European Journal of Orthodontics 22 (2000) 499–508

An evaluation of active shape models for the automatic identification of cephalometric landmarks

Tim J. Hutton, Sue Cunningham and Peter Hammond
Eastman Dental Institute, UCL, London, UK

SUMMARY This paper describes an evaluation of the application of active shape models to cephalometric landmarking. Permissible deformations of a template were established from a training set of hand-annotated images and the resulting model was used to fit to unseen images. An evaluation of this technique in comparison to the accuracy achieved by previous methods is presented.

Sixty-three randomly selected cephalograms were tested using a drop-one-out method. On average, 13 per cent of 16 landmarks were within 1 mm, 35 per cent within 2 mm, and 74 per cent within 5 mm. It was concluded that the current implementation does not give sufficient accuracy for completely automated landmarking, but could be used as a time-saving tool to provide a first-estimate location of the landmarks. The method is also of interest because it provides a framework for a range of future improvements.

Introduction

Cephalometry was first introduced by Broadbent (1931) and subsequently revolutionized the analysis of malocclusion and the underlying skeletal structures. Lateral cephalograms may be traced manually, but more recently computers have been used. It is widely acknowledged that both manual and computerized methods are time-consuming and error prone (Baumrind and Frantz, 1971; Broch *et al*., 1981). In addition, the process is open to considerable subjectivity. The associated errors were classified by Baumrind and Frantz (1971) as mechanical, projective, or due to problems in landmark identification. The greatest error lies in landmark identification and this, in turn, is related to the quality of the radiograph. Richardson (1981) compared digitization using an electro-mechanical device and manual tracing, and found the former to be slightly more reproducible.

There have been previous attempts to automate cephalometric analysis with the aims of reducing the time required to obtain an analysis, improving the accuracy of landmark identification, and reducing the errors due to clinician subjectivity. The bid to establish such a system has become topical once more with the move towards digital imaging and the development of 'film-free' hospitals. When on-screen digital images replace physical radiographs, as is already occurring, it will no longer be possible to manually trace cephalograms and landmark annotation using a computer will become standard. A system that can automatically identify landmarks and produce the required measurements would be of immense benefit.

Some of the earliest work on automating cephalometric analysis was published by Lévy-Mandel *et al*. (1986). They tracked edges in the image to locate landmarks on structures with well-defined outlines, such as the lower border of the mandible. *A priori* knowledge of the typical shape of the important edges was encoded in algorithms that followed the boundaries of different structures. Each algorithm was designed to find a specific structure, an approach termed 'hand-crafted algorithms'.

Their system was tested on two high-quality cephalograms scanned at 256×256 pixels with 256 grey levels. Only landmarks that lay on or near to edges in the image could be located. They

reported that 23 out of 36 landmarks could be determined on a good quality image, but gave no evaluation of the landmark accuracy as compared with an 'expert'.

Parthasarathy *et al*. (1989) presented a similar scheme, again using hand-crafted algorithms. They used 480 × 512 pixel images and created a four-level image pyramid for improving the efficiency of their search. No indication of the number of grey levels was given. Their system was tested on five cephalograms of varying quality. They compared the accuracy of their system to the landmarks as placed by two experts. Of nine landmarks, on average 18 per cent were located to within 1 mm, 58 per cent within 2 mm, and 100 per cent within 5 mm. It has been suggested that an error of 2 mm in landmark placement is acceptable (Rakosi, 1982), but that an accuracy of 1 mm is desirable (Forsyth *et al*., 1996).

Tong *et al*. (1990) presented an extension to the work of Parthasarathy *et al*. (1989). Their system looked for 17 landmarks, not including any of the seven of the earlier work. They indicated that their results could be combined to yield a full cephalometric analysis. If both systems were used together, as was intended, on the five cephalograms on average 40 per cent of 26 landmarks would be located to within 1 mm, 70 per cent to within 2 mm, and 95 per cent to within 5 mm.

Davis and Taylor (1991) described how hand-crafted algorithms could be integrated in a blackboard architecture, allowing back-tracking in the face of contradictions. A formal evaluation by Forsyth and Davis (1996) used 10 cephalograms scanned at 512 × 512 pixels with 64 grey levels. It excluded radiographs where unerupted teeth were overlying the apices of the incisors. On average, 63 per cent of 19 landmarks were located to within 1 mm and 74 per cent to within 2 mm.

In all these earlier works it was not stated whether the hand-crafting of the algorithms was performed with the testing images unseen. Indeed, it was suggested by Rudolph *et al*. (1998) that the algorithms were tested on the same images that were used to create them. To evaluate an image understanding algorithm it must be shown that it will perform acceptably on new images, not just those that have been used for designing the method. Without demonstrating that an approach can be generalized from the training set to a test set, a study is of little use. This flaw in scientific methodology means that direct comparison of the results of this investigation with those of the studies mentioned is perhaps inadvisable.

Each of the studies above used some algorithmic encoding of anatomical knowledge to locate landmarks on or near to edges in the image. While this approach gives good results for landmarks that are well-defined by strong edges, it tends to be unreliable in the face of image artefacts, variable image quality, and structure changes such as malocclusions. It is not clear how such systems can be made more robust to the presence of this variation without implementing ever more complex hand-crafted rules, together with error-checking and back-tracking mechanisms. The approach is intrinsically flawed with respect to these variations.

An alternative scheme, suggested by Cardillo and Sid-Ahmed (1994), used sub-image matching on hand-selected features from a training set of 40 images. Their images were scanned at 512 × 480 pixels with 256 grey levels. Seventy-six per cent of their 20 landmarks were located to within 2 mm. They did not specify how many were within 1 mm.

Rudolph *et al*. (1998) described the use of spatial spectroscopy to characterize the grey-level appearance around each landmark from a training set of hand-landmarked images. They used a 'drop-one-out' scheme to enable testing of 14 images using the other 13 as the training set. However, they used images that were just 64 × 64 pixels to make their scheme computationally feasible. They compared their system's performance with that of an expert using images of the same resolution. They reported that no statistical difference could be found between the manual error and the error of their automated system. For comparison, this implied that 100 per cent of the landmarks were located to within 4 mm. They suggested that, as the resolution increased, the landmarks would correspondingly become more accurate, but this remains to be shown.

Most recently, Chen *et al*. (1999) used neural networks together with genetic algorithms to search for sub-images that contained each of the cephalometric landmarks. However, they did not report on the accuracy of the landmark placement.

These three later methods used essentially the same approach: generating a model of the grey-levels around each point from a training set and then matching this model to a new image to locate the points of interest. While this approach does not depend on strong, continuous edges, it does rely on the image appearance around each landmark being reproducible across all cephalograms seen. Around the upper incisor tip, for example, the grey-level appearance may vary greatly between different malocclusions.

Active shape models (ASMs) were described in Cootes *et al*. (1995) and first reported by Hill *et al*. (1992). They use a model of the spatial relationships between the important structures, a template, to help search the image for features of interest. The key innovation is that the variation in shape is modelled, enabling the synthesis of plausible new examples of the structures seen. By matching the deformable template to the structure in a particular image not only can the landmarks of interest be located, but a complete atlas of all features can also be given. The approach has been used to tackle other image processing problems within the field of dentistry (Hutton *et al*., 1999).

The ASM approach uses both a local model of grey-level appearance and a global model of the spatial relationship between the points that define the target. The global model is not hand-crafted as in the earlier studies, but is learned directly from the training set. The grey-level model is statistically derived from the images in the training set, as with Cardillo and Sid-Ahmed (1994), Rudolph *et al*. (1998), and Chen *et al*. (1999).

The aim of this study was to evaluate the accuracy of the ASM approach as applied to the automatic location of cephalometric landmarks.

Materials and methods

Experimental design

The method for selecting the cephalograms to be tested is of prime importance if an evaluation is to be relevant for clinical use. Few of the early studies have borne this in mind, some deliberately used only radiographs that were judged to be of high quality. Ideally, the cephalograms to be tested should be randomly selected from the files of a typical radiology department. It is desirable also that they should be made accessible for later comparison by other researchers. For this study, both of these criteria have been fulfilled.

Sixty-eight pre-treatment cephalograms were randomly selected from the notes of patients currently undergoing orthodontic treatment within the department. Radiographs were used regardless of quality and included those from a range of ages, racial groups, and malocclusions (Table 1).

The cephalograms were scanned using a Microtek ScanMaker 4 flatbed scanner (Microtek Lab, Inc., Redondo Beach, CA, USA) at 100 dpi (1 pixel = 0.25 mm). This yielded images that were approximately 750 × 950 pixels with

Table 1 Distribution of ages, norm groups and dentitions in the training set.

Mean age (range)	Norm group			Incisor relationship				Dentition[1]	
	Caucasian	Afro-Caribbean/ Black African	Oriental	I	II/1	II/2	III	Mixed	Permanent
15.3 years (9–43 years)	59	6	3	12	32	7	17	11	57

[1]Mixed dentition = deciduous teeth still present.
Permanent dentition = no deciduous teeth retained.
NB: permanent dentition included some patients with unerupted canines.

256 grey levels, potentially capturing all the information present in the original radiographs (Macrì and Wenzel, 1993; Forsyth *et al*., 1996).

Five of the images were not of adequate quality for tracing, either manually or automatically, and these were rejected. The remaining 63 cephalograms were landmarked on computer by an experienced orthodontist (S.C.) to provide a 'gold standard' for evaluating the performance of the software.

For testing the accuracy of the algorithm, a drop-one-out scheme was used with the training set (Rudolph *et al*., 1998). Each example was removed from the training set and the model recomputed before testing. Thus, the model was based on 62 different tracings each time. This method of testing is statistically acceptable, makes maximum use of the training set, and provides a maximum number of tested images. The accuracy of a cephalometric landmark was measured using its distance from the corresponding landmark in the 'gold standard' tracing.

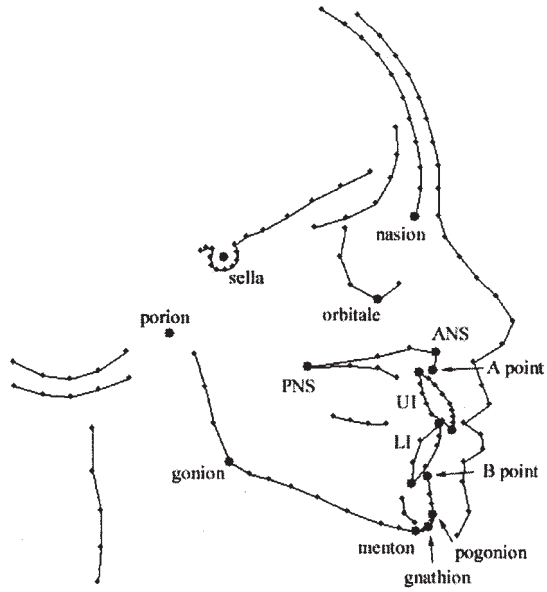

Figure 1 The cephalometric tracing used for the shape template.

Building the model

The ASM required a training set of hand-annotated images. The annotation took the form of a set of points joined with line segments, giving an outline of the major structures in the image. The design of this template is fairly intuitive in that structures such as the lower border of the mandible are represented by a line. The exact configuration of the model is not crucial for the algorithm to work, as long as all the important features are present. For this study the template shown in Figure 1 was used. It comprised recognized cephalometric landmarks plus a set of subsidiary landmarks evenly spaced along the strong edges that regularly appear in the images. In total, 137 points were used to define the model. The landmarks used were sella, nasion, porion, orbitale, ANS, PNS, point A, point B, pogonion, gnathion, menton, gonion, upper and lower incisor tips, and root apices. Other landmarks could be added without difficulty.

After manual annotation by the expert, each image in the training set was overlaid by the template of Figure 1 with all the points aligned with their corresponding structure in the image. The co-ordinates for the points that defined each template were analysed by applying a principal components analysis (PCA). This standard statistical technique is useful for reducing the high dimensionality of the input into a much smaller set of parameters by exploiting the correlations that naturally exist in the data.

The first step was to ensure that the templates were all aligned with each other. A standard method of doing this is the Procrustes algorithm (Goodall, 1991), which iteratively computes the mean shape and aligns all the examples to it until convergence. After alignment, the PCA yielded a set of 'modes of variation', single parameters that each represent a correlated movement of all the points in the model: a 'deformation'. By applying these deformations to the mean shape it was possible to synthesize new examples of plausible cephalometric tracings. Each mode had an associated proportion, the first few modes combined explaining a sizeable percentage of the variation that was seen. By drawing only on the first few modes all the examples in the training set could be closely approximated.

Landmarking on unseen images

In order to find the landmarks in a new image, the mean template was first overlaid onto the image. The method described by Cootes *et al.* (1995) was used, a local search at each point in the model determines whether there is a nearby place in the image that matches to the intensity profile model learned from the training set. Each point thus contributes a 'pull vector' from its current position to a nearby point in the image (Figure 2). For example, if the majority of the pull vectors are pointing to the left then the template will provide a better fit if it is translated to the left. Likewise the scaling and rotation of the template can readily be corrected.

The deformation required to produce a better fit was computed by correlating the pull vectors against the modes produced by the PCA. A key feature of the approach is that the deformation is strictly limited to be within a similar range as seen in the training set. In this way the template remains a plausible cephalometric configuration and is less likely to be 'thrown off' by confusing features in the image. This process is iterated, the template moves and deforms to give a best match to that found in the image. When a convergence criterion is reached, the process stops and the final configuration of the template is the cephalometric tracing. The fitting process is summarized in Figure 3.

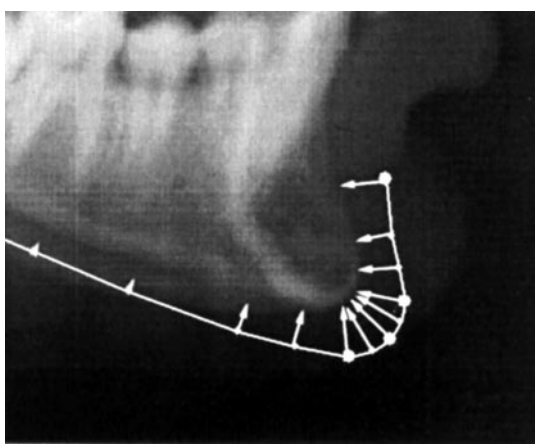

Figure 2 The template uses a local search in the image to determine the pull vectors at each point.

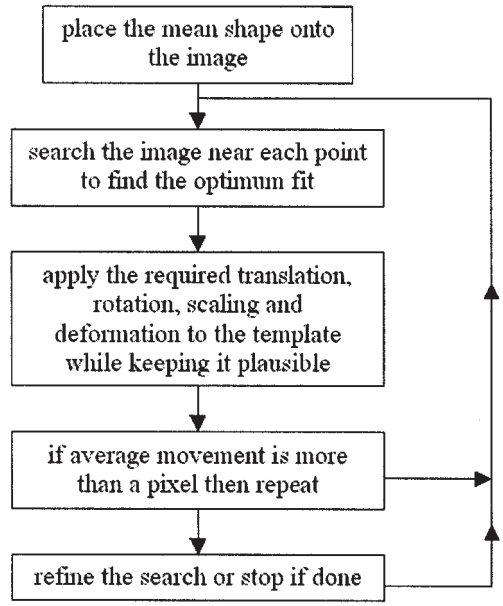

Figure 3 Flow diagram of the active shape model fitting procedure.

Results

Figure 4 shows the mean shape produced by averaging the co-ordinates of the templates after

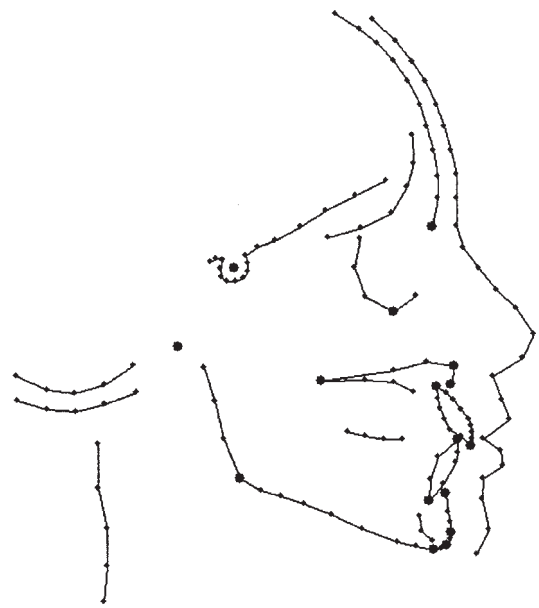

Figure 4 The mean shape.

all 63 examples had been aligned. It should be noted that this corresponds with the informal notion of the average person. To the authors' knowledge this is the first time that shape analysis techniques have been used to compute the mean for an entire cephalometric tracing, not just the angles and measurements. Singh *et al*. (1997) reported the results of a shape analysis of a lateral head radiograph, but only for the cranial base. Of course, this average is just for the 63 examples in the training set, but the method could be used to compute means over larger populations.

Figure 5 shows the modes of variation that were produced by the PCA analysis of the training set. Each row illustrates the mean shape deformed by one of the modes. The first mode explains 20.4 per cent of the variation seen across the 63 examples. Together the first three modes account for 45.3 per cent, nearly half of the variation that was seen.

Modes 1 and 2 captured antero-posterior and vertical variations in shape, whilst mode 3 captured mainly antero-posterior discrepancies. To account for 95 per cent of the variation, 29 modes were required.

Figure 6 shows the fitting algorithm on an unseen image. The soft tissue outline was not used for the fitting process in the present study. The final configuration of the template was a reasonably satisfactory cephalometric tracing for this image.

For comparison with previous studies the percentage of the landmarks that were located to within various radii is reported: 1, 2, and 5 mm.

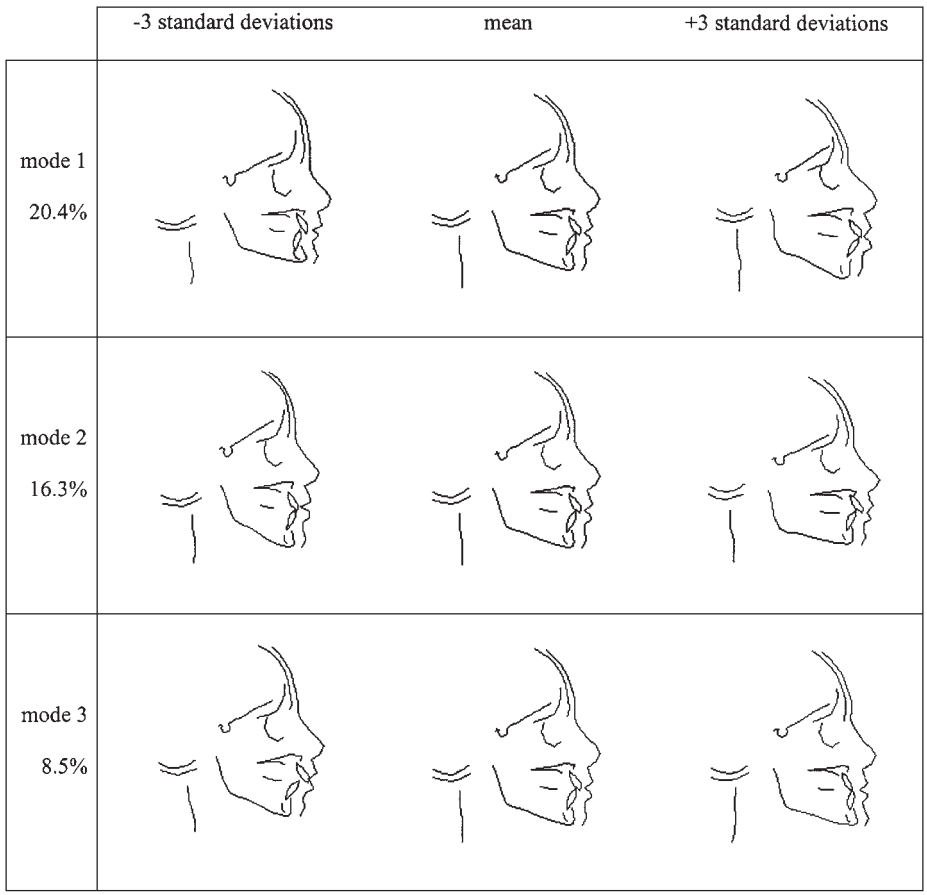

Figure 5 Modes of variation.

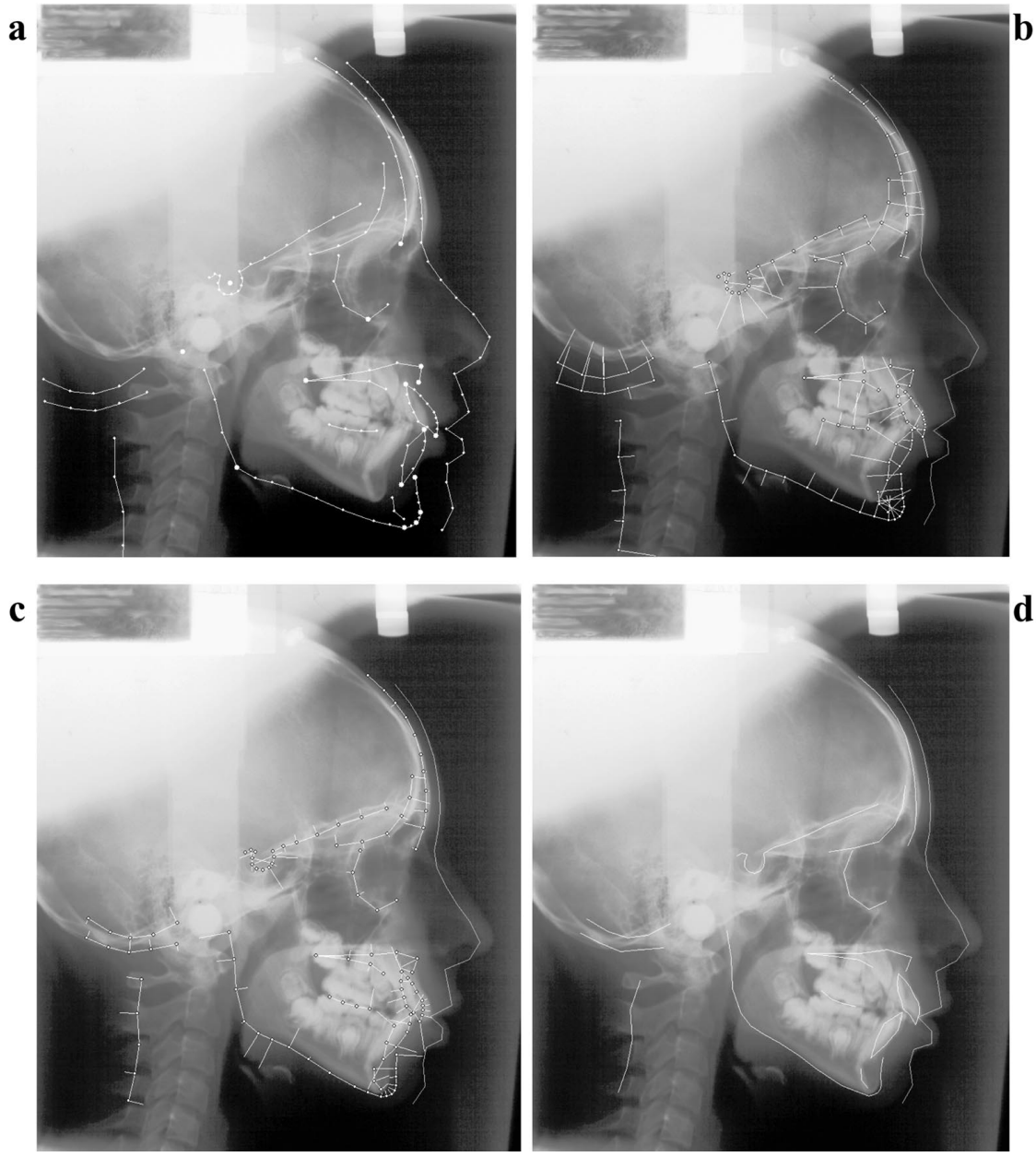

Figure 6 The fit in progress, showing the pull vectors. The soft tissue outline is not used for the fit. (a) Initial placement; (b) during the fit; (c) later in the fit; and (d) final configuration.

On the image above, eight of the 16 landmarks were correct to within 2 mm and two to within 1 mm. All the landmarks except porion were correct to within 5 mm.

When the fit was repeated on all the images, it was found that, on average, 13 per cent of the landmarks were within 1 mm, 35 per cent within 2 mm and 74 per cent within 5 mm. Table 2 shows the errors in millimetres of each of the landmarks over the training set.

The fit for each image took on average 80 seconds, with a range between 30 seconds and

Table 2 Landmark errors across the tested images.

	Average (mm)	Min (mm)	Max (mm)	SD (mm)
Porion	7.3	0.71	47.1	6.5
Sella	5.5	0.26	43.7	6.8
Nasion	5.6	0.57	23.3	3.9
Orbitale	5.5	0.82	21.3	3.4
UI root	2.9	0.40	15.7	2.6
UI tip	2.9	0.36	22.4	3.8
Gonion	5.8	0.56	37.0	6.0
Menton	2.7	0.18	20.0	3.6
Gnathion	2.7	0.24	19.3	3.4
Pogonion	2.7	0.39	18.6	3.4
Point B	2.6	0.03	16.1	2.7
PNS	5.0	0.31	24.6	4.1
ANS	3.8	0.21	12.0	2.2
Point A	3.3	0.38	14.8	2.4
LI root	3.9	0.53	11.9	2.7
LI tip	3.1	0.08	13.0	2.3

4 minutes. It should be noted however that the implementation was not optimized for speed.

Discussion

The introduction of more 'film-free' hospitals means that automatic cephalometric landmarking will become common-place, and more landmarking algorithms will be produced in an attempt to maximize accuracy and speed. To facilitate a direct comparison of the performance of these algorithms, the scanned cephalograms used in this study will be made available, on request, to other researchers. Patient identity will be concealed.

The ASM approach is different from the methods previously employed to automate cephalometric analysis. It combines a global shape model with a local search at each point, allowing image matching around each landmark within a framework that encapsulates learned knowledge of the permissible spatial relationships between landmarks. The fit is effectively driven by a voting system between the template points, making it robust to errors caused by false matches.

The local search method is of prime importance if the template is to converge on the correct structure. The method suggested by Cootes *et al.* (1995) was adopted, whereby a statistical model of the grey-levels along a profile at right angles to the template at each point was derived from the training set images. A statistical match between the grey-level model for each point and that found in the image yields the magnitude of each pull vector.

Two parameters need to be well chosen for the search to be optimal. The first is the size of the grey-level model that is used to define the search at each point. After experimentation, a value of two pixels either side (1.25 mm in total) was found to be the most successful. The second parameter is how far from each template point the search should extend. The best solution was to start with a large search distance, while the template fitting was still very approximate, then to reduce it as the fitting converged.

Another possibility is that, until a reasonable match has been found, it is better to use a smaller number of modes, introducing more later to refine the fit. The rate at which this should happen would ideally be dependent on how well the fitting has converged, as with the search distance. It is not clear how these refinements should be synchronized. After some experimentation, a hybrid scheme was produced that gave the best results. The settings shown in Table 3 were used until each had converged in turn.

The software implementation is entirely generic, allowing different cephalometric tracing

Table 3 Parameters for the search as the template converges.

No. modes used	0	1	2	3	4	5, 6, 7 … 29
Search distance (pixels)	67	57	47	37	27	17

styles, as well as different types of structure to be analysed and fitted. The design of the template and manual annotation are performed interactively by dragging template points into place over images using a mouse pointer. Images can be enlarged to enable close inspection of small features. The shape model produced by the PCA is visualized in an interactive manner, with visual controls to animate each of the modes of variation.

This study has shown that the ASM algorithm, as it stands, is not sufficiently accurate to be used to automate cephalometric analysis—this would require the majority of cephalometric landmarks to be acceptably placed in most images. However, the approach is well grounded in that it uses a model of what is being sought in the image, along with a model of what variation is permissible. In the earlier edge-based studies, it was not made clear how robust the algorithm is to variation in shape; the approach is somewhat fragile to such variation. Also, with the ASM approach the search algorithms do not need to be hand-crafted for each template structure, as they do with edge-based approaches.

A direct comparison with all but one (Rudolph *et al*., 1998) of the previous studies is unwise because the method for selection of the radiographs to be tested was not disclosed, it may be that the ones of optimum quality were selected. Also, some of the previous investigations used a very small number of test images, making their results statistically unreliable.

Sixty-three cephalograms were used to build the model of variation in this study. It may be that a larger training set would yield a more reliable statistical model encompassing more of the natural variation. This would improve the performance in images where the variation in shape is larger than or of a different nature to that seen in the present training set.

The ASM approach provides a framework for improvements, for example, local sub-image matching could be used to pinpoint the location of each landmark after the template has converged or while the fit proceeded. Another addition would be to augment the ASM fitting method with the more recent active appearance model (Cootes *et al*., 1998). This approach uses a grey-level model of the entire image to drive the template matching and should make the search more robust.

Conclusions

While ASMs do not provide sufficient accuracy for cephalometric analysis, the approach is of interest, and several possibilities exist to improve both accuracy and robustness. The implementation as it stands can be used as a time-saving first step before minute adjustments are made by an experienced orthodontist. The software can make measurements of the desired angles and ratios directly from the tracing.

This evaluation should provide a model for future studies and the image database can be used to make direct comparisons between the performance of algorithms proposed in the future.

Address for correspondence

Tim Hutton
Dental and Medical Informatics
Eastman Dental Institute
UCL
256 Gray's Inn Road
London WC1X 8LD, UK
http://www.eastman.ucl.ac.uk/~dmi/MINORI

Acknowledgements

This study was assisted by funding from the Matsumoto Dental University.

References

Baumrind S, Frantz R C 1971 The reliability of head film measurements. I. Landmark identification. American Journal of Orthodontics 60: 111–127

Broadbent H B 1931 A new x-ray technique and its application to orthodontia. Angle Orthodontist 1: 45–66

Broch J, Slagsvold O, Røsler M 1981 Error in landmark identification in lateral radiographic headplates. European Journal of Orthodontics 3: 9–13

Cardillo J, Sid-Ahmed M A 1994 An image processing system for locating craniofacial landmarks. IEEE Transactions in Medical Imaging 13: 275–289

Chen Y T, Cheng K S, Liu J K 1999 Improving cephalogram analysis through feature subimage extraction. IEEE Engineering in Medicine and Biology 18: 25–31

Cootes T F, Taylor C J, Cooper D H, Graham J 1995 Active shape models—their training and application. Computer Vision and Image Understanding 61: 38–59

Cootes T F, Edwards G J, Taylor C J 1998 Active appearance models. In: Burkhardt H, Neumann B (eds) Proceedings of the European Conference on Computer Vision 2. Springer, Stuttgart, pp. 484–498

Davis D N, Taylor C J 1991 A blackboard architecture for automating cephalometric analysis. Journal of Medical Informatics 16: 137–149

Forsyth D B, Davis D N 1996 Assessment of an automated cephalometric analysis system. European Journal of Orthodontics 18: 471–478

Forsyth D B, Shaw W C, Richmond S, Roberts C T 1996 Digital imaging of cephalometric radiographs, Part 2: Image quality. Angle Orthodontist 66: 43–50

Goodall C 1991 Procrustes methods in the statistical analysis of shape. Journal of the Royal Statistical Society B53: 285–339

Hill A, Cootes T F, Taylor C J 1992 A generic system for image interpretation. In: Proceedings of the 3rd British Machine Vision Conference. Springer-Verlag, Stuttgart, pp. 276–285

Hutton T J, Hammond P, Davenport J C 1999 Active shape models for customised prosthesis design. In: Horn W, Shahar Y, Lindberg G, Andreassen S, Wyatt J (eds) Proceedings of the Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making, Lecture Notes in Artificial Intelligence 1620. Springer-Verlag, Stuttgart, pp. 448–452

Lévy-Mandel A D, Venetsanopoulos A N, Tsotsos J K 1986 Knowledge-based landmarking of cephalograms. Computers and Biomedical Research 19: 282–309

Macrì V, Wenzel A 1993 Reliability of landmark recording on film and digital lateral cephalograms. European Journal of Orthodontics 15: 137–148

Parthasarathy S, Nugent S T, Gregson P G, Fay D F 1989 Automatic landmarking of cephalograms. Computers and Biomedical Research 22: 248–269

Rakosi T 1982 An atlas of cephalometric radiography. Wolfe Medical Publications, London

Richardson A 1981 A comparison of traditional and computerized methods of cephalometric analysis. European Journal of Orthodontics 3: 15–20

Rudolph D J, Sinclair P M, Coggins J M 1998 Automatic computerized radiographic identification of cephalometric landmarks. American Journal of Orthodontics and Dentofacial Orthopedics 113: 173–179

Singh G D, McNamara J A, Lozanoff S 1997 Thin-plate spline analysis of the cranial base in subjects with Class III malocclusion. European Journal of Orthodontics 19: 341–353

Tong W, Nugent S T, Gregson P G, Jensen G M, Fay D F 1990 Landmarking of cephalograms using a microcomputer system. Computers and Biomedical Research 23: 358–379