

Genetic Programming for Landmark Detection in Cephalometric Radiology Images

Vic Ciesielski
School of Computer Science
and Information Technology
RMIT University
GPO Box 2476V
Melbourne 3001,
Vic, Australia
vc@cs.rmit.edu.au

Andrew Innes, Sabu John
School of Aerospace,
Mechanical and
Manufacturing Engineering
RMIT University
ainnes@cs.rmit.edu.au
sabu.john@rmit.edu.au

John Mamutil
Braces Pty Ltd
404 Windsor Road,
NSW 2153
Australia
jrg@bigpond.net.au

Abstract

This paper describes the use of genetic programming to evolve object detection programs for craniofacial features in digital X-rays. The evolved programs use a feature set of pixel statistics of regions customised to the shapes of the landmark idiosyncrasies. The features are obtained from a square input window centred on the landmark and large enough to contain key landmark features. The detection program is applied, in moving window fashion, across the X-ray and the output of the program is interpreted as the presence/absence of the landmark at each position. During training a weighted combination of the detection rate and the false alarm rate is used as the fitness function. The method was tested on 4 landmark points, ranging from relatively easy to very difficult. Detection performance on the easier points was excellent and the performance on the very difficult point was quite good and the results suggest that a more careful crafting of the region shapes for the difficult point will lead to better detection. We believe that the methodology can be used successfully on other difficult real world detection problems.

1 Introduction

Advances and affordability in digital radiographic imaging have recently seen a demand for medical professions to automate analysis and diagnosis tasks that were once performed manually. Currently a cephalometric analysis is manually intensive, and it can take an experienced orthodontist up to thirty minutes to analyse one X-ray. The analysis involves finding a number of craniofacial landmarks and determining the distances and angles between them. Treatment planning is based on the results of this analysis. Figure 1 shows the landmarks specific to the cephalometric analysis performed by Braces Pty Ltd. The landmarks are located in both bony structure and soft tissue.

The aim of this paper is to establish whether genetic programming can be used to evolve programs for locating cephalometric landmarks in digital X-rays by using pixel level features. The four regions shown in Figure 2 each contain a cephalometric landmark. The regions have been selected to provide increasing difficulty in terms of variability of background noise, contrast and clutter. The regions containing the landmarks vary from easy to very difficult and contain the nose, upper lip, incisal upper incisor and the sella landmarks (refer to Figure 2). These regions have been delimited heuristically by using prior knowledge of the geometry of the human face and the easy to find bottom left corner point of the ruler. For example, at the scale of Figure 2, for most people, the tip of the nose will be in an area of about 10mm \times 12mm about 12mm below the bottom left corner of the ruler. This is region 1 in Figure 2. This method of cutting out the regions means that the landmark could be anywhere in the region, not just near the centre as shown in Figure 2.

Previous attempts at locating landmarks have had limited success [1, 2], but accurately locating landmarks for a large database has been unsuccessful. We believe that the best hope of success is to use an approach that involves some kind of learning. Previous work on the use of genetic programming for object detection has generally focused on simple objects. We would like to find out whether the genetic programming technique can be used for complex objects in difficult real world images. We hope to improve on previous attempts to find cephalometric landmarks by developing a robust method that can locate landmarks on a large database of images within a tolerance acceptable for a cephalometric analysis. Rakosi [3] suggests that an error of ± 2 mm is acceptable when locating landmarks. Error is defined as the difference between the location of the automatically detected landmark and the location as specified by an experienced orthodontist.

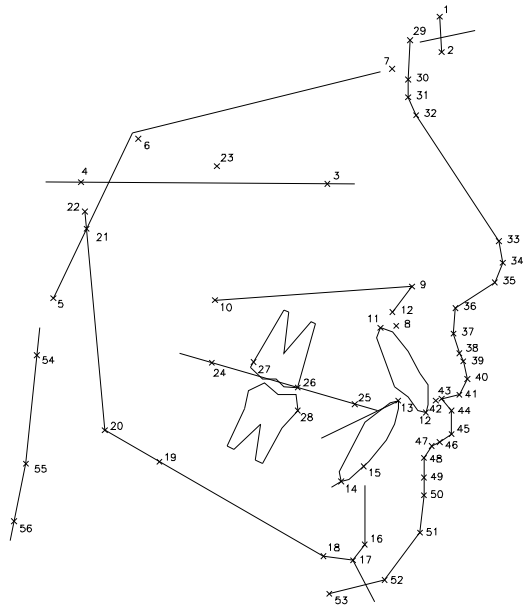


Figure 1: Line tracing of the fifty six landmarks required for a cephalometric analysis. Soft tissue landmarks are represented by points 29 -53.

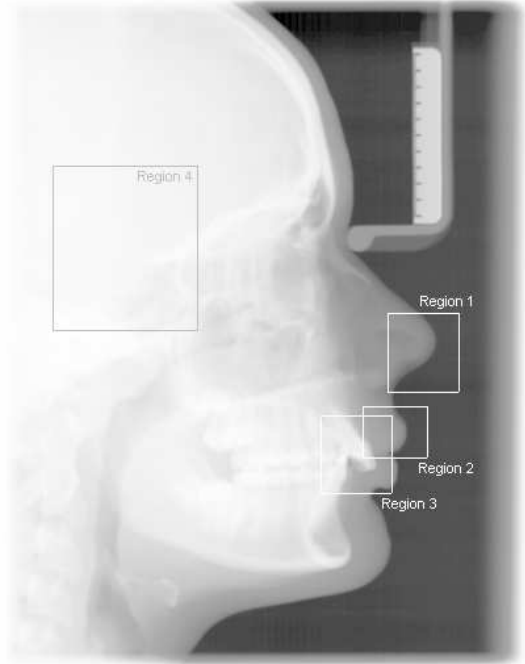


Figure 2: A digital cephalogram depicting four regions containing the mid nose, upper lip, incisal upper incisor and sella landmarks.

2 Previous Work

Traditionally the location of landmarks for a cephalometric analysis is performed manually using a tracing from X-ray film. More recently the film has been digitised and semi-automatic systems have allowed the orthodontist to plot landmarks directly to the screen. As a natural progression and whilst not a new concept, an automated cephalometric analysis was proposed by Husain et al. [4] in 1985.

Automatic cephalometric analysis has been attempted by no fewer than 20 independent researchers with varying degrees of success. The research can be categorised by two methodologies: prior knowledge and learning. Before 1989 the prior knowledge approach was dominant. The focus was to develop hand crafted programs for locating landmarks using image processing techniques in conjunction with prior knowledge of the cranial structure [5, 6]. During the past decade research has focused on artificial intelligence approaches to assist in detecting landmarks [1, 2, 7, 8]. Although some of the recent work has produced promising results [1, 2], to date, an automatic cephalometric analysis is still not forthcoming.

Earlier methods showed that hand crafted techniques could be improved using better image processing techniques, but the underlying flaw is that they are limited to a small population of images unless the code is written to handle variations in biological shapes. Some of

the other shortcomings may have been a result of detail lost when digitising film X-rays, or working with low resolution images. In any case, the hand crafted methods are limited to a small population of images defined by the code. Later attempts using learning approaches achieved better results due to a more flexible approach to the problem. The work by Cardillo et al. [1] has produced the most promising results locating 85% of landmarks within 2mm, although the results are based on a set of 40 images. Recent research by Hutton et al. [2] while not producing results as promising as Cardillo et al. uses a larger set of 63. The general consensus is that the difficulties in accurately locating landmarks are caused by the large variations in biological features, abnormalities, areas of soft tissue and areas with subtle changes in grey scale. There is also the variability due to the differences in signal-to-noise ratios of the digital X-ray procedure.

Recent methods of detecting landmarks show greater success when using a learning approach [1, 2]. This is due to a greater ability to cope with a large variety of biological shapes. Some of the recent methods have had limited success with small test sets [1], however the size of the test set suggests an inability to generalize to a larger variety of images. The most likely way to improve automatic landmarking is a learning approach using prior knowledge of facial geometry with an emphasis on pre-processing. Our goal is to develop an automated system that will be successful on a large number

of test images. The system will use prior knowledge of face geometry to determine the region in which a landmark should be located and then genetic programming to evolve a program to find the landmark in this region. This paper describes our work on four landmarks of increasing difficulty - the mid nose, the upper lip, the incisal upper incisor and the sella.

3 Methodology

Our approach involves dividing the landmark detection problem into 56 independent sub problems. Each of these problems involves finding a specific landmark in a region of the X-ray. For each landmark and each region we wish to evolve a program that can be placed over a small window and give a positive response if the window is centred on the landmark. The program will then be applied to the region, in moving window fashion, to find the landmark. Inputs to the evolved program will be a set of features based on partitioning the area surrounding a landmark into specific shapes (refer to Figure 3) individual to the landmark characteristics. The features consist of the mean and standard deviation of pixel intensities for each shape. The approach is based on [9] which used genetic programming to locate and classify objects such as heads/tails of different Australian coins in large images, and haemorrhages and micro-aneurisms in retina images.

The original image is first scaled down to 20% of the original image. This reduces the number of genetic program evaluations during training and reduces the effect of the Gaussian noise on the image.

A brief outline of the method we have used is:

1. Assemble a database of images with the known positions of landmarks to be located.
2. Use domain knowledge to extract regions in which the landmarks are expected to be.
3. Reserve some sub-images as 'unknowns' for measuring detection performance as the test set.
4. Determine `SQUARE_SIZE`, the size of a square centred on the landmark that will contain enough distinguishing information to permit the landmark to be identified. This is the size of the input window that will be used by the evolved program.
5. Manually determine the appropriate shapes (feature map) in the input window that are expected to apply to the training images and discriminate the landmark point from the background (refer to Figure 3).
6. Invoke an evolutionary process to generate a program which can determine whether a landmark in its input field.
7. Apply the generated program as a moving template to the reserved test images from step 3 and obtain the positions of the landmarks. Calculate the detection rate and the false alarm rate on the test set as the measure of performance.

3.1 Genetic Programming

3.1.1 The Terminal Set

In the context of genetic programming in object detection problems, terminals correspond to image features. Because each landmark is distinct in shape, gray scale and contrast, a different set of features or shapes is required for each landmark. The features presented in this paper correspond to the different shapes with their resulting means and standard deviations as shown in Figure 3. Note that the same feature maps are used for the nose and the lip points. In addition to these features, a terminal that generates a random number in the range of [0,255] was included.

3.1.2 The Function Set

The function set $\{+,-,*,/\}$ consists of four arithmetic operations. The $+$, $-$ and $*$ have their usual meanings, while $/$ represents a protected division which is the usual division operator except that a divide by zero produces zero. Additional operators were considered, but not used at this stage. Previous work by Zhang [10] has shown that additional operators (dabs, sin and exp) in the function set did not improve the detection rate. Convergence was also shown to be slightly slower when training to detect objects in easy pictures, however additional operators were shown to improve the rate of convergence when training to detect objects in more difficult images.

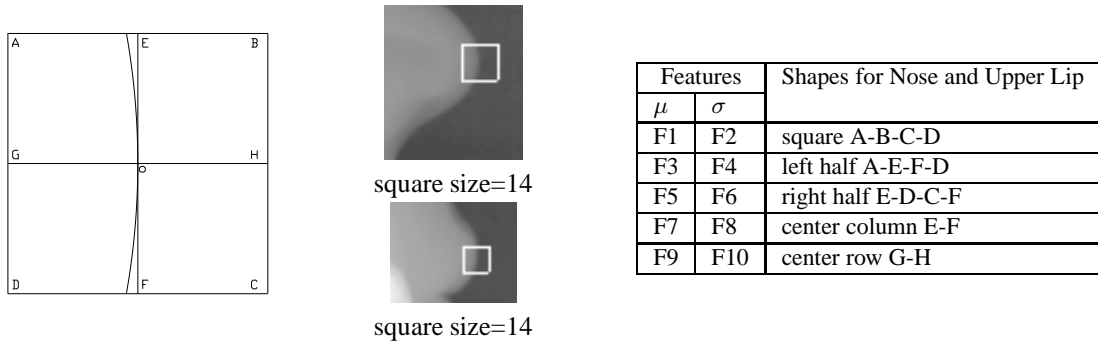
The output of the genetic program, *ProgOut*, is a floating point number which is interpreted as the likelihood that the evaluated position from the image is a landmark centre or background. During training the highest value of *ProgOut* from each image is used as the predicted position of the landmark. The predicted position given by the genetic program is then compared with the known true location and the example is classified as a true positive, a false alarm or unclassified (refer to section 3.1.3).

3.1.3 The Fitness Function

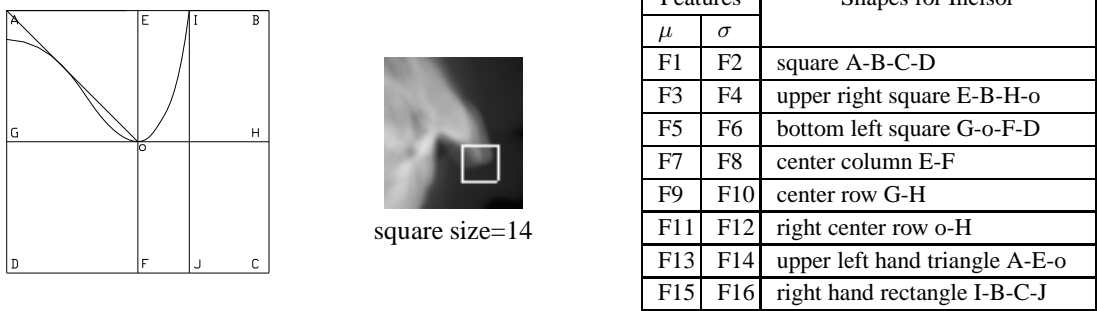
The fitness of a program during training is calculated by using detection rate and false alarm rate. The fitness is calculated as follows:

1. The program is applied as a moving window to each training image and the program output (*ProgOut*) evaluated at each pixel location. The predicted position of the detected landmark is recorded as the location corresponding to the highest value of *ProgOut* for each image.
2. If another position of the image has a *ProgOut* value within 5% of the highest evaluation (providing the position is not within `SQUARE_SIZE/2` of the recorded position), the landmark for that image is recorded as unclassified.

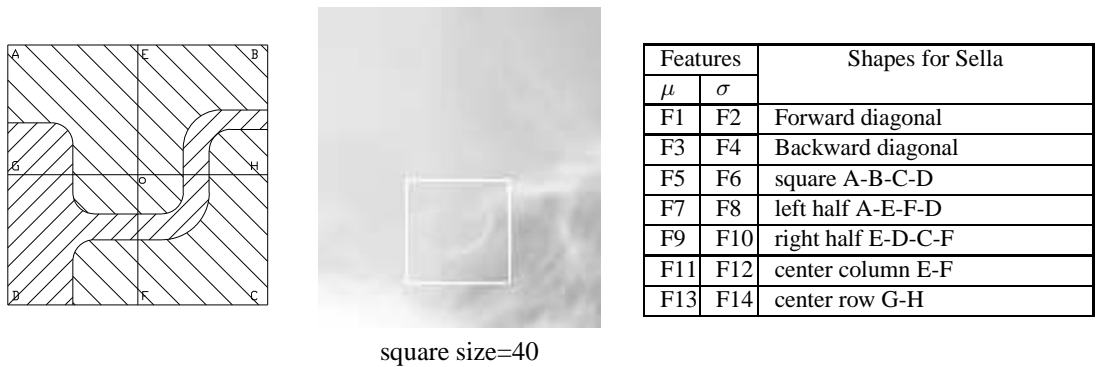
NOSE AND UPPER LIP



INCISOR



SELLA



Feature Map Feature Map positioned over landmark Features

Figure 3: Feature Maps and Features. The middle column shows the size of the feature map (Square size \times Square size, shown as the white square) relative to the size of the region.

3. If the image is not unclassified, then a comparison is made between the recorded position and the known true location of the landmark. A match (true positive) occurs when the comparison is within a set TOLERANCE of 5 pixels or 2mm. If the comparison is not within the set TOLERANCE then the landmark for the respective image is recorded as a false alarm.
4. At the conclusion of evaluating the program for each image in the training set the detection rate (Dr) and false alarm rate (Fr) are calculated.
5. The fitness is computed as per equation 1.

$$fitness = A \times Fr + B(1 - Dr) \quad (1)$$

where Fr is the false alarm rate, Dr is the detection rate, and A and B are constants that reflect the relative importance of the false alarm rate and detection rate.

The fitness function defined in equation 1 is constructed so that low fitness values are desirable. A perfect program will have a Dr of 1 and a Fr of 0 giving a fitness value of 0. The fitness function is also designed

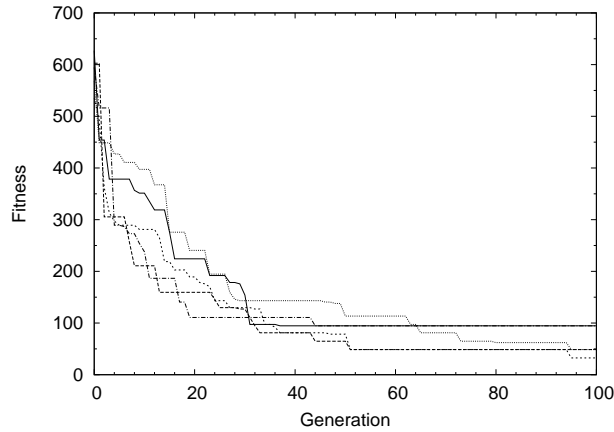


Figure 4: Best Fitness of Genetic Programs for the Incisor Point for 5 Runs over 100 Generations.

to evolve programs in which the *progout* value will be significantly higher at the predicted position of the landmark than at any other position. The graph in figure 4 shows the best fitness for five runs of the genetic program for the incisal upper incisor landmark over 100 generations. Note that optimal training was not achieved and that after 60 generations the improvement in program fitness was marginal. The fittest program at the end of 100 generations resulted in a detection rate of 94.6% and a false alarm rate of 5.4% when applied to the training images.

3.1.4 Genetic Programming Parameters

The values used in the genetic programming training are shown in Table 1. Population size is the number of individuals in the population, elitism (%) is the percentage of best individuals in the current population copied to the next generation, cross rate (%) is the percentage of individuals to be produced by cross over, mutation rate (%) is the percentage of individuals to be created by mutation, cross chance term (%) is the probability that in a crossover operation two terminals will be swapped, cross chance func (%) is the probability that in crossover operation random sub-trees will be swapped, maximum depth is the maximum depth allowed for programs, maximum generations is the termination point of the evolutionary process. A, B, Tolerance and Square size have already been defined. As shown in the table, the same parameter values have been used for all 4 landmarks, except for square size.

4 Results

To determine whether the proposed approach can be used to locate landmarks ranging from easy to most difficult, four landmarks were selected - the nose, upper lip,

incisal upper incisor and sella landmarks. Landmarks are classed as easy if the features surrounding the landmark are subject to minimal biological variation such as the nose landmark, whereas difficult landmarks such as the sella point are subject to greater biological diversity and clutter. A landmark is classified as found if the evolved program locates the landmark position within 2mm (5pixels) of the actual position as found by an orthodontist. If the landmark position is not within 2mm of the actual position the landmark is recorded as a false alarm. When a landmark has not been found by the genetic program it is unclassified (neither found or false alarm). Accuracy estimation was performed by 3 fold cross validation. The results are summarised in Table 2.

4.1 Region 1 (Nose)

The program shown in Figure 5 appeared in generation 56 in one of the runs and achieved a detection rate of 100% with no false alarms on the training data, resulting in early stopping of the evolutionary process. It also achieved 100% detection with no false alarms on the test data. The images in Figure 5 are indicative of the variations in size and shape of the nose. The white dot corresponds to the location of the nose as found by the genetic program. Note that, due to the heuristic used to cut out the regions, the tip of the nose can appear anywhere in a region.

(+ (+ (+ (* f9 f10) (* (+ (- f3 f7) f1) (* (- f10 f8) (- f3 f7)))) (* (+ (- f3 f7) f1) (* (- f10 f8) (- f3 f7)))) (* (+ (+ (- f3 f7) f9) f1) (- f10 f8)))

Figure 5: A generated program for the nose landmark.

4.2 Region 2 (Upper lip mid)

The program in Figure 7 is a result of the evolutionary process stopping after 100 generations. Training results produced a detection rate and false alarm rate of 83.8%

Parameters	Region 1, 2, 3, 4
Population size	100
Elitism (%)	10
Cross rate (%)	70
Mutation rate (%)	20
Cross chance term (%)	15
Cross chance func (%)	85
Maximum depth	6
Maximum generations	100
A	200
B	1000
Tolerance (pixels)	5 (2mm)
Square size (pixels)	14, 14, 14, 40

Table 1: Parameter Values.

	Nose		Lip		Incisor		Sella	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
No. of objects	73	36	74	36	74	36	73	36
No. of objects found	72.7	36.0	62.0	28.7	70.0	34.7	44.0	23.0
No. of false alarms	0.3	0.0	8.7	4.7	3.0	1.3	24.3	11.0
No. unclassified	0.0	0.0	3.3	2.7	0.0	0.0	4.7	2.0
Detection rate (%)	99.5	100.0	83.8	79.6	94.6	96.3	60.3	63.9
False alarm rate (%)	0.0	0.0	11.7	12.0	4.0	3.7	33.3	20.6

Table 2: Results of 3-fold cross validation.

and 11.7% respectively. Results from test images unseen during training produced a detection rate of 79.1% and a false alarm rate of 12.0%. The images in Figure 8 are indicative of the types of images containing the upper lip from the test set. The white dot corresponds to the location of the upper lip landmark as found by the genetic program shown in Figure 7. Although the nose and upper lip points are of similar difficulty detection accuracy for the lip point is lower because of the ambiguity between top and bottom lip.

4.3 Region 3 (Incisal upper incisor)

The program in Figure 9 resulted from 100 generations of the evolutionary process. Training results produced a detection rate and false alarm rate of 94.6% and 4.0% respectively. Results from test images unseen during training produced a detection rate of 96.3% and false alarm rate of 3.7%. The images in Figure 10 are indica-

tive of the types of images containing the incisal upper incisor from the test set. Detection was generally better on images subject to overbite than on images containing under-bite.

4.4 Region 4 (Sella)

The genetic program in Figure 11 is a program that has evolved after 100 generations. Training results produced a detection rate and false alarm rate of 60.3% and 33.3% respectively. Results from test images unseen during training produced a detection rate of 63.9% and false alarm rate of 30.6%. The images in Figure 12 are indicative of the types of images containing the sella point from the test set. The results for the sella are not as good as for the other landmarks. This is expected as there is clearly more variation and complexity in grey levels surrounding this point.

5 Discussion

As can be seen from Table 2 the training and test accuracies are generally close to each other. This indicates that no overtraining, a major problem when using machine learning techniques with images of this kind, has occurred.

An evolved program can be viewed as a classifier which discriminates a landmark of interest (class 1) from the other points (class 2) in the images. Thus there will be a decision boundary in feature space between these two classes. The boundary is the multi-variate polynomial represented by the program. It is reasonable to ask whether the same detection performance could be

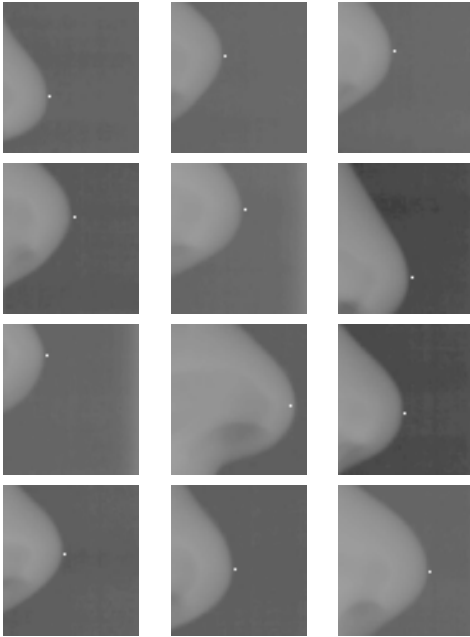


Figure 6: Twelve randomly selected nose test images. The white dot indicates the position found using the program shown in Figure 5. The genetic program was successful for all of these images.

$$\begin{aligned}
 & (/ (+ (+ (* 57.9886 (- (- f9 f5) (/ f9 f6))) (- (+ 131.325 \\
 & f7) (/ (* f6 f6) (/ f4 (- 37.5046 f5)))))) (+ (+ (+ (* 57.9886 \\
 & (- (- f9 f5) (/ f9 f6))) (- (+ 131.325 f7) (/ (* f6 f6) (/ f4 \\
 & f1)))) (+ (+ f3 f9) (+ (* f6 f5) (/ f6 f2)))) (+ (* f6 f5) (/ \\
 & f6 f2)))) (- (* (/ (- (* (- (+ 131.325 f7) (/ f4 f7)) (+ f5 (* \\
 & f9 f2))) (+ (* f6 f5) f10)) (/ (- f3 f3) (* f9 f6))) (* f3 (+ \\
 & f3 (* f8 f2)))) (/ f10 (- (/ (/ f7 f4) (- 57.9886 f8)) (/ (/ f3 \\
 & f10) (+ f4 (/ f4 f7))))))
 \end{aligned}$$

Figure 7: A generated program for the upper lip landmark.

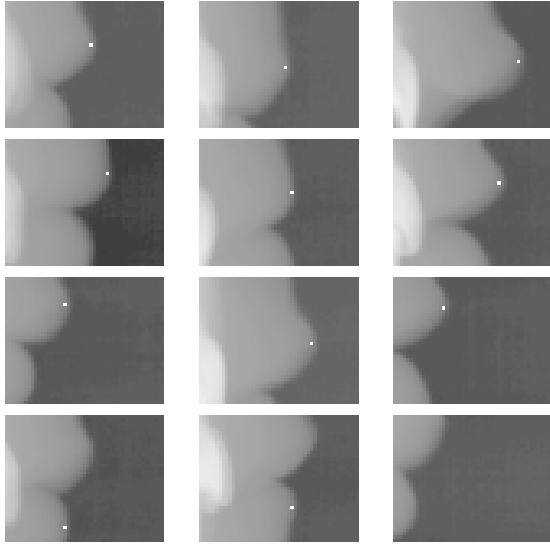


Figure 8: Upper lip images. The top three rows are images where the program found the landmark and the bottom left and bottom middle images are false alarms. The bottom right image is unclassified.

```
(- (- (+ f2 (- f12 f5))) (/ (* (+ (/ f4 f10) (+ (/ f6 f10) (/ f3 (- f9 250.8)))) f10) (/ (- f12 f1) f13))) (/ (* (+ (/ f4 f10) (+ (/ (- f12 f1) (/ f11 f10)) (+ 250.8 f4)) (/ f3 (- f9 250.8)))) f10) (/ (- f12 f1) (+ f13 f4))))
```

Figure 9: A generated program for the incisal upper incisor.

achieved with a simpler boundary. Earlier work [11, 10] suggests that a complex boundary is needed to avoid a high false positive rate. In [10], neural network classifiers trained by back propagation were used in place of genetic programs. The networks were able to find objects of interest but gave a significant number of false positives. The false positive rate could be reduced by subsequent refinement of the weights using a genetic algorithm with a similar fitness function to equation 1. It appears that a complex decision boundary found by penalising false positives during training is necessary for accurate detectors.

Table 3 shows the frequency of occurrence of features in the the most successful evolved programs. For the nose point, for example, 42 runs out of 84 terminated with a detection rate of 100% and feature F1 appeared at least once in 35 of them. In as much as occurrence of a feature is a rough measure of its usefulness in discriminating a landmark from the background, this table gives an indication of the importance of each feature.

6 Conclusion

The genetic programming methodology described in this paper has been successfully used to evolve detection programs for a number of cephalometric landmarks.

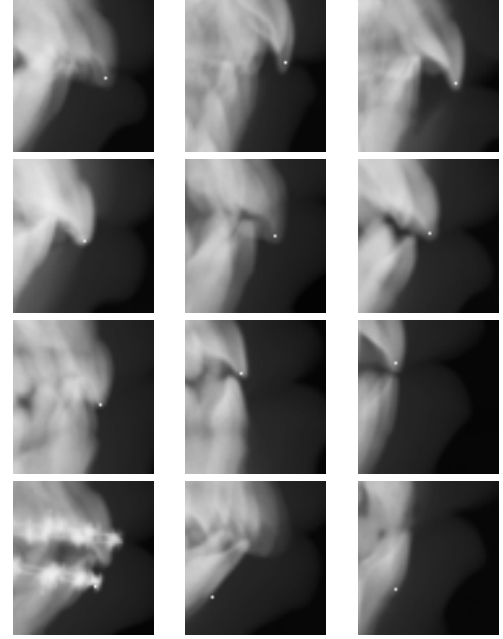


Figure 10: Incisor images. The top three rows are images where the program found the landmark and the bottom row are false alarms.

```
(/ (/ (- (- f1 (- (+ f2 f11) (- f1 f14))) (- f3 (- (- f1 (- (+ f2 f11) (- f1 f14))) (- f3 f2)))) (/ (- (+ f2 f13) (- f1 f2)) (/ (* f2 f1) f8) (/ (* f2 f1) f14)))) (+ f2 f1))
```

Figure 11: A generated program for the Sella landmark.

Detection performance on the easier landmarks was excellent and the performance on the very difficult landmark was very promising. While some of the landmarks have been classified as ‘easy’ it is important to note that there is a large variation in human shapes and sizes and that the accuracy obtained is a non trivial achievement. A drawback of the method is that run times for the evolutionary process are high, one run of 100 generations taking around 3 hours on a 1400MHz Pentium 4 computer. However this is a once-only cost. Applying the evolved program to an image is very fast, taking around 0.1 seconds. Given the coarseness of the features used, particularly for the nose tip and incisor points, the detection accuracy achieved is surprising and suggests that with more attention to the features the approach will be successful on the more difficult landmarks. In future work we plan to investigate learning methods for finding good shapes and corresponding features to use in the input window of the evolved object detection programs.

Acknowledgements

This work is funded by the Australian Research Council (ARC) SPIRT grant scheme in partnership with Braces Pty Ltd from grant no. C00107119 and supported by grant EPPNRM054 from the Victorian Partnership for

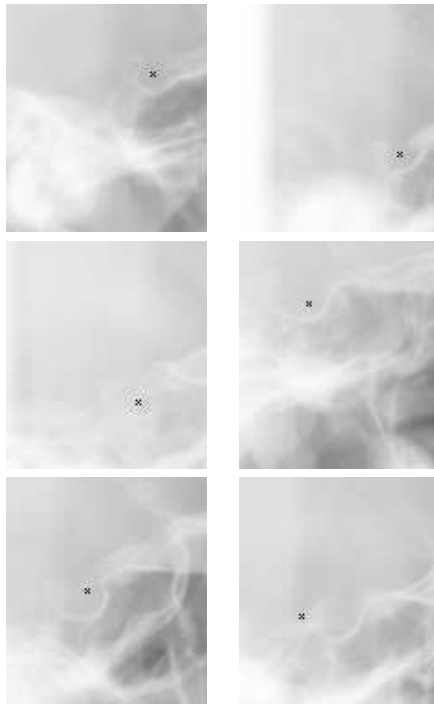


Figure 12: Sella images. The black dot indicates the position found by the program in Figure 11. The top two rows are images where the program found the landmark and the bottom row are false alarms.

Advanced Computing.

References

- [1] J. Cardillo and M.A. Sid-Ahmed, "An image processing system for locating craniofacial landmarks," *IEEE transaction on Medical Imaging* June 1994, vol. 13, no. 2, pp. 275–289, 1994.
- [2] T.J. Hutton, S. Cunningham, and P. Hammond, "An evaluation of active shape models for the automatic identification of cephalometric landmarks," *European Journal of Orthodontic*, vol. 22, no. 5, pp. 499–508, October 2000.
- [3] T. Rakosi, *An atlas of cephalometric radiography*, Wolfe Medical Publications, London, 1982.
- [4] Z. Hussain and H.H.S. Ip, "Automatic identification of cephalometric features on skull radiographs," *ACTA polytechnica Scandinavica-Applied physics series*, , no. 150, pp. 194–197, 1985.
- [5] A.D. Levymandel, A.N. Venetsanopoulus, and J.K. Tsutos, "Knowledge-based landmarking of cephalograms," *Computers and Biomedical Research*, vol. 19, no. 3, pp. 282–309, June 1986.
- [6] W. Tong, S.T. Nugen, P.H. Gregson, G.M. Jensen, and D.F. Fay, "Landmarking of cephalograms using a microcomputer system," *Computers and*

Terminal	Nose	Lip	Incisor	Sella
Success Rate	42/84	19/77	17/48	32/263
Dr Achieved	100%	77%	94%	52%
F1	35/42	17/19	15/17	31/32
F2	28/42	16/19	13/17	32/32
F3	36/42	16/19	10/17	23/32
F4	32/42	17/19	14/17	16/32
F5	33/42	18/19	11/17	24/32
F6	30/42	14/19	9/17	21/32
F7	35/42	17/19	10/17	18/32
F8	29/42	13/19	16/17	19/32
F9	26/42	19/19	9/17	15/32
F10	30/42	14/19	12/17	15/32
F11			9/17	19/32
F12			9/17	14/32
F13			11/17	27/32
F14			9/17	19/32
F15			14/17	
F16			9/17	

Table 3: Frequency of Appearance of Features in Successful Runs.

Biomedical research, vol. 23, no. 4, pp. 358–379, August 1990.

- [7] Y. Chen, K. Cheng, and J. Liu, "Improving cephalogram analysis through feature subimage extraction," *IEEE engineering in Medicine and Biology*, pp. 25–31, 1999.
- [8] M. Desvignes, B. Romaniuk, R. Demoment, M. Revenu, and M.J. Deshayes, "Computer assisted landmarking of cephalometric radiographs," in *Proceedings of the 4th IEEE Southwest Symposium on Image Analysis and Interpretation*, 2000, pp. 296–300.
- [9] M. Zhang and V. Ciesielski, "Genetic programming for multiple class object detection," in *Proceedings of the 12th Australian Joint Conference on Artificial Intelligence*. 1999, pp. 180–191, Springer-Verlag.
- [10] M Zhang, *A domain independent approach to 2D object detection based on the neural and genetic paradigms*, Ph.D. thesis, RMIT University, 2000, <http://www.cs.rmit.edu.au/~vc/papers/zhang-phd.ps.gz>.
- [11] M. Zhang and V. Ciesielski, "Using the back propagation algorithm and genetic algorithms to train and refine neural networks for object detection," in *Proceedings of 10th International Conference and Workshop on Database and Expert Systems Applications (DEXA99)*, T. Bench-Capon, G. Soda, and A.M. Tjoa, Eds. Aug 1999, vol. 1677, Lecture Notes in Computer Science, pp. 626–635, Springer, Heidelberg.