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 Density (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.

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 osteopenic 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—2 pixels). 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 $\chi^2$ 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 $\theta_i$ at point $i$, reflecting it in a notional symmetry axis to obtain $\theta'_i$, and the actual angle at the corresponding point on the other side $\theta''_i$. 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:

$$M_S = \log(1 - F(\frac{z}{\sqrt{2}}))$$

where $F(x) = 0.5(1 + \text{erf}(x))$, and $I$ contains the indices of a subset of 20 (equispaced) points between MF and AG. The mean and standard deviation $\mu_S$, $\sigma_S$ are derived from the training set. Cut-off thresholds 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 $\mu$ is the median, and $\sigma$ 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 comparison 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 predominantly 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 osteopertics, 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\textsuperscript{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.

Table 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

![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.](image-url)
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 $\chi^2$ distribution (6.63) for femoral neck diagnoses, and when comparing osteoporosis at any site the lowest McNemar test statistic was 6.26 (for femoral neck diagnoses, and when comparing osteoporosis excess of the 99% point of the McNemar test [10] (Table 4) All test statistics were far in excess of the 99% point of the $\chi^2$ distribution (6.63) for femoral neck diagnoses, and when comparing osteoporosis at any site the lowest McNemar test statistic was 6.26 (for femoral neck diagnoses, and when comparing osteoporosis excess of the 99% point of the McNemar test [10] (Table 4) All test statistics were far in excess of the 99% point of the $\chi^2$ distribution (6.63) for femoral neck diagnoses, and when comparing osteoporosis at any site the lowest McNemar test statistic was 6.26 (for femoral neck diagnoses, and when comparing osteoporosis excess of the 99% point of the McNemar test [10] (Table 4) All test statistics were far in excess of the 99% point of the $\chi^2$ distribution (6.63) for femoral neck diagnoses, and when comparing osteoporosis at any site the lowest McNemar test statistic was 6.26 (for femoral neck diagnoses, and when comparing osteoporosis excess of the 99% point of the McNemar test [10] (Table 4) All test statistics were far in excess of the 99% point of the $\chi^2$ distribution (6.63) for femoral neck diagnoses, and when comparing osteoporosis at any site the lowest McNemar test statistic was 6.26 (for femoral neck diagnoses, and when comparing osteoporosis excess of the 99% point of the McNemar test [10] (Table 4) All test statistics were far in excess of the 99% point of the $\chi^2$ distribution (6.63) for femoral neck diagnoses, and when comparing osteoporosis at any site the lowest McNemar test statistic was 6.26 (for femoral neck diagnoses, and when comparing osteoporosis excess of the 99% point of the McNemar test [10] (Table 4) All test statistics were far in excess of the 99% point of the $\chi^2$ distribution (6.63) for femoral neck diagnoses, and when comparing osteoporosis at any site the lowest McNemar test statistic was 6.26 (for femoral neck diagnoses, and when comparing osteoporosis excess of the 99% point of the McNemar test [10] (Table 4) All test statistics were far in excess of the 99% point of the $\chi^2$ distribution (6.63) for femoral neck diagnoses, and when comparing osteoporosis at any site the lowest McNemar test statistic was 6.26 (for femoral neck diagnoses, and when comparing osteoporosis excess of the 99% point of the McNemar test [10] (Table 4) All test statistics were far in excess of the 99% point of the $\chi^2$ distribution (6.63) for femoral neck diagnoses, and when comparing osteoporosis at any site the lowest McNemar test statistic was 6.26 (for femoral neck diagnoses, and when comparing osteoporosis excess of the 99% point of the McNemar test [10] (Table 4) All test statistics were far in excess of the 99% point of the $\chi^2$ distribution (6.63) for femoral neck diagnoses, and when comparing osteoporosis at any site the lowest McNemar test statistic was 6.26 (for femoral neck diagnoses, and when comparing osteoporosis excess of the 99% point of the McNemar test [10] (Table 4) All test statistics were far in excess of the 99% point of the $\chi^2$ distribution (6.63) for femoral neck diagnoses, and when comparing osteoporosis at any site the lowest McNemar test statistic was 6.26 (for femoral neck diagnoses, and when comparing osteoporosis excess of the 99% point of the McNemar test [10] (Table 4) All test statistics were far in excess of the 99% point of the $\chi^2$ distribution (6.63) for femoral neck diagnoses, and when comparing osteoporosis at any site the lowest McNemar test statistic was 6.26 (for femoral neck diagnoses, and when comparing osteoporosis excess of the 99% point of the McNemar test [10] (Table 4) All test statistics were far in excess of the 99% point of the $\chi^2$ distribution (6.63) for femora...