A Multi-Stage Approach to Facial Feature Detection

David Cristinacce, Tim Cootes and Ian Scott
Dept. Imaging Science and Biomedical Engineering
University of Manchester, Manchester, M13 9PT, U.K.
david.cristinacce@stud.man.ac.uk

Abstract

We describe a novel shape constraint technique which is incorporated into a multi-stage algorithm to automatically locate features on the human face. The method is coarse-to-fine. First a face detector is applied to find the approximate scale and location of the face in the image. Then individual feature detectors are applied and combined using a novel algorithm known as Pairwise Reinforcement of Feature Responses (PRFR). The points predicted by this method are then refined using a version of the Active Appearance Model (AAM) search, which is tuned to edge and corner features. The final output of the three stage algorithm is shown to give much better results than any other combination of methods. The method outperforms previous published results on the BIOID test set [11].

1 Introduction

Accurate localisation of facial features is important in many computer vision applications. For example in face recognition, accurate feature finding is necessary to compare two facial images. Facial feature finding can also be used to track the facial expressions of an actor to automate the creation of computer graphic characters in films or computer games.

In our system, the face is located using the boosted cascaded classifier method due to Viola and Jones [17]. The whole face region predicts approximate locations for each facial feature. These regions are searched using a suitable local detector. We then use PRFR to combine the resulting candidates using pairwise probabilistic constraints, which encode the reliability of each local detector. The predicted points are then refined using the Active Appearance Model (or AAM). The AAM was originally developed by Cootes *et al.* [2]. However, the AAM method used in this paper is a variation due to Scott *et al.* [16], which models edge and corner features instead of normalised pixel values. The edge/corner AAM [16] is shown to outperform the original AAM formulation [2]. Using PRFR to initialise the edge/corner AAM is shown to give superior results to using the edge/corner AAM alone.

2 Background

The task of facial feature location has generally been addressed by algorithms that combine shape and texture modelling. For example, Burl *et al.* [1] use multi-scale Gaussian

BMVC 2004 doi:10.5244/C.18.30

derivative filters to detect facial features and then select the combination of features which represent the most likely instance of a statistical shape model [6]. A similar approach is adopted by Yow and Cipolla [18], except that shape is modelled using a grouping method based on belief networks [14]. Hamouz *et al.* [9] use Gabor filters and test triplets of appropriate configurations of features using a SVM model of facial appearance [13].

An alternative strategy for finding facial features is to treat face finding and feature finding as two separate tasks. This coarse-to-fine approach is adopted by Jesorsky *et al.* [11], who use a three stage method to find eye points. The first stage detects the whole face using the Hausdorff distance [15] between edges found in the image and a model of face edge locations. The second stage uses a smaller model of the eyes. The third stage uses a Multi-Layer Perceptron (MLP) to refine the eye pupil locations. Similarly Feris *et al.* [7] describe a two stage approach to facial feature finding based on Gabor Wavelet Networks (GWNs). The first stage matches to the whole face, whilst the second stage matches to individual features.

Another influential approach is the AAM search algorithm due to Cootes *et al.* [2]. However, the AAM is only suitable for local search. The AAM combines shape and texture in a PCA space, then searches a new image iteratively by using the texture error to drive the model parameters. Given a good enough initialisation the AAM converges to the correct solution, but is otherwise prone to local minima.

The approach described in this paper combines the robustness of the Boosted Cascade Face Detector [17], estimates feature locations using a novel shape constrained detection technique and then refines the feature points using a variation of the AAM due to Scott *et al.* [16]. The method is shown to give improved results compared to the authors' previous work (see Section 5.2) and outperform previous published results on the BIOID test set (see Section 5.3).

3 Methodology

3.1 Face Finding

The face is localised in the image by applying the Boosted Cascade Face Detector due to Viola and Jones [17]. This algorithm utilises a boosting method known as AdaBoost [8] to select and combine a set of features, which can discriminate between face and non-face image regions. The detector is run over a test image and the image window with the highest face score[1] deemed to be the location of the face in the image.

3.2 Feature Detectors

Detectors are built for 17 facial features using a manually labelled training set consisting of 1055 images collected in our lab. An example marked up face is shown in Figure 1(a). Images patches are extracted around each manually labelled point (excluding the chin and temples) and used to train a Boosted Cascade Detector for each individual feature. Example training patches are shown in Figure 1(b). The patches are sampled 5 times with small random rotations and scale changes, to provide 5275 positive training examples for each feature detector.

[1]Calculated by summing the classifier scores from each level of the cascade

(a) Example training image

(b) Example feature detector training patches

Figure 1: Example of feature patch training set

During training a bounding box is computed on the range of each feature location within the region found by the face detector (for successful searches). Given the region computed by the face detector, feature detection can then proceed by merely searching within the bounded regions and the best match taken as the location of each feature. However, as shown in the authors' previous work [4], such an approach does not work well. Search accuracy can only be improved by employing a shape constraint to force the configuration of points returned by the feature detectors to form a valid face shape [4] [5]. Typically this is achieved using a statistical shape model [6].

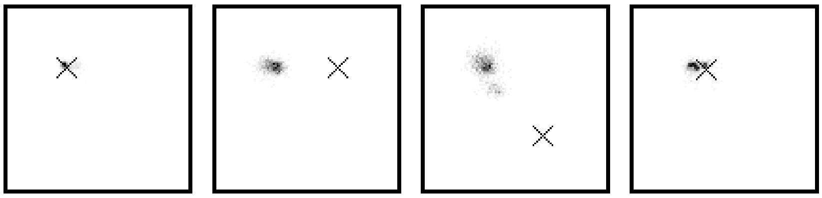
3.3 PRFR Model

In this paper a novel shape constraint is employed, known as Pairwise Reinforcement of Feature Responses (PRFR). This method does not use an explicit shape model, rather it models shape implicitly by learning the pairwise distribution of all true feature locations relative to the best match of each individual feature detector. When searching, the location of each feature is predicted by multiple detectors. The combination of multiple predictions makes the final prediction of each feature point more robust compared to individual feature search.

The pairwise distribution $P_{ij}(\mathbf{x}_i|\mathbf{x}_j)$ is defined as the distribution of the true location of feature i given the best match for feature detector j in the reference frame defined by the whole face region. In practice we use histograms of the form $H_{ij}(\mathbf{x}_i - \mathbf{x}_j)$ as an approximation to $P_{ij}(\mathbf{x}_i|\mathbf{x}_j)$. These distributions must be learnt for all possible pairs of feature detector and true feature locations. There are 17 feature detectors, trained to search for 17 feature locations, therefore 289 (=17x17) pairwise histograms are required.

Learning of histograms is achieved by applying the global face detector, followed by unconstrained feature detection, to a verification set of face images. For each verification image, the true location of all features within the global candidate frame is recorded along with the best match of each feature detector. The ensemble of true feature locations and detector matches allows relative histograms H_{ij} to be computed for the distribution of true feature location i relative to detector j.

Relative histograms H_{ij} for the right eye pupil location, are shown in Figure 2. Each diagram plots the distribution of true feature locations relative to the best match of a feature detector (marked with a cross). For example, the spread of true right eye locations relative to a right eye detection are shown in Figure 2(a). The spread of right eye locations relative to a left eye detection are shown in Figure 2(b).



(a) Right eye detector

(b) Left eye detector

(c) Left mouth corner detector

(d) Inner right eye corner detector

Figure 2: Right eye pupil location histograms relative to the best match of four different feature detectors (black pixels indicate peaks in each histogram)

Using non-parametric histograms allows realistic pairwise statistics to be modelled and makes no prior assumptions as to the distribution of any feature location relative to any particular feature detector. For example Figures 2(c) and 2(d) show multi-modal histograms which encode variation in the right eye pupil location relative to the more noisy left mouth corner and inner right eye corner feature detectors. This information may have been lost if simpler single Gaussian modelling had been used.

One disadvantage of using histograms is that a reasonably large amount of training data is required to obtain a representative sample of feature location/feature detection pairs. The number of samples required increases with the number of histogram bins. In our experiments, 100x100 bins were used for the whole candidate frame region, trained with 500 verification faces. It may be possible to approximate the distribution histograms using a Gaussian Mixture Model (GMM), if insufficient verification data is available. This would also produce a more compact model. However, in this section we make no Gaussian assumptions.

3.4 PRFR Search

Given an order list of detections for each feature detector we wish to predict the location $\hat{\mathbf{x}}_i$ of feature i by combining feature responses with the pairwise distributions $P_{ij}(\mathbf{x}_i|\mathbf{x}_j)$ as follows:-

$$\hat{\mathbf{x}}_i = \arg\max \sum_{j=1}^{n} \sum_{t=1}^{k} P_{ij}(\mathbf{x}_i|\mathbf{q}_{jt}) \tag{1}$$

Here $\mathbf{q}_{jt}$ is the position of the t^{th} maxima in the response image for feature detector j. We sum the probabilities (effectively voting) rather than multiplying, as this generally gives more robust results. Multiplication would be appropriate if all features were independent, which in this case they are not. Note that the prior distribution $P(\mathbf{q}_{jt})$ of each feature detector is ignored here and only raw matches to the current face region are used to predict the final feature locations ($\hat{\mathbf{x}}_i$).

The first k matches of each feature detector j are used instead of just the best match. This helps to protect against spurious false matches and provides more robust results. By empirical testing a suitable value of k is found to be 3. Similar results are obtained, using any value of k in the range $(3,10)$. However taking more detections into account increases the time taken to perform PRFR.

In practice the pairwise distributions $P_{ij}(\mathbf{x}_i|\mathbf{x}_j)$ are represented by relative histograms $H_{ij}(\mathbf{x}_i - \mathbf{x}_j)$. When searching, the PRFR algorithm projects the top k feature locations from the j^{th} detector into the histogram frame. Given the feature locations $\mathbf{q}_{jt}$ the relative histogram H_{ij} can be used to predict distributions D_{ijt} of likely locations for feature i. The most likely location $\hat{\mathbf{x}}_i$ is determined by simply summing over all predicted distributions D_{ijt} and selecting the highest ranking pixel in the histogram frame. The predicted feature locations $\hat{\mathbf{x}}_i$ in the histogram frame can then be mapped back to the corresponding location in the image being searched.

3.5 AAM Refinement

The AAM algorithm [2] can also be used to predict feature locations. The method attempts to match a shape and texture model to an unseen face by adapting the parameters of a linear model which combines shape and texture. The basic search algorithm is described by Cootes *et al.* [3] and is compared with a recent variation due to Scott *et al.* [16]. In this new approach, instead of normalising the raw pixel values within the region modelled by the AAM, four values are computed for each pixel. The four values are g'_x the normalised gradient in the x direction, g'_y the normalised gradient in the y direction, e' a measure of "edgeness" and c' a measure of "cornerness". A method based on the Harris corner detector [10] is used to compute the edge e' and corner values c'. For the precise details see Scott *et al.* [16].

4 Experiments

4.1 Test Data

The accuracy of feature search is assessed by applying search algorithms to a publicly available test set known as the BIOID database[2] (which is completely independent of the training set). This data set was first used by Jesorsky *et al.* [11], 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 3.

Some faces lie very close to the edge of the image in the BIOID data set, which prevents detection using the Boosted Cascade Face Detector. To avoid such edge effects, each BIOID image was extended by replicating edge pixels to create an artificial border around each image.

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

(a) Example 1 (b) Example 2 (c) Example 3 (d) Example 4

Figure 3: Examples from the BIOID test set

4.2 Proximity Measure

To assess the accuracy of feature detection the predicted feature locations are compared with manually labelled feature points. The average point to point error (m_e) is calculated as follows.

$$m_e = \frac{1}{ns} \sum_{i=1}^{i=n} d_i \tag{2}$$

Where d_i are the point to point errors for each individual feature location and s is the known inter-ocular distance between the left and right eye pupils. Here n is the number of feature points modelled. The search error m_e computed over the 17 features shown in Figure 1 is referred to as m_{e17}. The search error can also be computed for the eye pupils and mouth corners only, when it is referred to as m_{e4}.

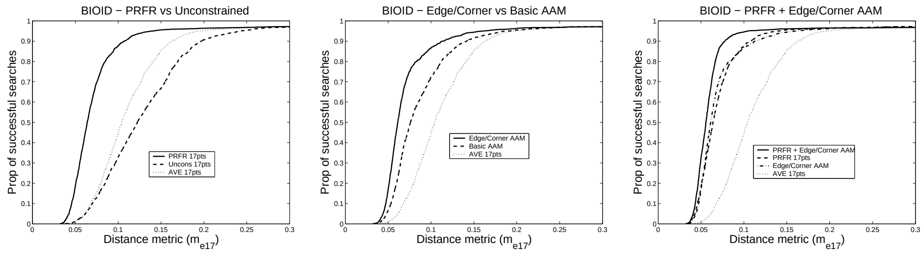
5 Results

5.1 Comparison of Individual Methods

Figure 4(a) plots the cumulative distribution of m_{e17} over the BIOID test set and shows that the PRFR algorithm outperforms both unconstrained search and average point prediction. For example, using a proximity threshold of $m_{e17} < 0.15$ the PRFR algorithm is successful in 96% of cases, compared to 85% using average point prediction. Unconstrained feature detection performs very poorly achieving a success rate of only 68%.

Figure 4(b) compares the edge/corner texture sampling AAM (described in section 3.5) with the basic AAM texture sampling method, initialised with the average points. The graph shows that with $m_{e17} = 0.15$ the edge/corner AAM achieves as success rate of 95%, compared to a success rate of 90% using the basic AAM. The edge/corner AAM is more successful than the basic AAM at all values of m_{e17}, so is clearly superior. Both AAM approaches improve on the search accuracy of average point prediction.

Figure 4(c) compares the search accuracy of the edge/corner AAM, the PRFR method and PRFR followed by edge/corner AAM refinement. Figure 4(c) shows that the PRFR followed by edge/corner AAM search is far superior to any other method. For example with $m_{e17} = 0.1$, the PRFR+AAM search is successful for 96% of the BIOID images, whilst at the same accuracy threshold, both the PRFR method alone and the edge/corner AAM initialised with average points achieve only 87% success rate.



(a) PRFR vs Unconstrained

(b) Edge/corner vs Basic AAM

(c) PRFR + Edge/corner AAM

Figure 4: Search accuracy (m_{e17}) of various methods when applied to the BIOID test set

Therefore the edge/corner AAM is able to refine the feature points predicted using the PRFR method to improve search accuracy. Initialising the edge/corner AAM using PRFR is much more effective than initialising using average point prediction. Some example search results and associated search errors using the PRFR+AAM method are shown in Figure 7.

5.2 Comparison with Previous Results

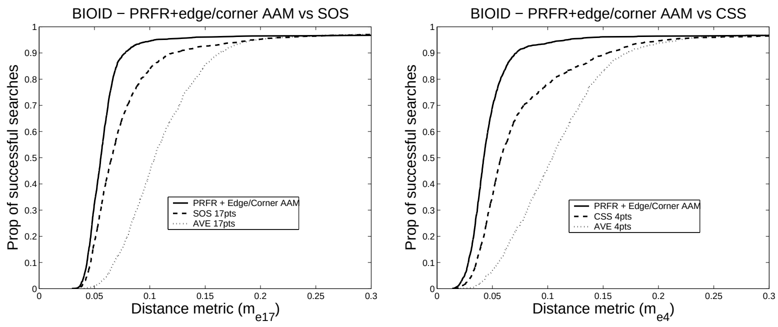
The PRFR+AAM search is compared with the authors' previous methods. For example, Figure 5(a) compares PRFR+AAM search with the search optimised search (SOS) algorithm described in [5], using the m_{e17} distance measure[3]. Figure 5(b) compares PRFR+AAM search with the combinatoric shape search (CSS) algorithm described in [4]. The CSS method only predicts the location of four features (the eye pupils and mouth corners), so the m_{e4} distance measure is used in Figure 5(b).

Figure 5(a) shows that with $m_{e17} = 0.1$ PRFR+AAM achieves a success rate of 96%, whilst the SOS algorithm only achieves 85%. The PRFR+AAM is more successful at all values of m_{e17}, so is clearly superior to SOS. Similarly, Figure 5(b) shows that PRFR+AAM is superior to the CSS algorithm at all values of m_{e4}.

5.3 Comparison with Other Published Results

Jersorsky *et al.* [11] first introduced the BIOID data set and published results on the eye pupil finding accuracy of their algorithm, which uses a face matching method based on the Hausdorff distance followed by a Multi-Layer Perceptron eye finder. Jesorsky *et al.* also present eye location accuracy results on the XM2VTS data set. Recently Hamouz *et al.* [9] also presented eye finding results on the BIOID and XM2VTS test sets, using a method which combines Gabor based feature detections to produce a list of face hypotheses, which are then tested using a SVM face model. These two methods can be compared with the PRFR+AAM algorithm for the task of eye pupil detection.

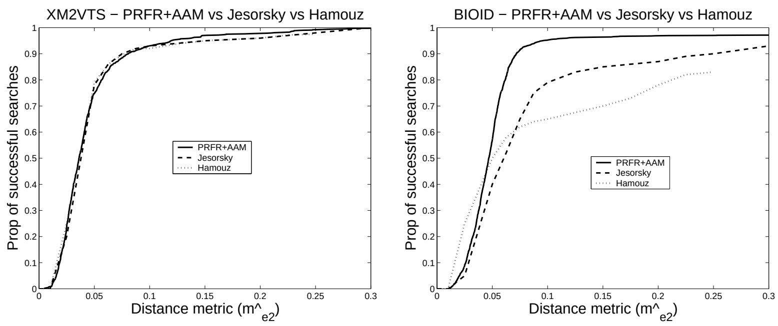
[3] Note, the equivalent graph in [5] is slightly different because it ignores cases where the global face detector fails.



(a) PRFR+AAM vs SOS (m_{e17}) (b) PRFR+AAM vs CSS (m_{e4})

Figure 5: Comparison with authors' previous search results on the BIOID test set

Jersorsky *et al.* [11] introduce a distance measure for eye search, which records the maximum point to point error over both eye point predictions, normalised by the known inter-ocular separation. We refer to this distance measure as $\hat{m}_{e2}$. Figure 6(a) plots $\hat{m}_{e2}$ for the first two sessions of the XM2VTS data set [12], which consists of 1180 images. Similarly Figure 6(b) plots $\hat{m}_{e2}$ for all three methods on the BIOID data set.



(a) XM2VTS (b) BIOID

Figure 6: Comparison with previously published results for eye finding on the XM2VTS [12] and BIOID test sets [11]

Figure 6(b) shows that when applied to the BIOID images, the PRFR+AAM is more successful than the Jesorsky method for all values of $\hat{m}_{e2}$. For example with $\hat{m}_{e2} = 0.1$, the PRFR+AAM search finds 96% of faces successfully compared to 79% using the Jesorsky approach. The Hamouz[4] method is more likely to find eye pupils very accurately (e.g.

[4]Here we take the best face match only, *not* the top 30 hypotheses, which are used by Hamouz *et al.* [9] for face verification.

$\hat{m}_{e2} < 0.05$), but is not very robust, sometimes failing to find the face completely and is the worst performing method on the BIOID data set for $\hat{m}_{e2} > 0.1$.

However the results are very different for the XM2VTS data set. Figure 6(a) shows that the accuracy of all three eye finding methods are very similar on the cleaner XM2VTS images. With $\hat{m}_{e2} = 0.1$, the Hausdorff+MLP search, Hamouz approach and PRFR+AAM achieve a success rate of around 93% and give very similar performance for all values of $\hat{m}_{e2}$. This indicates that the Hausdorff+MLP and Hamouz methods work well on relatively clean images under controlled conditions (e.g. the XM2VTS data set), but are less successful on the more complicated BIOID data set. The multi-stage Boosted Cascade Face Detector + PRFR+AAM search can find eye pupils reliably on both data sets.

The multi-stage approach is reasonably efficient requiring $\sim$1400ms to search a BIOID image using a relatively elderly PC (500Mhz PII processor). The PRFR step is the most time consuming operation requiring $\sim$800ms, due to the summation of 867 histograms (17*17*3) when predicting feature locations.

6 Summary and Conclusions

A multi-stage approach to facial feature detection has been presented, combining the Boosted Cascade Face Detector [17] , a novel constrained feature detection method (PRFR) and a refinement of the predicted points using the edge/corner AAM [16]. The method is found to predict accurate feature locations on the BIOID data set (see Figure 4(c)). This three stage approach is shown to outperform the authors' previous results (see Figure 5). When used to predict eye pupil locations only, the method is shown to give superior performance to results published by Jesorsky *et al.* [11] and Hamouz *et al.* [9] on the BIOID data set (see Figure 6(b)).

The edge/corner AAM approach is shown to outperform the original AAM approach when searching the BIOID data set, initialised using average feature points predicted from the Boosted Cascade Face Detector (see Figure 4(b)). However, it is also shown that far superior results can be obtained by initialising the edge/corner AAM with points predicted by PRFR (see Figure 4(c)). This indicates that the AAM needs a very good initialisation to avoid inaccurate matching due to false minima. PRFR point prediction is much more accurate than average point prediction, so more false minima can be avoided and the overall search performance improved when using PRFR+AAM. Some example search results are shown in Figure 7.

In conclusion, the three stage method described gives accurate point predictions for 17 feature locations, is reasonably efficient and able to cope with a challenging test set, which contains unconstrained lighting variation. The technique is very applicable to the tasks of face/expression recognition and automatic labelling of human faces.

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. Statistical models of appearance for computer vision. Technical report, Dept of Imaging Science and Biomedical Engineering, February 2001.

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

[5] D. Cristinacce and T. Cootes. A comparison of shape constrained facial feature detectors. In 6^{th} *International Conference on Automatic Face and Gesture Recognition 2004, Seoul, Korea*, pages 375–380, 2004.

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

[7] R. S. Feris, J. Gemmell, K. Toyama, and V.Krüger. Hierarchical wavelet networks for facial feature localization. In 5^{th} *International Conference on Automatic Face and Gesture Recognition 2002, Washington, USA*, Washington D.C. , USA, May 2002.

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

[9] M. Hamouz, J. Kittler, J. K. Kämäräinen, and H. Kälviöinen. Affine-invariant face detection and localization using gmm-based feature detectors and enhanced appearance model. pages 67–72, 2004.

[10] C. Harris and M. Stephens. A combined corner and edge detector. In 4^{th} *Alvey Vision Conference*, pages 147–151, 1988.

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

[12] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. Xm2vtsdb: The extended m2vts database. In *Proc. 2nd Conf. on Audio and Video-based Biometric Personal Verification*. Springer Verlag, 1999.

[13] E. Osuna, R. Freund, and F. Girosi. Training support vector machines: An application to face detection. In *Computer Vision and Pattern Recognition Conference 1997*, 1997.

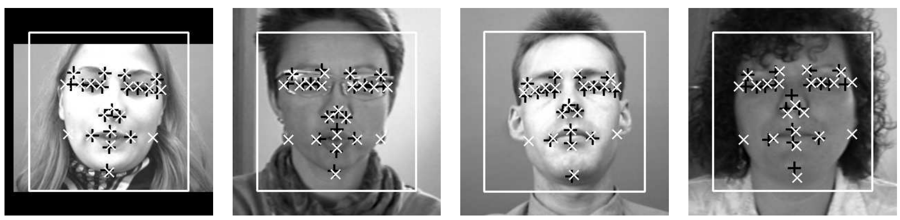
[14] J. Pearl. *Probablistic Reasoning in Intelligent Systems*. Morgan Kaufman, San Mateo, California, 1988.

[15] W. Rucklidge. Efficient visual recognition using the hausdorff distance. *Lecture Notes in Computer Science*, 1173, 1996.

[16] I. M. Scott, T. F. Cootes, and C. J. Taylor. Improving appearance model matching using local image structure. In *Information Processing in Medical Imaging*, 18^{th} *International Conference*, pages 258–269, July 2003.

[17] 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.

[18] K.C. Yow and R. Cipolla. A probabilistic framework for perceptual grouping of features for human face detection. In 2^{nd} *International Conference on Automatic Face and Gesture Recognition 1996*, pages 16–?, Killington, Vermont, USA, 1996. IEEE.



(a) $m_{e17} = 3.93\%$ (b) $m_{e17} = 4.45\%$ (c) $m_{e17} = 5.84\%$ (d) $m_{e17} = 7.14\%$

Figure 7: Example searches and search errors (m_{e17}), using PRFR+edge/corner AAM on the BIOID data set. Here “+”= manually labelled ground truth and “x”= points predicted using PRFR+edge/corner AAM search.