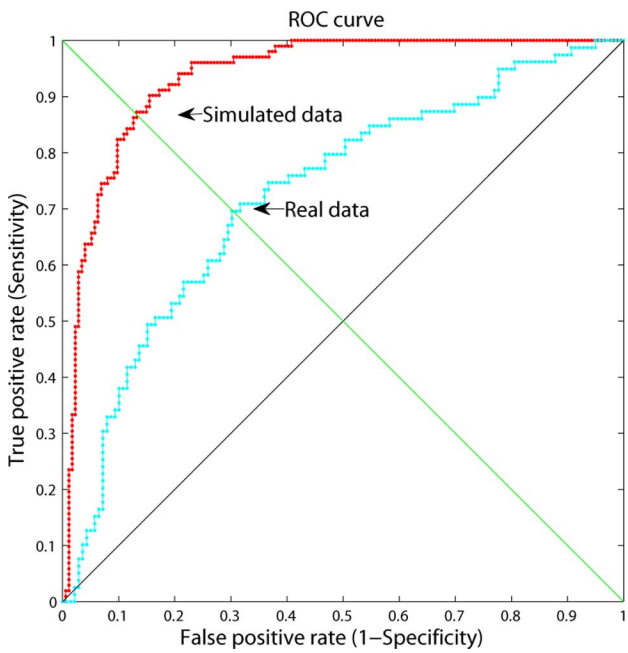


Fig. 9. ROC curve of the simulated data and real data for frame classification.

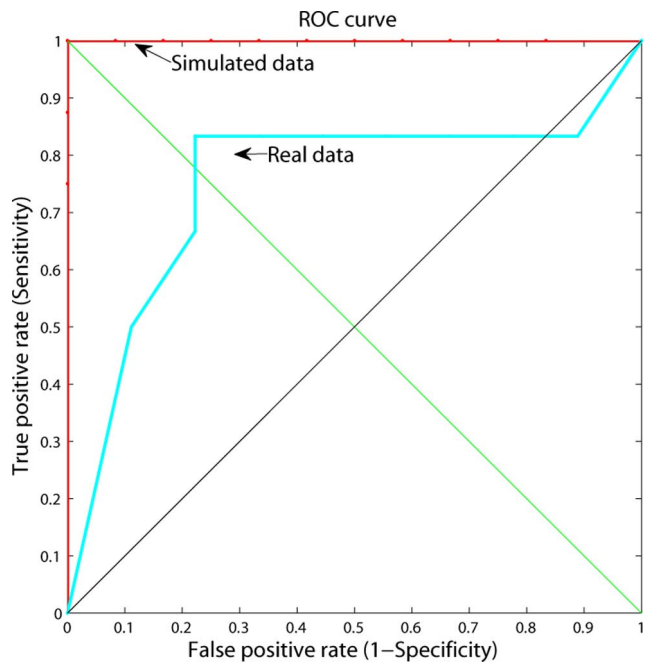


Fig. 10. ROC curve of the simulated data and real data for subject diagnosis.

(a much larger structure) based on biplane X-ray images, reporting root mean square errors of 1.48 ± 0.41 mm and 1.4 mm, respectively. In our case, the relative 3D distances between bones can be estimated more accurately. The error of the measured 3D Scaphoid–Lunate distances are 0.75 ± 0.50 mm for simulated data and 0.93 ± 0.47 mm for real data. In addition, the reliability of the registration can be estimated by comparing the deviation of each bone from the SPM model poses with the deviations of other frames in the same sequence.

We also proposed, and conducted a preliminary evaluation of a method for constructing a "standard" pathology measurement tool for automatically detecting Scaphoid–Lunate dissociation conditions, based on single-view fluoroscopic sequences. For the simulated data, it produced 100% sensitivity and specificity. For the real data, it achieved 83% sensitivity and 78% specificity. This tool could be a generic method for automatic, objective assessment of dissociation conditions. We have demonstrated its use with fluoroscopic video input. It appears that the limitation in accuracy arises largely from motion blurring effects in the video sequences. The method could equally well be applied using 2D radiographs at fixed positions. In a clinical setting, specified poses could be obtained using a fixation device.

The computational time for 1 frame was about 3 min running in Matlab on a 3.6 GHz machine. For a typical sequence, this would result in three to 5 h of computation, which would

be acceptable for an offline automatic analysis tool. If real-time feedback were required, faster computation would be necessary, which could be achieved by coding key parts in a compiled language, or use of GPU processing to parallelize the optimization process.

In further work, we will extend the current statistical model with more training data (in progress), and improve the measurement model by including more healthy subjects. A larger training set may allow us a different compromise between constrained model fitting and local refinement. Here we have sought to avoid local minima by restricting the SPM to only two modes of variation, relaxing the fit by local refinement. A larger dataset may result in a more specific model, making greater use of observed variability, reducing the need for the local refinement stage. However, if a range of abnormal conditions were to be included, the size of the training set might be prohibitive, requiring the retention of the local refinement. Our experience in this study indicates that it is a useful step in model fitting. On the basis of more data, we could further explore the relationship potentially associating the poses and shapes of bones. Nakamura *et al*. [24] have shown that carpal movement is affected by variation in the shape of the lunate. This raises the possibility that there may be more general relationships between bone shape and kinematics. It may be possible to build a more compact model by learning these relationships. We also intend to extend the framework to the (even) more challenging lateral views of flexion–extension motion, and further interpret the quantitative results for other wrist conditions. Acquiring a larger data set would also enable us to comprehensively test the classification performance. Currently training and evaluation are conducted using the same data in a leave-one-out fashion.

ACKNOWLEDGMENT

The authors are grateful for the helpful suggestions of the anonymous referees, which have contributed to this paper.

REFERENCES

[1] J. G. Snel, H. W. Venema, T. M. Moojen, M. Ritt, C. A. Grimbergen, and G. J. den Heeten, "Quantitative in vivo analysis of the kinematics of carpal bones from three-dimensional CT images using a deformable surface model and a three-dimensional matching technique," *Med. Phys.*, vol. 27, pp. 2037–2047, 2000.

[2] S. E. Sonenblum, J. J. Crisco, L. Kang, and E. Akelman, "In vivo motion of the scaphotrapezio-trapezoidal (STT) joint," *J. Biomechan.*, vol. 37, pp. 645–652, 2004.

[3] M. van de Giessen, G. J. Streekstra, S. D. Strackee, M. Maas, K. A. Grimbergen, L. J. van Vliet, and F. M. Vos, "Constrained registration of the wrist joint," *IEEE Trans. Med. Imag.*, vol. 28, no. 12, pp. 1861–1869, Dec. 2009.

[4] M. van de Giessen, M. Fournani, F. M. Vos, S. D. Strackee, M. Maas, L. J. van Vliet, K. A. Grimbergen, and G. J. Streekstra, "A 4D statistical model of wrist bone motion patterns," *IEEE Trans. Med. Imag.*, vol. 31, no. 3, pp. 613–625, Mar. 2012.

[5] M. Foumani, S. D. Strackee, R. Jonges, L. Blankevoort, A. H. Zwinderman, B. Carelsen, and G. J. Streekstra, "In-vivo three-dimensional carpal bone kinematics during flexion-extension and radio-ulnar deviation of the wrist: Dynamic motion versus step-wise static wrist positions," *J. Biomechan.*, vol. 42, pp. 2664–2671, 2009.

[6] C. Davatzikos, X. Tao, and D. Shen, "Hierarchical active shape models, using the wavelet transform," *IEEE Trans. Med. Imag.*, vol. 22, no. 3, pp. 414–423, Mar. 2003.

[7] J. J. Cerrolaza, A. Villanueva, and R. Cabeza, "Hierarchical statistical shape models of multiobject anatomical structures: Application to brain MRI," *IEEE Trans. Med. Imag.*, vol. 31, no. 3, pp. 713–724, Mar. 2012.

[8] J. Boisvert, F. Cheriet, X. Pennec, H. Labelle, and N. Ayache, "Geometric variability of the scoliotic spine using statistics on articulated shape models," *IEEE Trans. Med. Imag.*, vol. 27, no. 4, pp. 557–568, Apr. 2008.

[9] P. Marklj, D. Tomazevic, B. Likar, and F. Pernus, "A review of 3D/2D registration methods for image-guided interventions," *Med. Image Anal.*, vol. 16, pp. 642–661, 2010.

[10] G. P. Penney, P. G. Batchelor, D. L. G. Hill, D. J. Hawkes, and J. Weese, "Validation of a two- to three-dimensional registration algorithm for aligning preoperative CT images and intraoperative fluoroscopy images," *Med. Phys.*, vol. 28, pp. 1024–1032, 2001.

[11] D. B. Russakoff, T. Rohlfing, K. Mori, D. Rueckert, A. Ho, J. R. Adler Jr., and C. R. Maurer Jr., "Fast generation of digitally reconstructed radiographs using attenuation fields with application to 2D-3D image registration," *IEEE Trans. Med. Imag.*, vol. 24, no. 11, pp. 1441–1454, Nov. 2005.

[12] G. Zheng, "Statistically deformable 2D/3D registration for accurate determination of post-operative cup orientation from single standard X-ray radiograph," *Proc. MICCAI*, pp. 820–827, 2009.

[13] T. Whitmarsh, L. Humbert, M. De Craene, L. M. D. R. Barquero, and A. F. Frangi, "Reconstructing the 3D shape and bone mineral density distribution of the proximal femur from dual-energy X-ray absorptiometry," *IEEE Trans. Med. Imag.*, vol. 30, no. 12, pp. 2101–2114, Dec. 2011.

[14] N. Baka, M. de Bruijne, T. van Walsum, B. L. Kaptein, J. E. Giphart, M. Schaap, W. J. Niessen, and B. P. F. Lelieveldt, "Statistical shape model based femur kinematics from biplane fluoroscopy," *IEEE Trans. Med. Imag.*, vol. 31, no. 8, pp. 1573–1583, Aug. 2012.

[15] G. Zheng, S. Gollmer, S. Schumann, X. Dong, T. Feilkas, and M. A. G. Ballester, "A 2D/3D correspondence building method for reconstruction of a patient-specific 3D bone surface model using point distribution models and calibrated X-ray images," *Med. Image Anal.*, vol. 13, pp. 883–899, 2009.

[16] X. Chen, J. Graham, C. E. Hutchinson, and L. Muir, "Inferring 3D kinematics of carpal bones from single view fluoroscopic sequences," *Proc. MICCAI 2011*, vol. 6892/2011, pp. 680–687, 2011.

[17] X. Chen, J. Graham, and C. E. Hutchinson, "Integrated Framework for Simultaneous Segmentation and Registration of Carpal Bones," in *18th ICIP*, 2011, pp. 433–436.

[18] J. J. Craig, *Introduction to Robotics: Mechanics and Control*. Boston, MA: Addison-Wesley, 1989.

[19] R. H. Davies, C. Twining, T. F. Cootes, and C. J. Taylor, "Building 3D statistical shape models by direct optimisation," *IEEE Trans. Med. Imag.*, vol. 29, no. 4, pp. 961–980, Apr. 2010.

[20] N. Amenta, "The crust algorithm for 3D surface reconstruction," in *Proc. 15th Annu. Symp. Computat. Geometry*, 1999, pp. 423–424.

[21] M. J. Black and G. Sapiro, "Edges as outliers: Anisotropic smoothing using local image statistics," *Scale-Space Theories Comput. Vis.*, vol. 16822, pp. 259–270, 1999.

[22] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C++*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2002.

[23] [Online]. Available: http://personalpages.manchester.ac.uk/staff/xin.chen/CarpalReg.htm

[24] K. Nakamura, M. Beppu, R. M. Patterson, C. A. Hanson, P. J. Hume, and S. F. Viegas, "Motion analysis in two dimensions of radial-ulnar deviation of type I versus type II lunates," *J. Hand Surg.*, vol. 25, no. 5, pp. 877–888, 2000.

60. **Automatic generation of statistical pose and shape models for articulated joints.** X Chen, J. Graham, C.E. Hutchinson and L. Muir. *IEEE Trans Medical Imaging, 33(2), 372 – 383, 2014*. doi: 10.1109/TMI.2013.2285503

Automatic Generation of Statistical Pose and Shape Models for Articulated Joints

Xin Chen*, Jim Graham, *Member, IEEE*, Charles Hutchinson, and Lindsay Muir

***Abstract*—Statistical analysis of motion patterns of body joints is potentially useful for detecting and quantifying pathologies. However, building a statistical motion model across different subjects remains a challenging task, especially for a complex joint like the wrist. We present a novel framework for simultaneous registration and segmentation of multiple 3-D (CT or MR) volumes of different subjects at various articulated positions. The framework starts with a pose model generated from 3-D volumes captured at different articulated positions of a single subject (template). This initial pose model is used to register the template volume to image volumes from new subjects. During this process, the Grow-Cut algorithm [1] is used in an iterative refinement of the segmentation of the bone along with the pose parameters. As each new subject is registered and segmented, the pose model is updated, improving the accuracy of successive registrations. We applied the algorithm to CT images of the wrist from 25 subjects, each at five different wrist positions and demonstrated that it performed robustly and accurately. More importantly, the resulting segmentations allowed a statistical pose model of the carpal bones to be generated automatically without interaction. The evaluation results show that our proposed framework achieved accurate registration with an average mean target registration error of 0.34 ± 0.27 mm. The automatic segmentation results also show high consistency with the ground truth obtained semi-automatically. Furthermore, we demonstrated the capability of the resulting statistical pose and shape models by using them to generate a measurement tool for scaphoid-lunate dissociation diagnosis, which achieved 90% sensitivity and specificity.**

***Index Terms*—Articulated joint, carpal bones, segmentation, statistical pose model, statistical shape model, three-dimensional (3-D) image registration, wrist.**

I. Introduction

A NUMBER of recent studies have made use of statistical models for determining and quantifying abnormal articulated motion of anatomical joints (e.g., in the spine [2], [3], the femur [4], and inferring 3-D motion from 2-D video sequences in the wrist [5]). The principal underlying problem in building such statistical models is to establish correspondences representing important features across populations.

Davies *et al.* [6] have shown how such correspondences can be achieved automatically, given a segmented training set. However, for a complex joint like the wrist, segmentation of the wrist bones from CT volumes is a challenging task. They suffer from variation in intensity due to the nature of the trabecular bone and indistinct boundaries due to partial volume effects and the narrow gap between adjacent surfaces ([7]–[10]). It is a common experience, which we share, that "automatic" segmentation methods do not produce sufficiently accurate results and that "semi-automatic" methods such as those based on Graph-cuts often require complex interactions for every training example. The problem is frequently solved by tedious manual segmentation. Furthermore, once segmentation is achieved further registration is required to align different articulated positions of different subjects [11]. This is complicated in articulated joints as the relative poses of different components vary throughout the motion of the joint. The wrist is a particularly challenging example as it comprises eight carpal bones, the radius and the ulna (see Fig. 2) moving in a complex 3-D pattern. Recently, Cootes *et al.* [12] introduced a framework to compute dense correspondences across groups of images based on groupwise image registration, which has been successfully applied to face and brain images. However, for the registration of bones, the features mainly lie on the surface of the bone and the pose variation involves significant articulation of rigid parts. This is not accommodated well in [12].

In this paper, we present a method for automatic segmentation and registration of bones in an articulated joint (specifically the wrist) in a range of articulated positions across a group of individuals for the purpose of building a statistical pose model (SPM). There is a small body of research addressing this question. van deGiessen *et al.* [11] introduced a constrained registration of the wrist joint based on segmented 3-D surfaces using the iterative closest point (ICP) method, resulting in a 4D statistical model of wrist bone motion patterns [13]. The model represents local statistical properties between adjacent carpal bones by a set of predetermined point correspondences, and is used for detecting abnormal bone spaces. Boisvert *et al.* [2] studied spine variation using 3-D articulated pose models. The relative rigid transformation parameters of each vertebra with respect to the vertebra of the upper level were used to construct the articulated pose model. The spine variations between the same set of patients before and after treatment were compared using the model. Marai *et al.* [14] proposed a cost function based on distance fields for carpal bone registration, which is validated by

Manuscript received July 24, 2013; revised September 20, 2013; accepted October 04, 2013. Date of publication October 11, 2013; date of current version January 30, 2014. This work was supported by Medical Research Council, U.K., under Grant 87997. *Asterisk indicates corresponding author.*

*X. Chen is with the Centre for Imaging Sciences, Institute of Population Health, The University of Manchester, M13 9PT Manchester, U.K. (e-mail: xin.chen@manchester.ac.uk).

J. Graham is with the Centre for Imaging Sciences, Institute of Population Health, The University of Manchester, M13 9PT Manchester, U.K.

C. Hutchinson is with the Division of Health Sciences, University of Warwick, CV4 7AL Coventry, U.K.

L. Muir is with the Department of Hand Surgery, Salford Royal NHS Foundation Trust, M6 8HD Salford, U.K.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2013.2285503

0278-0062 © 2013 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requies IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

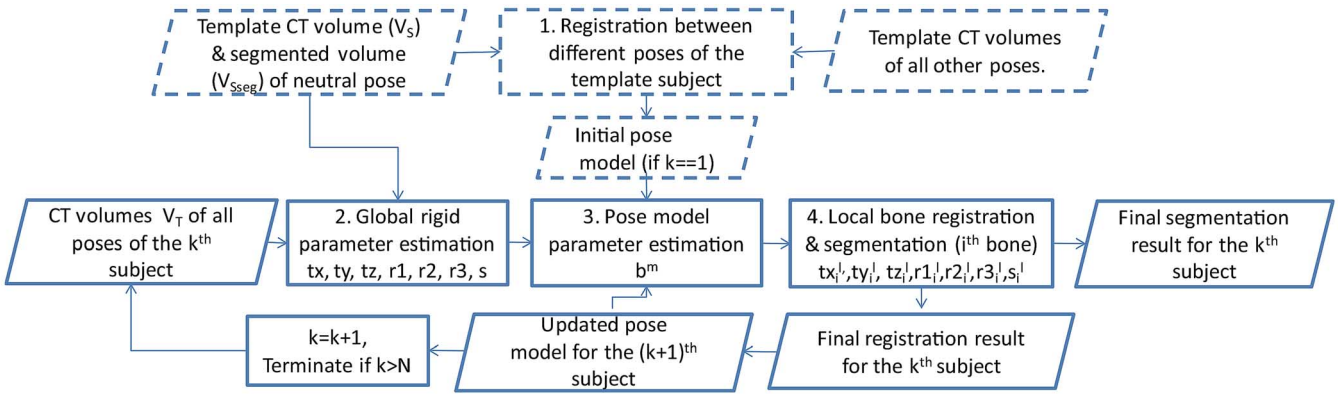


Fig. 1. Overview of the proposed system. Dashed boxes represent the initial registration and segmentation of the randomly selected template volume. Solid boxes represent the iterative registration and segmentation of the remaining examples.

aligning different poses for the same subject. In these studies, bone segmentations are performed either manually or semi-automatically prior to the registration. Here we describe a method that exploits the fact that identical bones from the same individual occur in different poses in the same data set. The different examples of these bones adopt different poses relative to the other bones as the joint moves. Examples of similar bones in similar poses are observed for different individuals. We have previously presented a framework for combined segmentation and registration [15]. That work is extended here by removal of the requirement for initial manual alignment prior to registration and incorporating the SPM building process into the framework.

The main contributions of this paper, distinguishing it from the aforementioned studies, are as follows. 1) it is an automatic framework where a statistical pose model (SPM) and a statistical shape model (SSM) can be generated via integrated segmentation and registration methods. 2) A consistent global scale factor is estimated by simultaneous registration performed on all articulated positions of the same subject, which leads to an accurate and compact SPM. 3) In contrast to [13], a statistical description of the *global* motion pattern of all carpal bones is calculated, from which the local pattern of motion between adjacent bones can also be described directly. Additionally, the SPM parameters can be used to align different wrist positions from different subjects, without the requirement (as in [13]) for a set of predefined positions. 4) The segmentation results are produced by a combination of data from all the wrist positions, which produces more reliable and consistent results than only using one wrist position for segmentation (e.g., [15]). 5) We avoid the requirement for interactive alignment (as in [15]) by basing the registration and segmentation on the pose model. This model, initially based on a single subject, grows incrementally as further subjects are registered and segmented. This results in fully automatic registration and segmentation. 6) The use of SPM and SSM as a measurement tool for pathology detection is demonstrated based on the Scaphoid–lunate dissociation condition.

The proposed framework, illustrated in outline in Fig. 1, is designed to align a template volume V_S to N target subjects, each with a number of different wrist positions. In our training datasets, CT data from 25 subjects, each at five different wrist positions were used (neutral position and four extreme positions in radial–ulnar and flexion–extension movement). The process consists of four steps. Step 1 is a preprocessing step that only needs to be done once. We randomly select one subject from the training data sets as the template, and segment a CT volume from one of the wrist positions (e.g., neutral position) using the grow-cut interactive segmentation method [1]. The segmented position is then registered to other positions within the template subject using the method described in [15] (see Sections IV-A and IV-B). The registration result is used to derive a pose model (described in Section III). In steps 2–4, the template is propagated to all the positions of the kth target subject simultaneously, providing estimates of the global rigid parameters, pose model parameters and local rigid parameters of each bone. In step 4, the Grow-Cut multi-label segmentation method is integrated with the registration process, which improves the robustness of the registration [15] and also generates a final segmentation. The successful registration result is then used together with the previously available (k-1) registration results to produce an updated pose model for the next iteration. An outlier rejection strategy is used for pose model updating. The whole process is terminated when all subjects in the training data set are registered to the template. Detailed descriptions of each step are given in the following sections.

II. Problem Parameterization

A coordinate system is defined (see Fig. 2) across all the subjects, in order to represent a consistent wrist motion. The origin of the coordinate system is defined with respect to the centroid of the radius bone. The X and Y coordinates are the corresponding coordinates of the centroid. As the length of the radius present in the image varies from subject to subject, the Z-coordinate (along the length of the radius) is defined, arbitrarily, to be 30 voxels above the lowest point in the radius of the template subject. The orientations of the X, Y, and Z axis are defined by the original CT volume coordinate system. All bone motions are represented relative to the origin. Three sets of parameters need to be estimated during image registration in order to interpret the true 3-D pose of each carpal bone. 1) Global wrist pose which is estimated by aligning the radius. It includes rigid transformation parameters and a global scale factor, denoted by $\theta = \{tx, ty, tz, r1, r2, r3, s\}$. $\mathrm{t} = [tx, ty, tz]^T$ denotes the

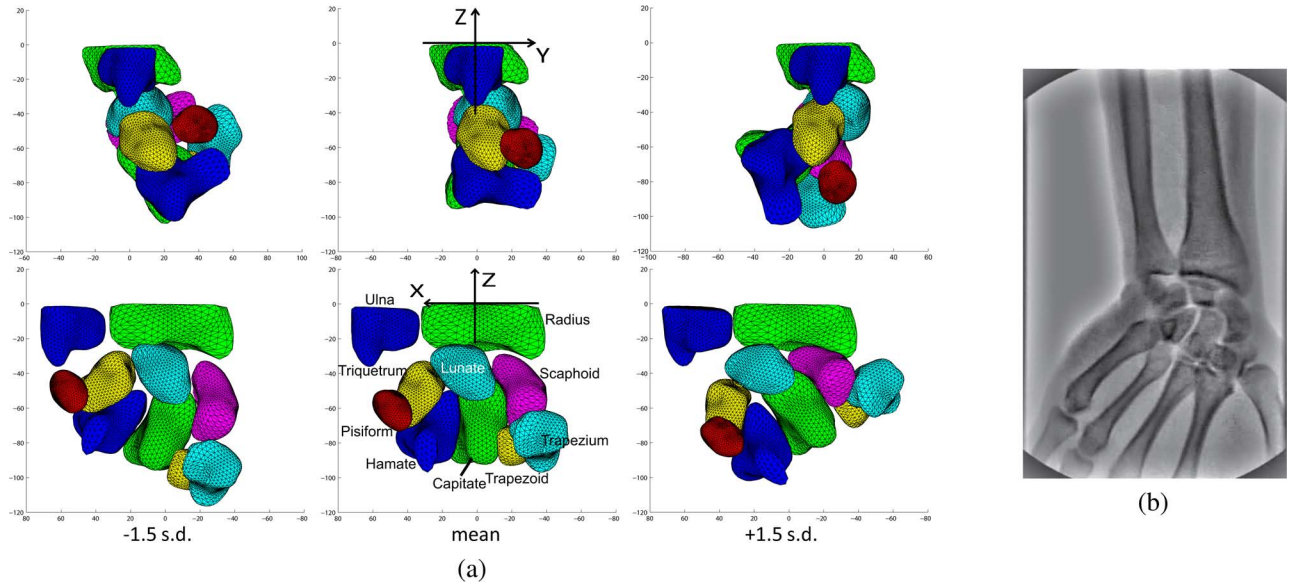


Fig. 2. (a) Poses of the first component (top row, lateral view) and the second component (bottom row, AP view) of the simple pose model. In each case the mean ± 1.5 s.d. are shown. (b) X-ray image showing the anatomical context of the carpal bones.

global translations, and $\mathrm{r} = [r1, r2, r3]^T$ is the Rodrigues parameter [13], [16] representing the global orientations. s controls the distance from the centroid of each bone to the origin in the radius, and the global size of the bones. 2) Poses of the carpal bones that are controlled by the statistical pose model parameters b^m [(6), m is a notation indicating the model parameters], which provides a rough alignment of the carpal bones. 3) Local rigid transformation parameters and a local scale factor for each bone $\beta_i = \{tx_i^l, ty_i^l, tz_i^l, r1_i^l, r2_i^l, r3_i^l, s_i^l\}$ (i is an index identifying each of the carpal bones, the radius and ulna), which provide refined alignment of bones based on the results from (2).

Using homogenous coordinates, the ith bone in the template volume coordinate system can be transformed to the target volume coordinate system by

$$A_i = PD_i \begin{bmatrix} ss_i^l Q_i \\ 1 \end{bmatrix} \tag{1}$$

where Q_i indicates the coordinates for the region of the ith bone in the template volume with respect to its own centroid. A_i is the transformed coordinates of the ith bone in the target volume. s_i^l is the local scale factor that controls the size of the ith bone. D_i is the pose matrix of the ith bone estimated using the pose model and the local pose refinement. P is the global rigid transformation matrix defined by

$$P = \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix} \tag{2}$$

where t is the translation vector $[tx, ty, tz]^T$. R is the 3×3 rotation matrix represented by Rodrigues parameters [13], [16], calculated as

$$R = I + K \sin|r| + K^2(1 - \cos|r|) \tag{3}$$

where $|r|$ is the magnitude of the rotation vector $[\mathrm{r1}, \mathrm{r2}, \mathrm{r3}]^T$. I is the identity matrix, and K is the skew-symmetric matrix normalized by $|r|$, expressed as

$$K = \begin{bmatrix} 0 & r3 & -r2 \\ -r3 & 0 & r1 \\ r2 & -r1 & 0 \end{bmatrix} \Big/ |r|. \tag{4}$$

In (1), D_i is calculated as

$$D_i = \begin{bmatrix} R_i^g & t_i^g \\ 0 & 1 \end{bmatrix} \tag{5}$$

where $t_i^g = s[tx_i^m, ty_i^m, tz_i^m]^T + [tx_i^l, ty_i^l, tz_i^l]^T$ is the summation of translation vectors estimated from the pose model and local bone refinement. $R_i^g = R_i^m R_i^l$ is the 3×3 rotation matrix that combines the rotations estimated from the pose model and local bone refinement respectively. R_i^m and R_i^l can be calculated individually by (3) using their corresponding Rodrigues parameters.

The use of Rodrigues parameters to represent bone orientations is convenient for pose model building and parameter optimization. More importantly, unlike the quaternion representation, it does not require vector normalization. Nor does it suffer from the singularity problem raised by the Euler angle rotations.

In the first step illustrated in Fig. 1, the template volume (V_S) is randomly selected from the training data sets. V_S is then segmented semi-automatically using the multi-label grow-cut method [1]. By using V_S and its corresponding segmentation result $V_{S\mathrm{seg}}$, each bone in V_S can be registered to other volumes at different wrist positions within the same subject using the method described in [15]. First, we manually adjust the transformation parameters of each bone in $V_{S\mathrm{seg}}$ to roughly align with the target volume, and then automatically refine the transformation parameters via intensity based registration [17]. The refinement is iterated until an acceptable alignment is achieved. The bones to be registered belong to the same subject, so there are no significant shape differences. While the bone poses may be

very different in different wrist positions, the initial rough interactive alignment makes it easy to achieve accurate registration. The interaction for each wrist position is also quick and efficient. This interactive step is carried out only once for a given template volume. The result is an initial pose model, which is used for subsequent automatic registration.

III. Statistical Pose Model Generation

The kinematics of the carpal bones is complex and significant pose difference can be introduced as the joint adopts different positions. In this section, we introduce a method for constructing the pose model of the carpal bones that enables the reproduction of valid poses with a small number of parameters. When carpal bones from different wrist positions of the same and different subjects are aligned, the pose model can be constructed from the transformation parameters of each bone with respect to a common reference coordinate system. The method of registration is described in Sections IV and V. After the registration, we use the six rigid transformation parameters for each bone to train the SPM. The common coordinate system for all bone poses has an origin in the radius of the template volume (Section II). Hence, for the SPM building, the transformation parameters for radius are always zero. The sizes of all the wrists are normalized to the template volume scale by using the estimated global scale factor. The pose of one subject is described by $(tx_1, ty_1, tz_1, r1_1, r2_1, r3_1, \ldots, tx_{10}, ty_{10}, tz_{10}, r1_{10}, r2_{10}, r3_{10})^t$ (eight carpal bones, one radius, and one ulna). Based on a set of training subjects, the pose model can be parametrized as

$$h = \mu^m + \nu^m b^m \tag{6}$$

where the mean pose μ^m (m is a notation indicating the model parameters) and the principal subspace matrix ν^m are computed using PCA. The vector b^m represents the pose parameters that describe the pose of h along each principal direction. For the initial pose model, only the volumes at different positions of the template subject are used. The first two significant components are shown in Fig. 2, which represent 99% of the variation. As more subjects are aligned with the template, the pose model is updated by including more training samples, and is used for subsequent registrations.

IV. Global Rigid Parameter and Pose Model Parameter Estimation

As illustrated in Fig. 1, by using V_S, $V_{S\text{seg}}$ and the pose model, the global rigid pose and poses of the individual carpal bones can be estimated in sequence, aligning the corresponding bones in V_S and V_T (target volumes). The registration process is to find the pose parameters that best align the corresponding bones in V_S and V_T by producing an optimum similarity value. The cost function that measures the similarity, and optimization method for estimating the pose parameters are described in the following subsections.

A. Similarity Measurement

To evaluate the similarity between the corresponding bone regions in V_S and V_T, we investigated several forms of the cost function (normalized correlation coefficient, sum of squared differences and mutual information based on intensities), achieving the best results from the one shown in (7), based on the difference of the normalized gradient magnitude of the two images. We define the normalized sum of squared difference (NSSD) between two images G_S and G_T as

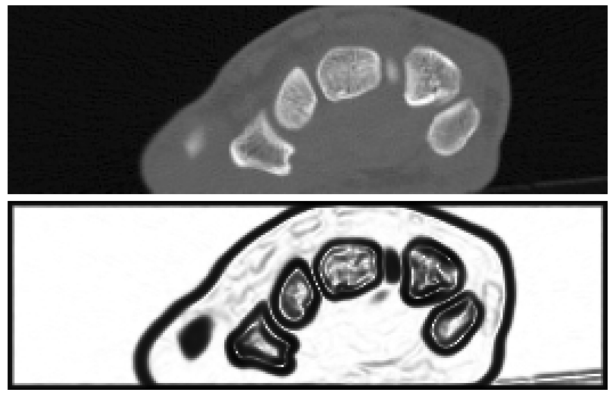


Fig. 3. Top: CT slice, bottom: corresponding normalized gradient magnitude.

$$E = \left(\frac{\sum_{j \in C} (G_S(T(j)) - G_T(j))^2}{w} \right)^{0.5} \tag{7}$$

where C represents the region of interest (ROI) corresponding to each specific bone and j indexes the voxels of C in G_T. The ROI is a region slightly larger than the bone volume, obtained by a dilation of 10 voxels along the three axes. w is the total number of voxels in C. $G_S(T(j))$ and $G_T(j)$ are the normalized values derived from the image gradient in the transformed template ROI image and the corresponding target image respectively, which are generated by

$$G_S(T(j)) = \frac{1}{(1 + \mu M_S(T(j)))}$$
$$G_T(j) = \frac{1}{(1 + \mu M_T(j))} \tag{8}$$

where T is the transformation (inverse to our estimated pose transformations) applied to C from target volume to template volume. M_S and M_T represent the gradient magnitude of the smoothed V_S and V_T respectively. V_S and V_T are smoothed by a $7 \times 7 \times 7$ Gaussian kernel with variance equal to 1 voxel (0.5 mm). μ was experimentally set to 0.1. Example images of the original CT slice and corresponding normalized gradient magnitude images are shown in Fig. 3.

B. Optimization

The global rigid pose parameter θ can be estimated by registration of the radius bone, since a point (defined in Section II) in the radius of the mesh model is used as the origin of the coordinate system for all motions. Other carpal bones and ulna are registered by estimating their individual pose parameters. We have target volumes captured at different positions (five wrist positions in our case) that belong to the same subject. They have different translation and rotation parameters, but should have a consistent scale factor s with respect to the template. Hence, we register the template bones to those target volumes simultaneously at each iteration (Step 3 in Fig. 1) When the translation and

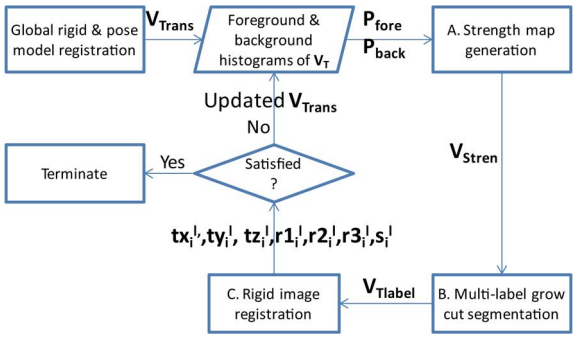


Fig. 4. Overview of the integrated segmentation and registration system corresponding to box 4 in Fig. 1.

rotation parameters are estimated, only the volume with corresponding wrist position is used to evaluate the cost function. In estimating the scale factor, s, on the other hand, volumes from all wrist positions are used to evaluate the cost function (summation of the cost function values). At the end of the registration, a set of translation and rotation parameters are obtained that correspond to the poses of the individual bones, and a single global scale factor is calculated for all wrist positions of the same subject. Individual scale factors for each bone are calculated in the local refinement stage (Section V). Scale invariance is important for pose model generation.

The optimization method we use is a simplified version of the Brent-Powell method [18], requiring a smaller number of optimization steps. We use parabola fitting to replace the Brent line search in the Brent–Powell method. The multi-dimensional search space ($\theta = \{tx, ty, tz, r1, r2, r3, s\}$, and b^m) is explored by iterative individual 1-D line searches. For each parameter search, the cost function is evaluated at the current position, positive and negative neighbor positions (defined by a search range), then an optimum is found by fitting a parabola to the three evaluated positions. The optimum is iteratively refined by reducing the search range until convergence. More details can be found in [19].

V. Simultaneous Registration and Segmentation for Local Pose Refinement

After performing the global rigid and pose model transformation, the template bones are approximately aligned with the bones in the target volume. Some local misalignment may still remain requiring a further step to refine the local pose of each bone. In this section, we introduce an integrated segmentation and registration method, which combines the multi-label Grow-Cut segmentation [1] and intensity-based registration. This method, illustrated in Fig. 4, is developed from that described in our previous paper [15], improved in several respects. To make this paper self contained, the following description includes details of the previously published version.

The main objective of the method is to estimate $\beta_i = \{tx_i^l, ty_i^l, tz_i^l, r1_i^l, r2_i^l, r3_i^l, s_i^l\}$ for the ith bone, improving its registration accuracy. The use of combined segmentation and rigid registration is preferred over nonrigid registration methods for this application. Finding the accurate pose parameters to align the bones is important for the SPM building. Nonrigid registration tends to deform the shape rather than finding the optimum pose. If rigid registration is performed individually, the topology of the bones may not be preserved and the bone volumes may overlap. This is overcome by combining the registration with multi-label Grow-Cut segmentation. In Grow-Cut, multiple labels are calculated simultaneously at each iteration and region overlapping is forbidden. It helps to make the registration more robust to the initial starting pose, and also acts as a soft constraint to preserve the topology of the bones. Subsequently, the segmentation results $V_{T\text{label}}$ can be used to build the statistical shape model of each bone. The overview of the framework is illustrated in Fig. 4; each of the key steps is described in the following subsections.

A. Strength Map Generation

There are two key elements in Grow-Cut segmentation [1]. They are the current label at each voxel and a strength map associated with the image. The strength map indicates the "energy" of the corresponding voxel, which is used to determine if the corresponding label can be propagated to its neighbors at each iteration. Since each labeled bone from $V_{S\text{seg}}$ (segmentation of V_S) has been roughly aligned with the target volume V_T through previous registrations, the labels for V_T can be therefore initialized using the transformed $V_{S\text{seg}}$, denoted as V_{Trans}. In V_{Trans}, all overlapped bone areas are set to zero, as new labels shouldn't be introduced. The initial label V_{Trans} will be evolved according to the associated strength map.

Here, we present a novel method for initializing the strength map V_{Stren} for Grow Cut. The objective is to initialize this map with values of 1 (high certainty) and 0 (low certainty) of being either bone or nonbone. To obtain the V_{Stren}, an initial binary volume, V_{bwTrans} (bone $= 1$, nonbone $= 0$) is generated from V_{Trans}. The normalized foreground and background histograms calculated from the overlap of V_{bwTrans} and the target volume V_T allow us to calculate the probability that a voxel belongs to the foreground (P_{fore}) or background (P_{back}). Using (9), we calculate the likelihood (V_L) of classifying each voxel as bone (positive) or nonbone (negative), from which (10) and (11) generate new binary volumes ($V_{\text{bw}L1}, V_{\text{bw}L2}$) representing high certainty regions of bone and nonbone respectively. The thresholds of 0.9 and -0.5 were determined empirically. $V_{\text{bw}L3}$ (12) represents the region of V_T that is not classified as bone either in V_{bwTrans} or $V_{\text{bw}L1}$. Equation (13) identifies the regions that are identified with certainty to be bone or nonbone, based on the histograms (P_{fore} and P_{back}), constrained to be within the respective bone and nonbone regions defined by V_{bwTrans}. Following Grow Cut relabelling, V_{bwTrans} and V_{Stren} are calculated for each iteration step

$$V_L = \frac{(P_{\text{fore}}(V_T) - P_{\text{back}}(V_T))}{\max(P_{\text{fore}}(V_T), P_{\text{back}}(V_T)))} \tag{9}$$

$$V_{\text{bw}L1} = \begin{cases} 1, & \text{if } V_L > 0.9 \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

$$V_{\text{bw}L2} = \begin{cases} 1, & \text{if } V_L < -0.5 \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

$$V_{\text{bw}L3} = 1 - (V_{\text{bwTrans}} \cup V_{\text{bw}L1}) \tag{12}$$

$$V_{\text{Stren}} = (V_{\text{bw}L2} \cap V_{\text{bw}L3}) \cup (V_{\text{bwTrans}} \cap V_{\text{bw}L1}). \tag{13}$$

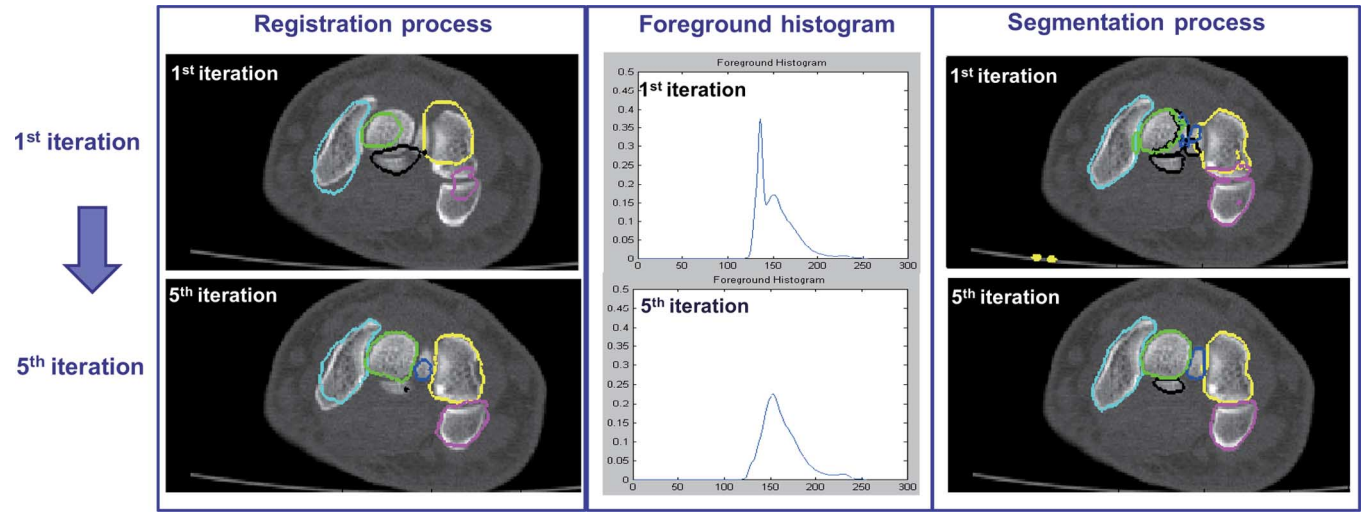


Fig. 5. Process of integrated segmentation-registration iteration for registering the template volume (colored contours) to a target volume of a different subject. At first (global) registration the correspondence between the template and target volumes is poor resulting in the inclusion of a significant region of background in the foreground histogram. The segmentation step improves the correspondence of the bones. At a later iteration the registration and the foreground histogram are improved. The resulting segmentation step results in good agreement between the segmentation and the target volume.

B. Multi-Class Grow Cut Segmentation

The advantages of Grow Cut in this application are its ability to obtain a multi-label solution in simultaneous iteration, and the capacity for fast parallel implementation. The segmentation labels of 10 bones (carpal bones plus radius and ulna) need to be updated simultaneously, helping to preserve bone topology. For efficiency the Grow Cut code was parallelized using NVidia Quadro FX 3800 Graphic Card via the CUDA API.[1]

In our proposed framework, the strength map V_{Stren} is initialised automatically in step A (Fig. 4), and V_{Trans} (from step A or updated from step C) is used as the labeled volume. Since, there is only a small number of uncertain voxels with $V_{\mathrm{Stren}} = 0$ at each iteration, it takes less than 2 s to complete the segmentation of a $141 \times 268 \times 169$ volume. The segmentation volume is denoted as $V_{T\mathrm{label}}$.

C. Rigid Image Registration

Following the segmentation, rigid image registration is performed. The cost function expressed in (14) is used as the similarity measurement in which a new term is added to the cost function described in (7). Since each bone has a unique label, the new cost function term tends to "drag" the template bones to the corresponding segmented regions, which preserves the topology of the bones

$$E_{\mathrm{local}} = E + \left(\frac{\sum_{j \in C} (B_S(T(j)) - B_T(j))^2}{w} \right)^{0.5} \qquad (14)$$

E is the gradient-based cost function in (7). B_S and B_T are the ROI binary image obtained from $V_{S\mathrm{seg}}$ and the corresponding binary image obtained from $V_{T\mathrm{label}}$, respectively. Other notations are the same as in (7). The optimization method is the same as described in Section IV-B. A new V_{Trans} is then obtained by using the updated transformation parameters.

[1]Available online: http://www.nvidia.com/object/cuda_home_new:html

D. Iteration and Termination

Step A to C are repeated; the segmentation volume, registration parameters and the intensity histograms coherently improve each other until the termination conditions are satisfied (the difference of the segmented volume $V_{T\mathrm{label}}$ between adjacent iterations stops decreasing). The iteration process is illustrated in Fig. 5. The foreground histogram, registration and segmentation result at the first and fifth (final) iterations are shown, where the colored contours from the template are superimposed on the target volume. The method described here differs from that described in [15] in that the transformation parameters and segmentations for all wrist positions are estimated in the same framework, and a consistent scale factor for each bone is calculated across all wrist positions. The final segmentation result is derived from all wrist positions. The labelled volumes at different positions are transformed to the template volume coordinate system. The overlapping area that is greater than 60% is used for the final label. Then the final label is transformed back to the volume at each position. The combination of segmentations in different wrist positions, and hence with different orientations relative to the sampling grid, reduces segmentation errors arising from partial volume effects. Combining the segmentation with the registration method makes the registration more robust than the registration only method, in terms of the sensitivity to the initial bone pose. This was evaluated in [15].

VI. Statistical Shape Model Generation

The shape of the bones varies among different subjects. A SSM of each bone in the wrist is also important for pathology diagnosis. The key step of generating a SSM is to establish correspondences across subjects. In some approaches this has been achieved using deformable registration (e.g., [12], [20], [21]). However, in these studies the principal aim is to establish shape, rather than pose correspondence. In the context of our framework, it is important to determine the correct relative bone poses, and deformable registration would tend to change

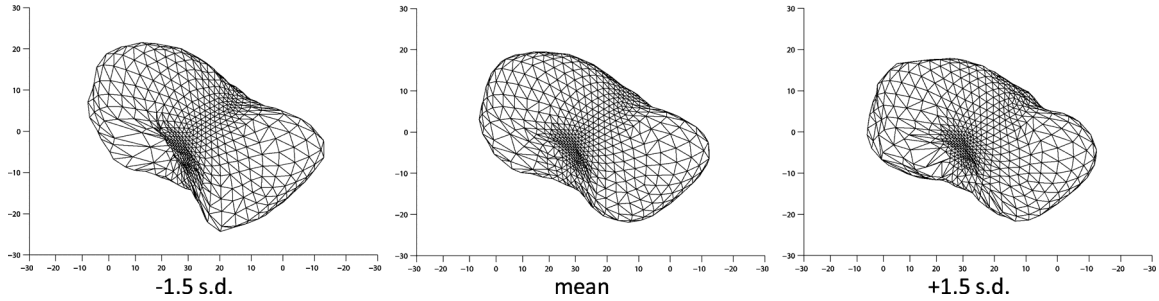


Fig. 6. First component of the shape model of the scaphoid. Mean ± 1.5 s.d. are shown.

the shape of the bones rather than finding the correct pose. This would result in a less accurate SPM, so rigid registration is preferred. It would be possible to apply deformable registration after rigid alignment. In this case the computational cost depends on the selected deformable model. The final result is highly dependent on the regularization method applied, and may be difficult to correct if increased accuracy is required. In our proposed framework, the shape differences between the corresponding bones of different individuals are accommodated by the segmentation process. Following the automatic registration and segmentation framework, the segmentation result can be directly used for the SSM construction. Our SSM is based on the Point Distribution Model (PDM- for example [22]). This requires the establishment of point correspondences between bones of different subjects, for which we use the well-known minimum description length (MDL) algorithm [6]. One training example is described by $(x_1, y_1, z_1, \ldots x_{1002}, y_{1002}, z_{1002})^t$ (1002 points on each bone). The coordinates of the shape points of each bone are expressed with respect to its own centroid. The statistical shape model, o_i, is then described as

$$o_i = \mu_i^q + \nu_i^q b_i^q \tag{15}$$

where μ_i^q and ν_i^q (q is a notation indicating the shape parameters) are the mean shape and the principal subspace matrix for the ith bone. b_i^q is the shape model parameter for generating new valid bone shapes. Fig. 6 shows the shapes that arise by varying the first component of SSM of the scaphoid.

VII. Framework Integration

Each part of the framework (Fig. 1) has been described in previous sections. Two important issues need to be further explained to complete the framework.

Firstly, in order to increase the robustness of the framework, the CT volumes are preprocessed to construct a multi-scale pyramid (downsampled by a factor of 2 at each level). In the optimization procedure, the same set of initial search ranges is used at each level for both the global and local registrations (described in Section VIII). The search ranges are divided by 2 each time the criteria are met, and the whole process is terminated when the maximum value of the search ranges is smaller than a preset threshold. To avoid the optimization becoming trapped in local minima, a stochastic optimization procedure is used for global parameter (wrist pose and scale factor) and pose model parameter (carpal bone pose) estimation, as follows.

1) Starting from zero transformation and orientation, optimize the poses of the bones of all input wrist volumes.
2) Record the best cost function value for each pose.
3) Randomly alter the starting value of the parameters for unsatisfied poses (defined in step 4) within the possible parameter space, and optimize again.
4) Repeat steps 2 and 3. Terminate the random process for the pose if the best function value of that pose remains unchanged for five times or the number of iterations exceeds 20.

This stochastic process is only performed on the lowest pyramid level for computational efficiency and robustness.

Secondly, at each iteration of the SPM updating procedure (see Fig. 1), an outlier rejection algorithm is applied to exclude inaccurate registrations. After the registration is finished for each subject, the cost function values for each bone across all wrist positions are compared. Since the data is in the same image modality and from the same subject, the cost function value for the corresponding bones should be similar, independent of bone poses. Only the wrist positions with the cost function value less than $1.2\times$ the best (smallest) cost function value are considered as successful registration. After all of the subjects are visited, those excluded subjects are revisited and aligned again by using the SPM generated from the included subjects. If successful registration is achieved, the revisited subject will be included to update the SPM. The process is repeated until the number of included subjects is unchanged. The unregistered volumes do not contribute to the model.

VIII. Evaluation

We evaluated our framework based on CT data from 25 subjects (10 female and 15 male, median age 51, age range 25–72 years), recruited from the hand clinic at Salford Royal Hospital, Greater Manchester, U.K. Eight of these subjects were diagnosed radiologically as suffering from scaphoid–lunate dissociation (referred to as the "abnormal" group in the following discussion), the remainder being assessed not to have this condition (referred to as the "normal" group). Each subject was imaged at five different wrist positions: neutral, and four extreme positions in radial–ulnar and flexion–extension movement. The wrist positions were held on a specially designed foam. Each of the CT volumes is captured by a GE LightSpeed VCT machine with a very low-dose exposure. The exposure from all five scans was 20 mGy. The acquisition parameters were: tube voltage of 80 kV, focal spot of 0.7 mm, slice thickness of 0.625 mm, pixel spacing of 0.29×0.29 mm^2. The volumes were resampled by

TABLE I
MEAN AND STANDARD DEVIATION REGISTRATION ERRORS (MEASURED BY MEAN TARGET REGISTRATION ERROR IN MM) AND THE SUCCESSFUL REGISTRATION RATE OF EACH BONE BASED ON USING EACH OF THE 25 SUBJECTS IN TURN AS THE INITIAL TEMPLATE

Bones	Ulna	Radius	Triquetrum	Lunate	Scaphoid
mTRE (mm)	0.29±0.24	0.22±0.17	0.33±0.26	0.42±0.34	0.43±0.33
SucRate (%)	80.2±3.7	99.8±0.5	95.6±3.1	96.7±2.0	97.3±2.2
Bones	Pisiform	Hamate	Capitate	Trapezoid	Trapezium
mTRE (mm)	0.39±0.30	0.28±0.21	0.4±0.34	0.38±0.33	0.30±0.25
SucRate (%)	95.3±3.1	97.1±1.0	93.1±7.6	90.9±3.1	97.1±1.9

tri-linear interpolation to iso-cubic volumes of $0.5 \times 0.5 \times 0.5$ mm^3 prior to the registration and segmentation. We found that higher resolution (e.g., $0.25 \times 0.25 \times 0.25$ mm^3) did not produce a much better accuracy of the segmentation and registration but required much larger memory and longer computational time.

Using the interactive method described in [15] we obtained the segmentation of each of these subjects in the neutral position and the transformation parameters that relate the neutral position to the extreme positions for that subject. The segmentations at each position were validated by an experienced clinician. These segmented and registered images were used as ground truth in the evaluation of the automated framework described here.

In our experiments, each of the 25 subjects was selected as the template and registered with the remaining 24 subjects in turn. The registration order to other subjects was randomly sequenced. In the optimization procedure, the same set of fixed initial search ranges was used at each level for both the global and local registrations (four voxels for translation, $4\pi/180$ for rotation, 0.2 for scale and one standard deviation for pose model parameters). The framework terminated when the largest search range was smaller than 0.1.

A. Registration Results

To evaluate the registration results, we transformed the mesh points of each bone in neutral position to other positions using the ground truth registration parameters and our estimated registration parameters respectively, for each subject. Then the 3-D Euclidean distances of each corresponding mesh point between the two transformed meshes for each bone are measured (known as mean target registration error (mTRE) [23]). The registration errors are presented in Table I, showing the measurements from 25 (different initial template) $\times$ 24 (subjects) $\times$ 5 (positions) tests. The errors were only calculated based on successful registrations, defined by the outlier rejection scheme (Section VII). Specifically, for each bone, the successful registration rate is the total number of instances of that bone included by the outlier rejection algorithm divided by the total number of tested volumes. As shown in Table I, registration achieved subvoxel accuracy (mean error of 0.34 ± 0.27 mm). The successful registration rate of each bone across all tests are also presented in Table I. The successful registration rate for most bones is very high. The main exception is the ulna which is much lower than the others. The shape of the ulna is highly symmetric and its movement varies greatly between individuals. In some individuals, there is relatively little movement, while in others there may be significant movement. Of the others, the lowest successful registration rate occurred in the trapezoid. This bone has a nearly spherical shape, making calculation of the orientation rather unstable. It is also the smallest of the carpal bones.

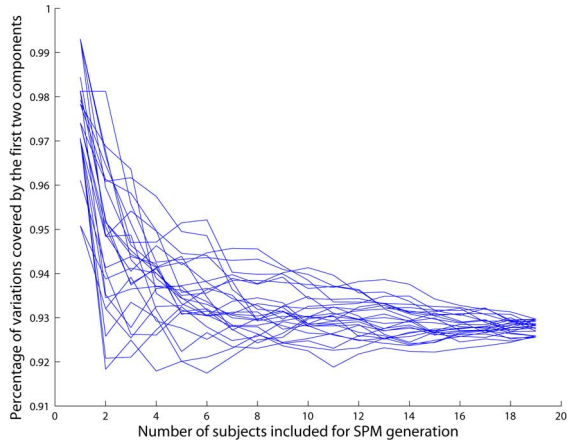


Fig. 7. Percentage of variation covered by the first two significant components in SPM updating process. Each line represents the use of a different subject as the template.

The standard deviation of the success rate indicates that the success rate and registration error are not very sensitive to the selection of initial template, as the SPM is updated each time a new subject is included, and the failed registrations are revisited in a larger loop. In our 25 independent tests, each based on a different initial template, 19 or 20 out of 25 subjects were consistently successfully registered (all positions successfully aligned) and used for final SPM generation. Subjects were excluded from model building if any of the positions for that subject were rejected by the automatic framework. In each rejected subject the failed positions arose because of misalignment of either the ulna or trapezoid (or both).

B. Statistical Pose Model Updating

Only the first two significant components of the SPM were used throughout the whole registration across all subjects. When the SPM is updated at each iteration, the percentage of variations captured by the first two significant components decreases. The percentage variation represented by the first two components in the evolving process of the SPM over the 25 independent tests is shown in Fig. 7. This converged to 92%–93% variation, irrespective of the registration sequence and initial template selection. An animation of generating intermediate poses by varying the first two components of the final SPM can be found in the supplementary material. In our experiments, including more PCA components did not increase the registration accuracy. This also indicates that the integrated registration and segmentation local refinement step works very well based on the starting pose provided by the pose model.

C. Segmentation Result

We compared the segmentation results with ground-truth using the Tanimoto coefficient (TC) [24] (also known as the Jaccard similarity coefficient [25]), presented in Table II, showing the mean and standard deviation from 25 independent tests using different initial template subjects. The Tanimoto

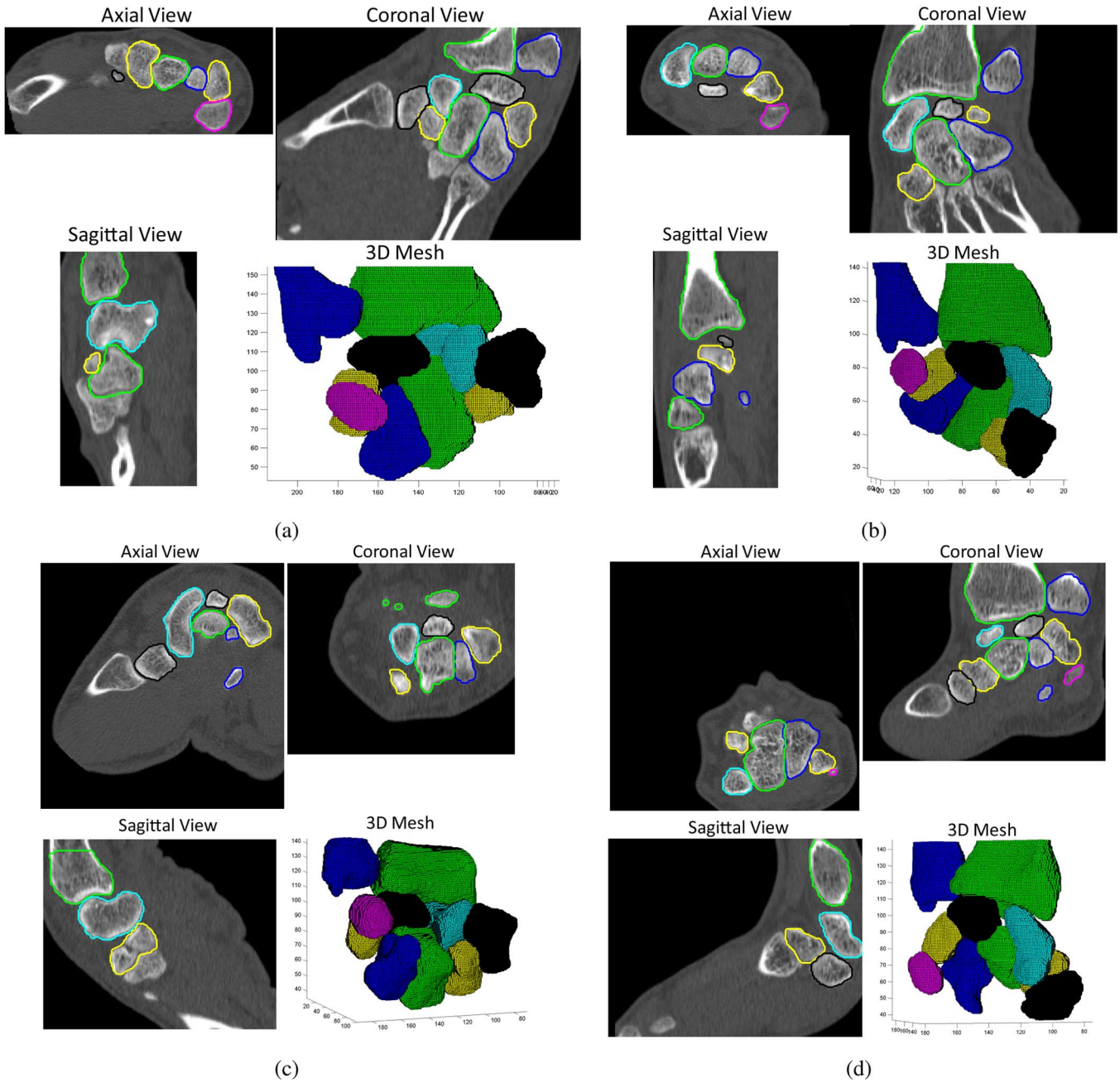


Fig. 8. The axial view, coronal view, sagittal view, and 3-D mesh model of automatically segmented wrists at (a) radial deviation, (b) ulnar deviation, (c) flexion, (d) extension positions of different subjects. Unlabelled bones that appear in some images are parts of the metacarpal bones which are not included in the segmented template. Colored lines of the 2-D-slices correspond to the colors in the 3-D mesh models.

TABLE II
TANIMOTO COEFFICIENT COMPARING THE GROUND TRUTH SEGMENTATION AND AUTOMATICALLY CALCULATED SEGMENTATION FOR EACH BONE. MEAN AND STANDARD DEVIATION ARE SHOWN, BASED ON 25 INDEPENDENT TESTS USING DIFFERENT INITIAL TEMPLATE SUBJECTS

Bones	Ulna	Radius	Triquetrum	Lunate	Scaphoid
TC	0.86±0.10	0.88±0.06	0.90±0.06	0.88±0.10	0.87±0.09
Bones	Pisiform	Hamate	Capitate	Trapezoid	Trapezium
TC	0.90±0.10	0.91±0.03	0.83±0.10	0.82±0.09	0.90±0.10

coefficient measures the similarity between two sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets. The coefficient is between [0, 1], where 1 indicates perfect overlap of the compared images. These segmentation results, arising entirely from the automated framework without further refinement, show a high level of segmentation accuracy and repeatability. An example of the automatically segmented carpal bones is presented in Fig. 8, showing the sagittal, axial, and coronal view with the 3-D mesh model.

D. Pathology Detection

In previous sections, we presented a complete framework for automatically generating a SPM and SSM across subjects at different positions. In this section, we demonstrate the effectiveness of the SPM and SSM by applying them to pathology detection and quantification. Using a "standard" SPM and SSM, the

relative poses of the carpal bones can be measured and recorded as "standard." One condition that may be assessed using this measurement is bone dissociation, where the 3-D distance between the bones is larger than normal. Scaphoid–Lunate dissociation is one of the most common of these and we use it as an example to demonstrate the method of using the SPM and SSM for diagnosis.

Since the bone spaces may vary at different wrist positions, the constructed "standard" measurement tool needs to be position dependent. One major advantage of using the SPM is that the wrist from different positions can be aligned directly based on the SPM values. The first two components of our SPM cover more than 90% of the observed variation in the full range of flexion–extension and radial–ulnar movements. The combination of the two components is also able to generate interpolated positions within the motion range. To simplify the problem, here we only demonstrate the pathology detection tool based on radial–ulnar deviation movement, which is the most appropriate for diagnosing Scaphoid–Lunate dissociation.

1) Building the "Standard" Scaphoid/Lunate Model: 14 subjects in the normal group, each at three different positions (neutral and extreme radial–ulnar deviation), were used to build the "standard" measurement model. Each of the 14×3 target volumes is aligned with the final SPM (based on 25 subjects, each at five wrist positions) from the registration framework (Section VII) to obtain the poses of the individual bones, and hence the scaphoid–lunate distance. Since there are not enough samples that cover the full range of continuous positions, we interpolated (cubic spline) the positions at integer intervals of the pose model parameter (Fig. 2) between the neutral position and two extreme positions of radial–ulnar deviation for each subject. This results in the training volumes being grouped at each integer interval of the second component of the SPM parameter. (The second component of the SPM mainly represents radial-ulnar movement—see Fig. 2).

To diagnose the Scaphoid–Lunate dissociation conditions, the "standard" range of distances between these two bones need to be calculated and recorded. The "standard" SSM represents a range of shapes for each bone. To maintain consistency in the measurement of distances, each carpal bone is represented by its mean shape. The SSM has the same number of surface points (1002) on each bone. The point correspondences between different bones are established by using the index of the mesh points on the surface. Here we used evenly down-sampled number of surface points N ($N = 100$ in our case) to reduce the memory usage and improve the computational efficiency. If each of the jth selected surface points on the Lunate and Scaphoid of the kth sample volume at position ϕ are represented as $l^k_{\phi,j}$ and $s^k_{\phi,j}$ respectively, the relative distance between the Scaphoid and Lunate was calculated as

$$d^k_{\phi,j} = l^k_{\phi,j} - s^k_{\phi,j} \tag{16}$$

$d^k_{\phi,j}$ is a 3×1 vector (X, Y, and Z axis). Then the mean differences $m_{\phi,j}$ and covariance matrix $C_{\phi,j}$ of the jth point pair based on all k samples at position ϕ can be obtained. Equation (16) shows that we calculate and record the distances for each pair of points. Subsequently, the average Mahalanobis distance (MD) between the newly accessed data and the model is calculated using

$$m_\phi = \frac{1}{N}\sum_{j=1}^{N}\sqrt{\left(d^{new}_{\phi,j} - m_{\phi,j}\right)^T C^{-1}_{\phi,j}\left(d^{new}_{\phi,j} - m_{\phi,j}\right)}. \tag{17}$$

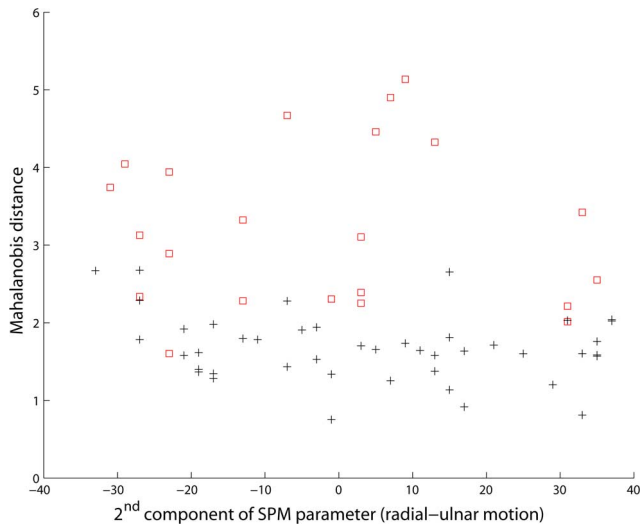


Fig. 9. Mahalanobis distances of 22 subjects at a range of wrist positions for Scaphoid–Lunate measurement. Red squares represent abnormal subjects; black crosses are normal subjects.

2) Wrist Diagnosis: To assess a new wrist, the image is firstly registered to the template volume using the proposed framework. After the SPM parameters are estimated, the measurement of the input image can be compared with the "standard" model at the corresponding wrist positions. Additionally, the mean shape of the SSM is used to represent each of the assessed carpal bones, where the point correspondences between Scaphoid and Lunate were already established. The MD (17) is then calculated to indicate the degree of abnormality of the subject. Seventy-five volumes (three wrist positions from each of the 25 subjects) were registered by the proposed framework. The MDs of 64 automatically and successfully registered volumes are presented in Fig. 9. The 11 unsuccessfully registered volumes (see Section VIII-A) came from three normal subjects and one abnormal subject. The MDs for the remaining 14 successfully registered normal wrists were calculated using leave-one-out experiments (13 subjects were used for "standard" model building). In this figure, the red squares represent abnormal subjects and the black crosses represent normal subjects. The accuracy of classifying individual volumes as abnormal is indicated in the receiver operating characteristic (ROC) curve shown in Fig. 10, obtained by varying the threshold of MD (the same threshold for all positions). The area under curve is 0.94. The MD threshold that produces the best classification is 2.04, which results in a 92% true positive rate (TPR) and 12% false positive rate (FPR) in identifying individual abnormal wrist position.

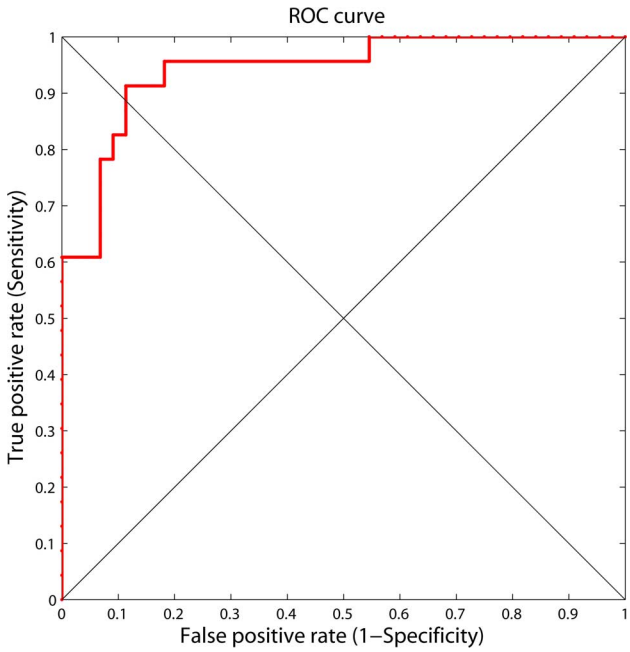


Fig. 10. ROC curve for individual wrist position diagnosis.

IX. Discussion

Statistical pose models of articulated joints in 3-D are potentially highly useful in imaging studies aimed at assessing abnormal kinematics [4], [13], [26]. We have previously described [26] the use of models built in this way for inference of 3-D kinematics based on 2-D image sequences, and demonstrated that this can be achieved with sufficient accuracy to allow meaningful clinical measurements to be made. However, the task of building such a model, requiring accurate segmentation of each bone in a complex joint in several articulated positions, is a daunting one. We argue that the model needs to incorporate the position variation arising from the articulation (kinematics) and the variation in shape of bones between individuals. In this study we have demonstrated a method for building such models, which exploits the facts that identical bones from the same individual are represented in different positions, and that the positions are similar between individuals. The method learns a statistical pose model, while simultaneously generating accurate segmentations. The use of rigid registration integrated with segmentation has allowed us to decouple the issues of pose and shape, as it is only the latter that is relevant for segmentation. Convergence of the iterative framework is assisted by using the evolving model to constrain the registration. This approach has something in common with combined group-wise registration and model building [12], where the registration avoids the selection of an arbitrary template. In this case the number of articulated components adds several degrees of freedom to the problem, and we have based the registration on a single template example. However, we have demonstrated that the combined registration and segmentation is insensitive to the template selection. The variation in pose of the bones at different joint positions results in a requirement for interactive initiation of the registration in the template example. The segmentation step works by having an initial approximate segmentation from the registration of the bones across different positions, which suggests the use of methods developed for semi-interactive segmentation (e.g., Graph-Cut [27], Grow-Cut [1], and Random Walker [28]). We evaluated some such methods; the efficient multi-label propagation of segmentation provided by grow-cut made it ideal for this purpose. The iterative refinement of the grow-cut strength map means that, should further interactive segmentation prove necessary after the final convergence, this can easily be accommodated within the framework. Our observed segmentations were sufficiently good to make this step unnecessary in this case. The segmentation showed high consistency with ground-truth and sub-voxel registration accuracy was achieved.

We demonstrate the effectiveness of the statistical pose and shape model built using our automatic framework by using it to identify scaphoid–lunate dissociation. Scaphoid–lunate dissociation is most apparent in wrist positions during radial–ulnar movement. The model was able to represent the kinematics with sufficient precision to allow the abnormal cases to be identified with high sensitivity and specificity. The use of a SPM for this purpose allows images from different sets of articulated positions to be aligned directly with the model using the SPM values. This confers an advantage over, for example, the method described in [11], where comparisons are made by aligning the wrists at a limited number of predefined positions.

References

[1] V. Vezhnevets and V. Konouchine, "Grow-cut—Interactive multi-label N-D image segmentation," *Proc. Graphicon*, pp. 150–156, 2005.

[2] J. Boisvert, F. Cheriet, X. Pennec, H. Labelle, and N. Ayache, "Geometric variability of the scoliotic spine using statistics on articulated shape models," *IEEE Trans. Med. Imag.*, vol. 27, no. 4, pp. 557–568, Apr. 2008.

[3] A. Rasoulian, R. N. Rohling, and P. Abolmaesumi, "A statistical multi-vertebrae shape+pose model for segmentation of CT images," in *Proc. SPIE Med. Imag.: Image-Guided Procedures, Robotic Intervent., Model.*, Mar. 12, 2013, 86710.

[4] N. Baka, M. de Bruijne, T. van Walsum, B. L. Kaptein, J. E. Giphart, M. Schaap, W. J. Niessen, and B. P. F. Lelieveldt, "Statistical shape model based femur kinematics from biplane fluoroscopy," *IEEE Trans. Med. Imag.*, vol. 31, no. 8, pp. 1573–1583, Aug. 2012.

[5] X. Chen, J. Graham, C. E. Hutchinson, and L. Muir, "Inferring 3-D kinematics of carpal bones from single view fluoroscopic sequences," in *Proc. MICCAI*, 2011, vol. 6892/2011, pp. 680–687.

[6] R. H. Davies, C. Twining, T. F. Cootes, and C. J. Taylor, "Building 3-D statistical shape models by direct optimisation," *IEEE Trans. Med. Imag.*, vol. 29, no. 4, pp. 961–980, Apr. 2010.

[7] T. B. Sebastian, H. Tek, J. J. Crisco, and B. B. Kimia, "Segmentation of carpal bones from CT images using skeletally coupled deformable models," *Med. Image Anal.*, vol. 7, no. 1, pp. 21–45, 2003.

[8] M. Koch, A. G. Schwing, D. Comaniciu, and M. Pollegeys, "Fully automatic segmentation of wrist bones for arthritis patients," in *Proc. Int. Symp. Biomed. Imag.: From Nano to Macro*, 2011, pp. 636–640.

[9] J. Duryea, M. Magalnick, S. Alli, L. Yao, M. Wilson, and R. Goldbach-Mansky, "Semiautomated three-dimensional segmentation software to quantify carpal bone volume changes on wrist CT scans for arthritis assessment," *Med. Phys.*, vol. 35, no. 6, pp. 2321–2330, 2008.

[10] A. Zhang, A. Gertych, and B. Liu, "Automatic bone age assessment for young children from newborn to 7-year-old using carpal bones," *Comput. Med. Imag. Graph.*, vol. 31, no. 4–5, pp. 299–310, 2007.

[11] M. van deGiessen, G. J. Streekstra, S. D. Strackee, M. Maas, K. A. Grimbergen, L. J. van Vliet, and F. M. Vos, "Constrained registration of the wrist joint," *IEEE Trans. Med. Imag.*, vol. 28, no. 12, pp. 1861–1869, Dec. 2009.

[12] T. F. Cootes, C. J. Twining, V. S. Petrovic, K. O. Balalola, and C. J. Taylor, "Computing accurate correspondences across groups of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 1994–2000, Nov. 2010.

[13] M. van deGiessen, M. Fournani, F. M. Vos, S. D. Strackee, M. Maas, L. J. van Vliet, K. Grimbergen, and G. J. Streekstra, "A 4D statistical model of wrist bone motion patterns," *IEEE Trans. Med. Imag.*, vol. 31, no. 3, pp. 613–625, Mar. 2012.

[14] G. E. Marai, D. H. Laidlaw, and J. J. Crisco, "Super-resolution registration using tissue-classified distance fields," *IEEE Trans. Med. Imag.*, vol. 25, no. 2, pp. 177–187, Feb. 2006.

[15] X. Chen, J. Graham, and C. E. Hutchinson, "Integrated framework for simultaneous segmentation and registration of carpal bones," in *Proc. 18th IEEE Int. Conf. Image Process.*, Belgium, 2011, pp. 433–436.

[16] J. J. Craig, *Introduction to Robotics: Mechanics and Control*. Boston, MA: Addison-Wesley, 1989.

[17] G. P. Penney, J. Weese, J. A. Little, P. Desmedt, D. Hill, and D. Hawkes, "A comparison of similarity measures for use in 2-D-3-D medical image registration," in *Proc. MICCAI*, 1998, vol. 1496, pp. 1153–1161.

[18] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C++*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2002.

[19] X. Chen, M. Varley, L. Shark, G. Shentall, and M. Kirby, "A computationally efficient method for automatic registration of orthogonal X-ray images with volumetric CT data," *Phys. Med. Biol.*, vol. 53, pp. 967–983, 2008.

[20] J. P. Thirion, "Image matching as a diffusion process: An analogy with Maxwell's demons," *Med. Image Anal.*, vol. 2, pp. 243–260, 1998.

[21] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: Application to breast MR images," *IEEE Trans. Med. Imag.*, vol. 18, no. 8, pp. 712–721, Aug. 1999.

[22] T. F. Cootes, D. Cooper, C. J. Taylor, and J. Graham, "Active shape models—Their training and application," *Comput. Vis. Image Understand.*, vol. 61, no. 1, pp. 38–59, 1995.

[23] E. B. van de Kraats, G. P. Penney, D. Tomazevic, T. van Walsum, and W. J. Niessen, "Standardized evaluation methodology for 2-d–3-d registration," *IEEE Trans. Med. Imag.*, vol. 24, no. 9, pp. 1177–1189, Sep. 2005.

[24] D. W. Shattuck, G. Prasad, M. Mirza, K. L. Narr, and A. W. Toga, "Online resource for validation of brain segmentation methods," *Neuroimage*, vol. 45, no. 2, pp. 431–439, 2008.

[25] P. Jaccard, "The distribution of the flora in the alpine zone," *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[26] X. Chen, J. Graham, C. Hutchinson, and L. Muir, "Automatic inference and measurement of 3-D carpal bone kinematics from single view fluoroscopic sequences," *IEEE Trans. Med. Imag.*, vol. 32, no. 2, pp. 317–328, Feb. 2013.

[27] Y. Boykov and M. P. Jolly, "Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images," in *Proc. Int. Conf. Comput. Vis.*, 2001, vol. 1, pp. 105–112.

[28] L. Grady, "Random walks for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1768–1783, Nov. 2006.