# Fully automatic cephalometric evaluation using Random Forest regression-voting

Claudia Lindner and Tim F. Cootes

Centre for Imaging Sciences, University of Manchester, UK

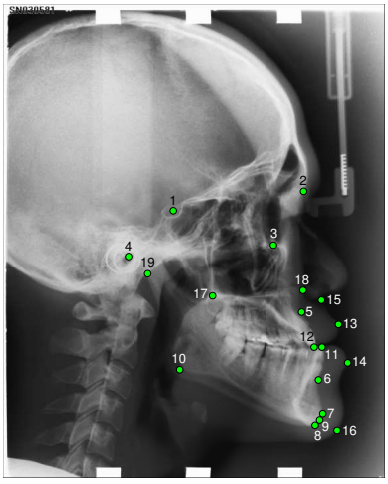
**Abstract.** Cephalometric analysis is commonly used as a standard tool for orthodontic diagnosis and treatment planning. The identification of cephalometric landmarks on images of the skull allows the quantification and classification of anatomical abnormalities. In clinical practice, the landmarks are placed manually which is time-consuming and subjective. This work investigates the application of Random Forest regression-voting to fully automatically detect cephalometric landmarks, and to use the identified positions for automatic cephalometric evaluation. Validation experiments on two sets of 150 images show that we achieve an average mean error of 1.6mm - 1.7mm and a successful detection rate of 75% - 85% for a 2mm precision range, and that the accuracy of our automatic cephalometric evaluation is 77% - 79%. This work shows great promise for application to computer-assisted cephalometric treatment and surgery planning.

## 1 Introduction

Cephalometric radiography is commonly used as a standard tool in orthodontic diagnosis and treatment planning as well as in corrective and plastic surgery planning. Cephalometric evaluation is based on a number of image landmarks on the skull and surrounding soft tissue, which are used for quantitative analysis to assess severity and difficulty of orthodontic cases or to trace facial growth. Figure 1 shows the cephalometric image landmarks used in this work.

Traditionally, these landmarks are placed manually by experienced doctors. This is very time-consuming, taking several minutes for an experienced doctor, and results are inconsistent. To overcome these limitations in clinical practice as well as in the research setting, attempts have been made to automate the landmark annotation procedure [6, 7, 13]. However, due to overlaying structures and inhomogeneous intensity values in radiographic images as well as anatomical differences across subjects, fully automatic landmark detection in cephalograms is challenging. A precision range of 2mm is accepted in the field to evaluate whether a landmark has been detected successfully. A number of outcomes have been reported for this range (e. g. [6] 73%, [7] 61%, [13] 71%) but results are difficult to compare due to the different datasets and landmarks used.

Recently, significant advances in automatically detecting landmarks (i. e. annotating objects) in radiographic images have been made by using machine learning approaches [2, 4, 8, 9]. Some of the most robust and accurate results based on shape model matching have been achieved by using Random Forests (RFs) [1] to




| | |
|---|---|
| L1 | sella |
| L2 | nasion |
| L3 | orbitale |
| L4 | porion |
| L5 | subspinale |
| L6 | supramentale |
| L7 | pogonion |
| L8 | menton |
| L9 | gnathion |
| L10 | gonion |
| L11 | lower incisal incision |
| L12 | upper incisal incision |
| L13 | upper lip |
| L14 | lower lip |
| L15 | subnasale |
| L16 | soft tissue pogonion |
| L17 | posterior nasal spine |
| L18 | anterior nasal spine |
| L19 | articulate |

**Fig. 1.** Cephalogram annotation example showing the 19 landmarks used in this work.

vote for the positions of each individual landmark and then to use a statistical shape model to regularise the votes across all landmarks. In [9], we presented a *fully automatic* landmark detection system (FALDS) based on RF regression-voting in the Constrained Local Model framework to annotate the proximal femur in pelvic radiographs. Here, we investigate the performance of this approach to detect cephalometric landmarks. To be able to compare the performance of our approach to other methodologies, we apply our FALDS as part of the *ISBI 2015 Grand Challenge in automatic Detection and Analysis for Diagnosis in Cephalometric X-ray Images* [12] which aims at automatically detecting cephalometric landmarks and using their positions for automatic cephalometric evaluation of anatomical abnormalities to assist clinical diagnosis.

We show that our FALDS achieves a successful detection rate of up to 85% for the 2mm precision range and an average cephalometric evaluation classification accuracy of up to 79%, performing sufficiently well for computer-assisted planning.

## 2 Methods

We explore the performance of RF regression-voting in the Constrained Local Model framework (RFRV-CLM) [8] to detect the 19 landmarks as shown in Figure 1 on new *unseen* images. In the RFRV-CLM approach, a RF is trained for each landmark to learn to predict the likely position of that landmark. During detection, a statistical shape model [3] is matched to the predictions over all landmark positions to ensure consistency across the set. We apply RFRV-CLM as part of a FALDS [9]: We use our own implementation of Hough Forests [5] to estimate the position, orientation and scale of the object in the image, and use this to initialise the RFRV-CLM landmark detection. The output of the FALDS can then be used for automatic cephalometric evaluation by using the identified landmark positions to calculate a set of dental parameters.

### 2.1 RF regression-voting in the Constrained Local Model framework

Recent work has shown that one of the most effective approaches to detect a set of landmark positions on an object of interest is to train RFs to vote for the likely position of each landmark, then to find the shape model parameters which optimise the total votes over all landmark positions (see [8, 9] for full details):

**Training:** We train the RF regressors (one for every landmark) from a set of images, each of which is annotated with landmarks $\mathbf{x}$ on the object of interest. The region of the image that captures all landmarks of the object is re-sampled into a standardised reference frame. For every landmark in $\mathbf{x}$, we sample patches of size $w_{patch}$ (width = height) and extract features $\mathbf{f}_i(\mathbf{x})$ at a set of random displacements $\mathbf{d}_i$ from the true position in the reference frame. Displacements are drawn from a flat distribution in the range $[-d_{max}, +d_{max}]$. We train a regressor $R(\mathbf{f}(\mathbf{x}))$ to predict the most likely position of the landmark relative to $\mathbf{x}$. Each tree leaf stores the mean offset and the standard deviation of the displacements of all training samples that arrived at that leaf. We use Haar features [11] as they have been found to be effective for a range of applications and can be calculated efficiently from integral images.

A statistical shape model is trained based on landmarks $\mathbf{x}$ in the set of images by applying principal component analysis to the aligned shapes [3]. This yields a linear model of shape variation which represents the position of each landmark $l$ using $\mathbf{x}_l = T_\theta(\bar{\mathbf{x}}_l + \mathbf{P}_l\mathbf{b} + \mathbf{r}_l)$ where $\bar{\mathbf{x}}_l$ is the mean position of the landmark in a suitable reference frame, $\mathbf{P}_l$ is a set of modes of variation, $\mathbf{b}$ are the shape model parameters, $\mathbf{r}_l$ allows small deviations from the model, and $T_\theta$ applies a global transformation (e. g. similarity) with parameters $\theta$.

**Landmark detection:** Given an initial estimate of the pose of the object, the region of interest of the image is re-sampled into the reference frame. We then search an area around each estimated landmark position in the range of $[d_{search}, +d_{search}]$ and extract the relevant feature values at every position. These will be used for the RF regressor to vote for the best position in an accumulator array where every tree will cast independent votes to make predictions on the position of the landmark. The forest prediction is then computed by combining all tree predictions, yielding a 2D histogram of votes $V_l$ for each landmark $l$.

Based on the 2D histograms $V_l$ from the RF regressors, we aim to combine the votes in all histograms given the learned shape constraints via maximising

$$Q(\{\mathbf{b}, \theta\}) = \Sigma_{l=1}^{n} V_l(T_\theta(\bar{\mathbf{x}}_l + \mathbf{P}_l\mathbf{b} + \mathbf{r}_l)). \quad (1)$$

We apply the technique described in [8] to solve this optimisation problem.

### 2.2 Automatic cephalometric evaluation

The 19 landmarks in Figure 1 allow calculation of a number of dental parameters that are used in cephalometric evaluation to classify types of anatomical abnormalities. Table 1 summarises the parameters used in this work. A Python script to automatically calculate the parameters and classify subjects based on a set of 19 landmark positions was provided by the challenge organisers.

**Table 1.** Overview of dental parameters used in the cephalometric evaluation.

| | **ANB**[1] | **SNB**[2] | **SNA**[3] | **ODI**[4] | **APDI**[5] | **FHI**[6] | **FMA**[7] | **MW**[8] |
|---|---|---|---|---|---|---|---|---|
| C1 | 3.2-5.7° | 74.6-78.7° | 79.4-83.2° | 68.4-80.5° | 77.6-85.2° | 0.65-0.75 | 26.8-31.4° | 2-4.5mm |
| C2 | >5.7° | <74.6° | >83.2° | >80.5° | <77.6° | >0.75 | >31.4° | =0mm |
| C3 | <3.2° | >78.7° | <79.4° | <68.4° | >85.2° | <0.65 | <26.8° | <0mm |
| C4 | – | – | – | – | – | – | – | >4.5mm |

[1] ANB: angle between point A (L5), nasion (L2) and point B (L6).
[2] SNB: angle between sella (L1), nasion (L2) and point B (L6).
[3] SNA: angle between sella (L1), nasion (L2) and point A (L5).
[4] Overbite depth indicator (ODI): sum of the angle between the lines from L5 to L6 and from L8 to L10 and the angle between the lines from L3 to L4 and from L17 to L18.
[5] Anteroposterior dysplasia indicator: sum of the angle between the lines from L3 to L4 and from L2 to L7, the angle between the lines from L2 to L7 and from L5 to L6 and the angle between the lines from L3 to L4 and from L17 to L18.
[6] Facial height index: ratio of the posterior face height (distance from L1 to L10) to the anterior face height (distance from L2 to L8).
[7] Frankfurt mandibular angle: angle between the lines from sella (L1) to nasion (L2) and from gonion (L10) to gnathion (L9).
[8] Modified Wits appraisal: $((x_{L12} - x_{L11})/|x_{L12} - x_{L11}|)\|\mathbf{x}_{L12} - \mathbf{x}_{L11}\|$.

## 3 Datasets

The challenge organisers provided two datasets: (a) A training dataset consisting of 150 images as well as ground truth annotations for 19 landmarks as in Figure 1 and ground truth classifications for the dental parameters listed in Table 1 for each image; we will refer to this as the Train1 dataset. (b) A testing dataset consisting of 150 images without ground truth annotations or classifications; we will refer to this as the Test1 dataset. All images were lateral cephalograms acquired from 400 subjects (age range: 6 - 60 years) with Soredex CRANEX©Excel Ceph machine (Tuusula, Finland) and Soredex SorCom software (3.1.5, version 2.0).

The resolution of all images was $1935 \times 2400$ pixels with a pixel spacing of 0.1mm. The ground truth annotations were the average of two sets of manual annotations for each training image, annotated by two experienced medical doctors. Table 2 shows the intra- and inter-observer variability of the ground truth annotations.

**Table 2.** Intra- and inter-observer errors of manual ground truth annotations for the training dataset: Mean radial error ± standard deviation in mm (as defined in Section 4).

| **Intra-observer variability** | | **Inter-observer variability** |
|---|---|---|
| Doctor1 | Doctor2 | Doctor1 vs Doctor2 |
| 1.73 ± 1.35 | 0.90 ± 0.89 | 1.38 ± 1.55 |

## 4 Experiments and evaluation

We conducted a series of experiments to evaluate the performance of the RFRV-CLM approach for automatic cephalometric analysis. Cross-validation experiments on the Train1 dataset were used to optimise the parameters for the detection

of cephalometric landmarks. The fully automatically detected landmark positions were then used to analyse their ability for automatic classification of anatomical abnormalities. We applied the optimised FALDS to the Test1 dataset and report the landmark detection and classification results provided by the challenge organisers.

All annotation results are reported as the mean radial error (MRE, in mm), obtained after rounding all landmark positions to the nearest whole pixel position and defined by $MRE = (\Sigma_{i=1}^{n} R_i)/n$ with $n$ being the number of images and $R = \|\mathbf{x}_{L_m} - \mathbf{x}_{L_a}\|$ for manually and automatically marked landmarks $L_m$ and $L_a$, respectively. We also report the successful detection rate (SDR, in %) which gives the percentage of landmarks detected within a certain precision range $z \in \{2.0mm, 2.5mm, 3.0mm, 4.0mm\}$: $SDR = \#\{i : (R_i \leq z)\}/n \times 100$ for $i \leq 1 \leq n$. For evaluation of the automatically obtained cephalometric classification results, we report the successful classification rate (SCR, in %) which gives the percentage of accurately classified images per dental parameter: $SCR = \#\{i : C_{m_i} = C_{a_i}\}/n \times 100$ with manually and automatically obtained classification types $C_{m_i}$ and $C_{a_i}$, respectively, and $i \leq 1 \leq n$.

### 4.1 Parameter optimisation via cross-validation experiments

In [10], it was shown that the RFRV-CLM approach generalises well across application areas. To investigate whether structure specific improvements can be achieved, we conducted a series of two-fold cross-validation experiments to optimise the parameters for cephalometric landmark detection. Here, we summarise the results. All cross-validation experiments were conducted on the Train1 dataset by randomly splitting the dataset into two disjoint sets of 75 images each, training on one set and testing on the other and then doing the same with the datasets switched. Results reported are the average of the two runs.

We train a FALDS following a coarse-to-fine approach (10+1 search iterations) and using the baseline parameters as suggested in [9]. We found that we can improve upon results using previously reported parameters by making the following changes for the second-stage, fine model: increase the patch size $w_{patch}$ from 20 to 30, increase the sampling range $d_{max}$ from 15 to 30 and halve the search range $d_{search}$. Increasing the parameters even further did not lead to any improvements. Figure 2 gives the results. These show that our FALDS was able to detect all landmarks across all 150 images with an average MRE of 1.6mm ($\pm$0.4mm standard deviation), and successful detection rates of 85%/98%/100% for the 2.0mm/2.5mm/3.0mm precision ranges, respectively.

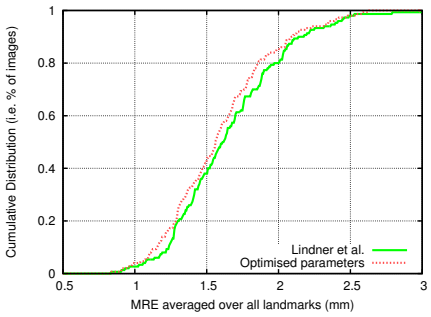



**Fig. 2.** Fully automatic landmark detection results comparing the parameters suggested in [9] to the optimised parameters proposed in this work.

### 4.2 Landmark detection results on the Test1 dataset

We used the optimised parameters to train a FALDS using all 150 images from the Train1 dataset for training, and applied the system to the 150 images in the Test1 dataset. The average runtime of our system to detect all 19 landmarks was less than 5s per image. Table 3 shows the landmark detection results. We also experimented with adding auxiliary landmarks to improve the landmark detection performance but did not find this to make a significant difference. When adding 10 additional landmarks along the jawline between landmarks L8 and L19 we achieved the following average MRE and SDR (2.0/2.5/3.0/4.0mm) values on the Test1 dataset: 1.67±1.48mm, 73.68% /80.21%/ 85.19%/ 91.47%.

**Table 3.** Landmark detection results on the Test1 dataset: Mean radial error (MRE ± standard deviation) and successful detection rates (SDR) for 2.0mm, 2.5mm, 3.0mm and 4.0mm precision ranges.

| **Landmark** | **MRE (mm)** | **SDR (%)** | | | |
|---|---|---|---|---|---|
| | | **2.0mm** | **2.5mm** | **3.0mm** | **4.0mm** |
| sella (L1) | 0.75 ± 0.95 | 97.33 | 97.33 | 97.33 | 98.00 |
| nasion (L2) | 1.71 ± 2.01 | 71.33 | 76.00 | 77.33 | 90.00 |
| orbitale (L3) | 1.73 ± 1.34 | 66.67 | 77.33 | 85.33 | 94.00 |
| porion (L4) | 3.05 ± 2.74 | 52.67 | 58.00 | 62.00 | 67.33 |
| subspinale (L5) | 2.31 ± 1.60 | 54.00 | 62.00 | 73.33 | 86.00 |
| supramentale (L6) | 1.61 ± 1.45 | 70.00 | 82.00 | 86.67 | 91.33 |
| pogonion (L7) | 1.07 ± 0.84 | 85.33 | 92.67 | 96.00 | 100.00 |
| menton (L8) | 1.02 ± 0.82 | 91.33 | 96.67 | 98.00 | 99.33 |
| gnathion (L9) | 0.83 ± 0.68 | 92.67 | 98.00 | 98.00 | 99.33 |
| gonion (L10) | 4.58 ± 3.12 | 23.33 | 28.67 | 36.00 | 48.00 |
| lower incisal incision (L11) | 0.92 ± 1.10 | 89.33 | 90.67 | 96.00 | 96.67 |
| upper incisal incision (L12) | 0.61 ± 0.90 | 97.33 | 98.00 | 98.00 | 99.33 |
| upper lip (L13) | 1.33 ± 3.90 | 90.67 | 94.00 | 98.67 | 98.67 |
| lower lip (L14) | 0.93 ± 0.91 | 92.67 | 98.00 | 98.00 | 99.33 |
| subnasale (L15) | 1.14 ± 1.02 | 86.00 | 90.00 | 93.33 | 97.33 |
| soft tissue pogonion (L16) | 2.22 ± 1.91 | 52.67 | 60.67 | 76.67 | 88.67 |
| posterior nasal spine (L17) | 1.01 ± 0.81 | 90.00 | 92.67 | 98.00 | 99.33 |
| anterior nasal spine (L18) | 1.84 ± 1.99 | 70.00 | 78.67 | 81.33 | 86.67 |
| articulate (L19) | 3.14 ± 3.31 | 50.67 | 54.00 | 56.67 | 64.67 |
| **AVERAGE** | **1.67 ± 1.65** | **74.95** | **80.28** | **84.56** | **89.68** |

The results show that the average MRE over all images and landmarks on the Test1 dataset is very similar to the results we obtained in our two-fold cross-validation experiments. However, for the Test1 dataset, the standard deviation for the error values is significantly higher and the obtained successful detection rates for the various precision ranges are lower. This suggests that Test1 might be a more "challenging" dataset or that the shape and appearance variation experienced in the Test1 dataset is not well represented in the Train1 dataset. An alternative explanation for these deviations in performance would be an inconsistency in the manual annotations between the Train1 and Test1 datasets.

Landmark L10 has a significantly lower detection rate than all other landmarks. A likely reason for this is that the left and right mandibular halves are not always well aligned in cephalograms which only provide a 2D projection of a 3D shape (see Figure 1 for examples). Qualitative analysis suggests that our FALDS sometimes annotates the "wrong" edge causing high error values if both mandibular halves appear far apart in an image. We believe that the performance on L10 could be improved upon by training the system on a larger representative dataset.

### 4.3 Cephalometric classification results

The automatically detected landmark positions can be used to calculate dental parameters for cephalometric evaluation as in Table 1. The successful classification rates (SCR) of comparing the classifications obtained from the automatically detected landmark positions with the ground truth classifications are shown in Table 4. We report the results for the classifications obtained from both the landmark detections for the cross-validation experiments on the Train1 dataset and the landmark detections on the Test1 dataset. The results show consistent high accuracy (on average 79%/77% for Train1/Test1) in the automatic evaluation of cephalometric abnormalities. We would like to point out that this is a rather challenging classification task considering the small ranges for the different classes. For example, the three classes for ANB are separated by less than 3°.

**Table 4.** Successful classification rates (SCR) for eight dental parameters.

| | **SCR (%)** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **ANB** | **SNB** | **SNA** | **ODI** | **APDI** | **FHI** | **FMA** | **MW** |
| Train1* | 76.0 | 91.3 | 70.0 | 74.0 | 76.0 | 82.7 | 82.0 | 82.0 |
| Test1 | 71.3 | 83.3 | 60.0 | 80.0 | 83.3 | 77.3 | 81.3 | 85.3 |

* These are two-fold cross-validation results based on randomly splitting the Train1 dataset into two disjoint subsets of equal size.

## 5 Discussion and conclusions

We have investigated the performance of RFRV-CLM as part of a FALDS [9] for the automatic detection of cephalometric landmarks and subsequent automatic cephalometric evaluation. We optimised the FALDS parameters to account for the isolated landmark positions in cephalograms, and show that the improved system achieves a successful landmark detection rate of 75% to 85% for the 2mm precision range when detecting 19 landmarks in two datasets of 150 images each – in less than 5s per image. On the *ISBI 2015 Grand Challenge* [12] Test1 dataset, this is an improvement from 73% to 75% when compared to the best performance of a similar *ISBI 2014 Grand Challenge* [6], though the results are not directly comparable as the training and testing datasets have been increased. We have also shown that the automatically detected landmarks can be used for automatic cephalometric evaluation, achieving a classification accuracy of 77% to 79%.

Comparing our landmark detection results with the intra- and inter-observer errors of the manual ground truth annotations as shown in Table 2 reveals that our fully automatic approach achieves a similar performance as the manual annotations. The performance of our FALDS is likely to improve if more accurate ground truth annotations were available. Moreover, we would like to point out that the datasets used in this work included subjects with an age range from 6 to 60 years which introduces significant variation in shape and appearance of the skull. We believe that the performance of our cephalometric FALDS could be further improved if more representative training data was available.

Given its high accuracy and low runtime, our FALDS shows great promise for application to computer-assisted cephalometric treatment and surgery planning.

**Acknowledgements.** C. Lindner is funded by the Engineering and Physical Sciences Research Council, UK (EP/M012611/1).

## References

1. Breiman, L.: Random Forests. Machine Learning 45, 5–32 (2001)
2. Chen, C., Xie, W., Franke, J., Grutzner, P., Nolte, L.P., Zheng, G.: Automatic X-ray landmark detection and shape segmentation via data-driven joint estimation of image displacements. Medical Image Analysis 18(3), 487–499 (2014)
3. Cootes, T., Taylor, C., Cooper, D., Graham, J.: Active Shape Models - Their training and application. Computer Vision and Image Understanding 61(1), 38–59 (1995)
4. Donner, R., Menze, B., Bischof, H., Langs, G.: Fast anatomical structure localization using top-down image patch regression. In: Proceedings MICCAI 2012 - Workshop MCV. Lecture Notes in Computer Science, vol. 7766, pp. 133–141. Springer (2013)
5. Gall, J., Lempitsky, V.: Class-specific Hough forests for object detection. In: Proceedings CVPR 2009. pp. 1022–1029. IEEE Press (2009)
6. Ibragimov, B., Likar, B., Pernus, F., Vrtovec, T.: Automatic cephalometric X-ray landmark detection by applying game theory and random forests. In: Proceedings ISBI 2014 Grand Challenge: Automatic Cephalometric X-Ray Landmark Detection Challenge. pp. 1–8 (2014)
7. Kafieh, R., Sadri, S., Mehri, A., Raji, H.: Discrimination of bony structures in cephalograms for automatic landmark detection. In: Advances in Computer Science and Engineering, vol. 6, pp. 609–620. Springer (2009)
8. Lindner, C., Bromiley, P., Ionita, M., Cootes, T.: Robust and Accurate Shape Model Matching using Random Forest Regression-Voting. IEEE TPAMI (2014), http://dx.doi.org/10.1109/TPAMI.2014.2382106
9. Lindner, C., Thiagarajah, S., Wilkinson, M., The arcOGEN Consortium, Wallis, G., Cootes, T.: Fully Automatic Segmentation of the Proximal Femur Using Random Forest Regression Voting. IEEE TMI 32(8), 1462–1472 (2013)
10. Lindner, C., Thiagarajah, S., Wilkinson, M., The arcOGEN Consortium, Wallis, G., Cootes, T.: Accurate Bone Segmentation in 2D Radiographs Using Fully Automatic Shape Model Matching Based On Regression-Voting. In: Proceedings MICCAI 2013. LNCS, vol. 8150, pp. 181–189. Springer (2013)
11. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings CVPR 2001. pp. 511–518. IEEE Press (2001)
12. Wang, C., Huang, C., Li, C., Chang, S.: ISBI 2015 Grand Challenge: Automated detection and analysis for diagnosis in cephalometric x-ray image. http://www-o.ntust.edu.tw/ cweiwang/ISBI2015/challenge1/index.html (2015), accessed: 10/02/2015
13. Yue, W., Yin, D., Li, C., Wang, G., Xu, T.: Automated 2-D cephalometric analysis on x-ray images by a model-based approach. IEEE TBME 53(8), 1615–1623 (2006)