A Comparison of Shape Constrained Facial Feature Detectors

D. Cristinacce

Imaging Science and Biomedical Engineering
University of Manchester, U.K
david.cristinacce@stud.man.ac.uk

T. F. Cootes

Imaging Science and Biomedical Engineering
University of Manchester, U.K
tim.cootes@man.ac.uk

Abstract

We consider the problem of robustly and accurately locating facial features. The relative positions of different feature points are represented using a statistical shape model. We construct an individual detector for each feature point, which is used to generate a feature response image. The quality of a given hypothesised shape can be evaluated quickly by combining values from each response image. We use global search to predict the approximate position of the face, then refine the hypothesis using non-linear optimisation. The result is an algorithm capable of robustly and accurately matching a face model to new images, which we refer to as Shape Optimised Search (SOS). We describe SOS in detail and compare the performance of the algorithm when three different classes of feature detectors are used. We demonstrate that the approach is capable of outperforming the well known Active Appearance Model method.

1. Introduction

This paper addresses the problem of accurately finding facial features, such as the eye pupils, corners of the mouth etc. This is important for many tasks, such as quantitative measurement, feature tracking and face recognition.

The face is first localised using a global face detector, which provides an approximate location and scale. In the first stage of Shape Optimised Search (SOS), feature detectors are applied to the face region and the response of each feature detector recorded as a fit quality surface. The shape model is initialised by the average feature points indicated by the global face location. The fitting parameters of the shape model are then optimised using a non-linear optimiser. The objective function is the sum of feature responses indicated by the current shape parameters, constrained so that the shape parameters form a valid shape. The non-linear optimiser is then able to drive the shape parameters to maximise the sum of feature responses, within shape constraints.

The SOS procedure is general in the sense that any feature detector can be used to provide a fit quality surface and any optimiser can be used to compute the optimal shape parameters. In this paper the optimiser is the Nelder-Meade simplex method [11]. Three different feature detection methods are compared, namely Viola-Jones feature templates [12], orientation maps [8] and normalised correlation . The technique is applied to faces, however SOS could easily be applied to other image interpretation tasks. The approach is shown to be robust, relatively quick and provide superior performance to the active appearance model matching method.

2 Background

The task of facial feature detection has generally been addressed by model based algorithms that combine shape and texture modelling [4] [2] [13].

A popular method is the active shape model (ASM) algorithm due to Cootes *et. al.* [4]. The ASM models grey-level texture using local linear template models and the configuration of feature points using a statistical shape model [6]. An iterative search algorithm seeks to improve the match of local feature points and then refine the feature locations by the best fit of the shape model. Fitting to a large number of facial feature points mitigates the effect of individual erroneous feature detections.

In later work Cootes *et. al.* [2] combine shape and texture in the active appearance model (AAM) approach to model matching. The shape and texture are combined in one PCA space. The model then iteratively searches a new image by using the texture residual to drive the model parameters. The AAM aims to form a more robust model compared with the original ASM approach, by explicitly modelling texture correlation over the entire model region, rather than relying on local texture profiles.

However a problem with both the ASM and AAM approaches is that a good initialisation, close to the correct solution is required, otherwise both methods are prone to local minima. This problem is partially address by Haslam *et. al.* [9] who use non-linear optimisation of shape model parameters to fit a joint model of grey-level texture profiles

and shape to the image data. However, this method is slow due to the need to resample image texture at each iteration and is hampered by the use of linear profile texture models.

Other authors have attempted to combine feature detections and shape modelling to avoid local minima. For example Burl *et. al.* [1] use multi-scale Gaussian derivative filters to detect facial features and then select the combination of features which provide the most likely shape. Wiskott *et. al.* [13] use "Gabor jets" and the inter-feature distances as weak shape constraints. The search proceeds in a coarse-to-fine manner, each stage allowing slightly greater shape variation.

In earlier work, Cristinacce and Cootes [5] detect facial features using boosted classifiers and accept the most probable set of feature responses that also pass a shape constraint. However, it is difficult to extend this method to more than four or five feature point due to combinatorial explosion. In this paper, the number of feature points is increased to seventeen, by incorporating the shape constraint into an optimisation scheme. The optimisation of shape parameters is related to Haslam *et. al.* [9]. However SOS constrains the search using a global face detector, uses stronger local feature detectors, precomputes the feature responses and only then optimises the shape parameters. We apply our approach to faces, whereas Haslam *et. al.* applied their algorithm to medical image data.

For global face detection we use the popular boosted classifier method, due to Viola and Jones [12]. For local search, any numerical optimisation or feature detection method can be used. In this paper, the shape parameters are optimised using Nelder-Meade Simplex [11]. Results are presented for three different local feature detectors, namely boosted classifiers of individual facial features [12], orientation maps , as described by Fröba and Küllbeck [8] and normalised correlation .

The SOS is efficient with any feature detector, due to precomputation of feature responses and is shown to outperform the AAM approach when applied to the task of facial feature localisation.

3 Methodology

3.1 Shape Modelling

To learn the plausible configurations of feature points a shape model is built using the methods introduced by Cootes *et. al.* [4]. The shape model is built from a set of shapes with corresponding landmarks, which are aligned into a common co-ordinate frame. Principle component analysis (PCA) is then used to reduce the dimensionality of the data, such that any shape $\mathbf{x}$ in the training set can be approximated as follows.

$$\mathbf{x} \approx \bar{\mathbf{x}} + \mathbf{\Phi b} \tag{1}$$

Here $\mathbf{b}$ is a vector of shape parameters, $\mathbf{\Phi}$ is a matrix of eigenvectors and $\bar{\mathbf{x}}$ is the mean training shape. By varying the elements of $\mathbf{b}$ the shape $\mathbf{x}$ can be varied. The variance of the i^{th} parameter, b_i, across the training set is given by the eigenvalue λ_i. By applying limits of $\pm 3\sqrt{\lambda_i}$ to the parameter b_i it is possible to ensure that the generated shape is similar to the original training set. As described later, the shape parameters $\mathbf{b}$ are varied in order to optimise the response at each feature point, but still provide a valid overall shape.

Figure 1: Landmarked training data

Our shape model is built from a set of landmarked faces as shown in Figure 1. In the following we describe results using 17 point shape models (all landmarked points are used except the temples and chin, see Figure 1).

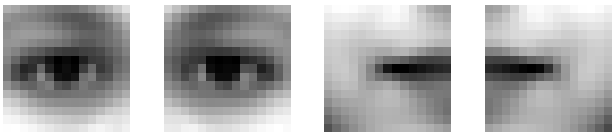
3.2 Feature Point Models

To locate individual features it is necessary to model local image patches and build feature detectors based on these regions. In this paper, 15x15 pixel regions around each landmark point are extracted from each face in the training set, where the face region is rescaled to 100x100 pixels. Some example training images for the eye and mouth corner detectors are shown in Figure 2.

Figure 2: Example training images for the right/left eye and right/left mouth corner regions

The three types of feature detector are trained on the data shown in Figure 2, collected from 1055 training set images.

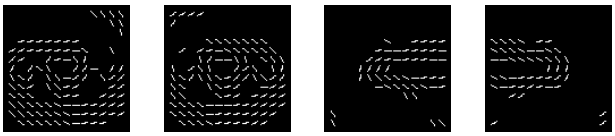
The simplest detector is a normalised correlation template, constructed by computing an average image for each feature and scaling such that the pixel values have zero mean and unit variance. The normalised template for the eyes and mouth corners are shown in Figure 3.



right eye left eye right m/cnr left m/cnr

Figure 3: Normalised correlation templates for eye and mouth corner features

Orientation maps are constructed by application of a sobel edge filter, to each averaged feature image, using the method described by Fröba and Küllbeck [8]. The orientation maps for the eyes and mouth corners are shown in Figure 4.



right eye left eye right m/cnr left m/cnr

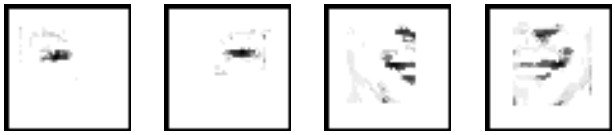
Figure 4: Orientation maps for eye and mouth corner features

The third feature detector is the boosted classifier due to Viola and Jones [12]. This method models an image region by creating a large set of simple Haar wavelet like features, then uses AdaBoost [7] to select features that differentiate between the object class and a set of background non-face images. The selected features are individual weak classifiers, i.e. they give poor performance, however the AdaBoost scheme weights the feature classifiers to produce one strong combined classifier.

The feature detector is cascaded to improve efficiency using a ten level cascade with ten features in each level. This form of feature detection has been successfully applied to the whole face region by Viola and Jones [12] and later applied to individual facial features by Cristinacce and Cootes [5].

3.3 Feature Search

Before searching for individual features the object of interest is assumed to be approximately located in the image. Therefore the approximate scale and orientation of the object is assumed to be known. Also each individual feature within the whole object is assumed to lie within a region, whose size can be estimated from the training set.



right eye left eye right m/cnr left m/cnr

Figure 5: Response images for the eyes and mouth corner detectors (black implies strong response)

In our system upright frontal faces are automatically detected using a Viola and Jones [12] style face detector. The range of each individual facial feature within the detected whole face region is also learnt in advance, by applying the face detector to a hand labelled verification set.

When searching an unseen image, the feature detectors are applied to their individual search regions and a quality of fit measure returned for each pixel within the region. Responses for the eye and mouth corner using Viola and Jones feature detectors are shown in Figure 5.

Dark regions in Figure 5 indicate a high likelihood of a correct match, whereas lighter regions indicate that a pixel is unlikely to be the correct feature location.

3.4 Non Linear Optimisation

Given a set of feature response images I_i, the mean feature locations within the object are used to initialise the shape. Here shape is represented as a concatenated vector of X and Y co-ordinates, as follows.

$$\mathbf{X} = (X_1, \ldots, X_n, Y_1, \ldots, Y_n)^T \tag{2}$$

The shape $\mathbf{X}$ is represented by the shape parameters $\mathbf{b}$ from Equation 1 and a transformation $T_\mathbf{t}$ from the shape model frame to the image frame. Therefore $\mathbf{X}$ is calculated from the shape parameters as follows.

$$\mathbf{X} \approx T_\mathbf{t}\left(\bar{\mathbf{x}} + \mathbf{\Phi}\mathbf{b}\right) \tag{3}$$

Here $T_\mathbf{t}(x)$ applies a similarity transform with parameters $\mathbf{t}$. We concatenate the shape and pose parameters into $\mathbf{p} = \left(\mathbf{t}^T|\mathbf{b}^T\right)^T$. The objective function f($\mathbf{p}$) for a given vector $\mathbf{p}$ is then defined as follows.

$$f(\mathbf{p}) \quad = \quad \begin{cases} \sum_{i=1}^{n} I_i(X_i, Y_i) & \text{if } |b_i| < 3\sqrt{\lambda_i}\ \forall i \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Where bilinear interpolation of the feature response image $I_i(X, Y)$ is used for non-integer (X, Y). Here it is implicitly assumed that the feature responses have similar ranges for each response image. A threshold of three standard deviations ($|b_i| < 3\sqrt{\lambda_i}$) is applied to each shape parameter b_i to ensure a valid shape.

Figure 6: Example test set images

The SOS algorithm is simply to vary the parameter vector $\mathbf{p}$ to maximise the fitting function $f(\mathbf{p})$. This can be achieved using any optimisation technique. In this paper we use the Nelder-Meade simplex method [11]. When the simplex method produces no significant change in the parameter vector $\mathbf{p}$, i.e. a maxima has been found, the algorithm terminates and the final shape parameters $\mathbf{p}$ determine the location of each feature.

4 Experiments

4.1 Test Data

The search algorithm is tested on a publicly available images set known as the BIOID database [1]. This data set was first used by Jesorsky *et. al.* [10], to evaluate face detection and eye finding algorithms, but is now available with a set of 20 manually labelled feature points. The BIOID images consist of 1521 images of frontal faces taken in uncontrolled conditions using a web camera within an office environment. The face is reasonably large in each image, but there is background clutter and unconstrained lighting. Example images from the BIOID data set are shown in Figure 6.

4.2 Testing Criteria

Each image is searched using a cascaded classifier [12] to detect upright faces. The candidates are ranked, using the sum of classifier scores from each level of the cascade and the highest ranking candidate selected as the location of the face. We consider the face location step to be successful if the mean distance from the found eye positions (as predicted by the detected region) to the true eye locations (annotated manually) is less than 30% of the true inter-ocular distance. To concentrate on the feature detection, we discard those examples in which the global search failed this threshold. Global search failed on seventy four of the BIOID images (74/1521=4.8%) [2], which were excluded from the analysis of feature detection performance given below.

[1] http://www.humanscan.de/support/downloads/facedb.php

[2] The relatively poor detection rate on the BIOID data set is due to many images in which the top of the face was cropped from the image

When evaluating the local feature search, we measure the positional error of each feature and express it as a proportion of the inter-ocular distance. There are seventeen feature points, as shown in Figure 1, excluding the temples and chin. The average point to point error of all seventeen feature points is used as a distance metric m_e to describe the accuracy of each individual search. The cumulative distribution of this error measure is then used to compare different search algorithms. For example see Figure 8.

4.3 Comparison with AAM

Different formulations of SOS are tested on the BIOID data set. The method is also compared with the well established Active Appearance Model (AAM) approach [2].

The AAM extends the shape model described in section 3.1, by also applying PCA to the grey level texture. The combined model then forms a compact representation of the training set variation, which can be fitted to a given image region. The AAM learns the relationship between the texture residual of the model near the optimal solution [2]. This allows the AAM to converge to the correct solution, given an unseen example object and a sufficiently accurate starting position.

The method describe in [3] is used to constrain the AAM search, by learning the isotropic errors on the feature points predicted by the global face detector. We experimented with different formulations of the AAM and present results using our best performing model.

5 Results

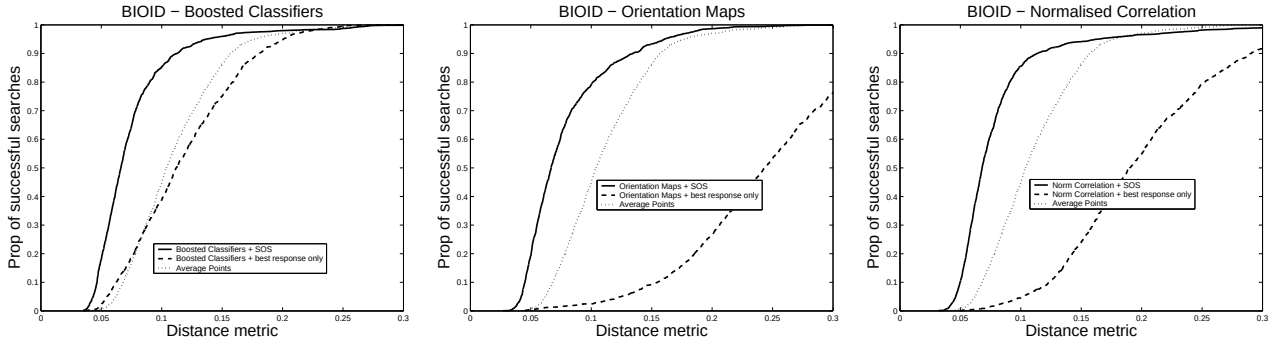
5.1 Shape Constraints

This section compares feature detection performance with and without shape constraints. Three different feature search algorithms are compared:-

1. Mean position predicted from full face match (no feature search) (dotted line).
2. Best feature response for each detector (no shape information used) (dashed line)
3. SOS using Nelder-Meade Simplex (solid line).

The mean positional error m_e is calculated for each BIOID image and the cumulative probability distributions for m_e plotted in Figure 7.

The graphs in Figure 7 show that the SOS approach is more accurate than the average points predicted by the global face detector. However, when shape is ignored and only the best feature response returned by each detector, the local search is actually worse than the average points.



(a) boosted feature detectors (b) orientation map feature detectors (c) normalised correlation feature detectors

Figure 7: Average mean positional error(m_e) of the 17 feature points comparing SOS, best response only and mean locations predicted by the global face detector

For example, with boosted detectors (see Figure 7(a)) and threshold $m_e < 0.15$ the shape optimised search is successful for 96% of faces, using average points works for 86% of faces, but only 75% of faces if the search is unconstrained. The reduction in performance when *not* using shape constraints for orientation maps and normalised correlation is more extreme, i.e. $95\% \rightarrow 9\%$ for orientation maps (see Figure 7(b)) and $95\% \rightarrow 22\%$ for normalised correlation (see Figure 7(c)).

5.2 Comparison of Feature Detectors

The final accuracy of SOS when using different feature detectors is shown in Figure 8.

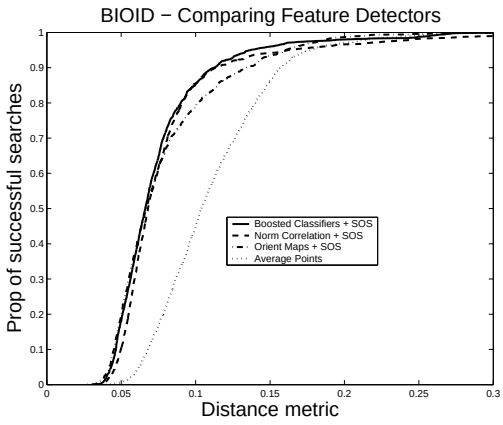


Figure 8: Average mean positional error(m_e) of 17 feature points when applying SOS with different feature detectors

Figure 8 shows that the proportion of successful searches with $m_e < 0.15$ on the BIOID images is 96% using boosted feature detectors and 95% using orientation maps or normalised correlation . Therefore the accuracy of SOS varies little when using different feature detectors. However the improvement in performance compared to unconstrained search as shown in Figure 7 is more dramatic with orientation maps and normalised correlation detectors.

5.3 Comparison with AAM

The performance comparison between the AAM search and SOS using boosted detectors is shown in Figure 9.

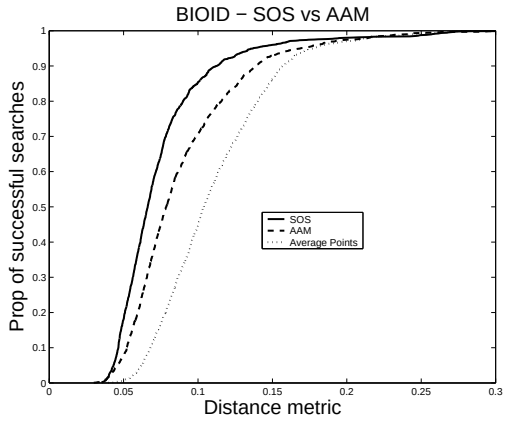


Figure 9: Average mean positional error(m_e) of 17 feature points, comparing shape optimisation and the AAM search

The best performing search is the shape optimised (SOS) approach. Figure 9 shows that with $m_e < 0.15$ shape optimisation is successful for 96% of faces, whilst the AAM works for 92% of faces. If the face features are required to be localised more accurately (i.e. $m_e < 0.1$) then the

shape optimised search works for 85% of faces, but only 70% using the AAM. Both the AAM and SOS algorithm give more accurate results than the whole face candidate prediction method. However, SOS is superior to the AAM search at all values of m_e, so is the the most robust method.

5.4 Speed of feature detection

Timings were carried out on the BIOID images (384*286 pixels) using modest hardware, a 500Mhz PII processor. The global search using Viola and Jones [12] requires $\sim 300ms$. The complexity of the SOS is mainly dependent on the time taken for the simplex algorithm to converge, the time to calculate the feature response images is less critical. SOS requires $\sim 150 - 500ms$. The AAM search converges in $\sim 200ms$. Therefore the search time for the SOS is comparable to the AAM and the whole image can easily be processed in less than a second, even on a very old machine.

6 Summary and Conclusions

A general method of face matching has been described, which we refer to as Shape Optimised Search (SOS). A shape model is used to learn the spatial distribution of features over the face. Individual detectors are trained on facial feature points and provide a quality of fit surface for each feature. The shape parameters are then optimised to maximise the sum of feature responses. The algorithm is general in the sense that any feature detectors can be combined with any non-linear maximisation scheme.

The algorithm is illustrated with boosted feature detectors, orientation maps , normalised correlation and combined with Nelder-Meade Simplex optimisation. The technique is used to model facial features, but could be applied to any other deformable object. On a large face data set SOS using boosted features is shown to outperform the well known AAM approach. The computational complexity of both methods is similar, but the new method is more robust and more accurate.

Further work will involve applying SOS to other deformable objects as well as faces. Another avenue of investigation is feature matching to partially occluded objects. Initial experiments show that SOS is extremely robust to semi-occluded faces. Some example searches are shown in Figure 10. The feature search is successful in most cases, where enough of the face is visible for the global search to detect the face. Manual landmarking of a test set will allow a more thorough investigation of feature finding on occluded objects.

In conclusion SOS is relatively fast, very robust and gives better performance than the AAM when tested on a large data set. Empirical experiments have suggested that the method is also highly robust to partial occlusion. We anticipate the technique will have wide applications in many areas of model based computer vision.

References

[1] M. Burl, T. Leung, and P. Perona. Face localization via shape statistics. In 1^{st} *International Workshop on Automatic Face and Gesture Recognition 1995*, Zurich, Switzerland, 1995.

[2] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In H.Burkhardt and B. Neumann, editors, 5^{th} *European Conference on Computer Vision*, volume 2, pages 484–498. Springer, Berlin, 1998.

[3] T. F. Cootes and C. J. Taylor. Constrained active appearance models. In 9^{th} *International Conference on Computer Vision*, pages 748–754, 2001.

[4] T. F. Cootes, C. J. Taylor, D.H. Cooper, and J. Graham. Active shape models - their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, January 1995.

[5] D. Cristinacce and T. Cootes. Facial feature detection using adaboost with shape constraints. In 14^{th} *British Machine Vision Conference*, pages 231–240, 2003.

[6] I. Dryden and K. V. Mardia. *The Statistical Analysis of Shape*. Wiley, London, 1998.

[7] Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *2nd European Conference on Computational Learning Theory*, 1995.

[8] B. Fröba and C. Küllbeck. Orientation template matching for face localization in complex visual scenes. In *International Conference on Image Processing ICIP2000*, pages 251–254, 2000.

[9] J. Haslam, C. J. Taylor, and T. F. Cootes. A probabilistic fitness measure for deformable template models. In E Hancock, editor, 5^{th} *British Machine Vision Conference*, pages 33–42, York, England, September 1994. BMVA Press, Sheffield.

[10] Oliver Jesorsky, Klaus J. Kirchberg, and Robert W. Frischholz. Robust face detection using the hausdorff distance. In 3^{rd} *International Conference on Audio- and Video-Based Biometric Person Authentication 2001*, 2001.

[11] J. A. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1965.

[12] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition Conference 2001*, volume 1, pages 511–518, Kauai, Hawaii, 2001.

[13] L. Wiskott, J.M. Fellous, N. Kruger, and C.von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):775–779, 1997.

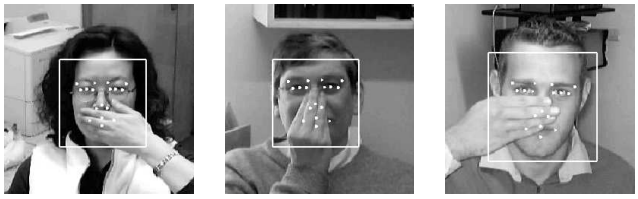


Figure 10: Example feature searches on partially occluded faces