Feature Detection and Tracking with Constrained Local Models

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

Abstract

We present an efficient and robust model matching method which uses a joint shape and texture appearance model to generate a set of region template detectors. The model is fitted to an unseen image in an iterative manner by generating templates using the joint model and the current parameter estimates, correlating the templates with the target image to generate response images and optimising the shape parameters so as to maximise the sum of responses. The appearance model is similar to that used in the AAM [1]. However in our approach the appearance model is used to generate likely feature templates, instead of trying to approximate the image pixels directly. We show that when applied to human faces, our Constrained Local Model (CLM) algorithm is more robust and more accurate than the original AAM search method, which relies on the image reconstruction error to update the model parameters. We demonstrate improved localisation accuracy on two publicly available face data sets and improved tracking on a challenging set of in-car face sequences.

1 Introduction

This paper describes a method of modelling a class of objects with a distinct set of corresponding features, for example the human face. The model is matched to new instances of an object using an iterative template generation and shape constrained search technique.

The template model and generation of feature templates is similar to that used in the popular Active Appearance Model (AAM) algorithm [1]. However the Constrained Local Model (CLM) approach described here learns the variation in appearance of a set of template regions surrounding individual features instead of triangulated patches and uses a different search algorithm.

Given current image points, the template generation proceeds by fitting the joint model of shape and appearance to regions sampled around each feature point. The current feature templates are then applied to the search image using normalised correlation. This generates a set of response surfaces. The quality of fit of the model is optimised using the Nelder-Meade simplex algorithm [10] to drive the parameters of the shape model in order to maximise the sum of responses at each point. Given a new set of candidate feature locations the templates are regenerated and the search proceeds iteratively.

This Constrained Local Model (CLM) approach, summarised in Figure 3, is shown to be robust, relatively quick and provide superior tracking performance compared to the Active Appearance Model matching method [1], when applied to human faces. The CLM template update method is also compared with previous shape constrained search methods [3] [4].

2 Background

There are many examples of computer vision techniques that combine both shape and texture to build models and match to unseen images [1][2][4] [5][7].

One popular approach is the Active Appearance Model (AAM) algorithm [1], which uses a combined statistical model of shape and texture. The AAM searches by using the texture residual between the model and the target image to predict improved model parameters to obtain the best possible match. In this paper, we propose a similar template appearance model, but use a more robust shape constrained search technique.

Given a set of region templates the CLM search proceeds in a similar manner to Shape Optimised Search (SOS) method described in [3] and the Template Selection Tracker (TST) method described in [4]. However in [3] the feature templates are fixed during search, whilst here we propose an iterative scheme which generates appropriate templates given the current feature points and target image. In [4] the feature templates are updated by selecting example templates from the training set using a nearest neighbour approach. The CLM uses a statistical model learnt from the training set to generate templates.

The search method is also related to the earlier Active Shape Model (ASM) algorithm [2], however the ASM also uses fixed templates and only uses the shape model to update the feature locations after computing the best match of each detector. Our approach utilises the whole response surface, allowing a better balance between shape and feature response.

Our algorithm is similar to the recent SMAT algorithm described by Dowson and Bowden [5]. The SMAT method tracks an object given an initialisation and generates new templates from a clustered set of templates sampled from previous frames and also uses a shape model to constrain feature configurations, but does not form a combined model of shape and texture. The CLM generates appropriate templates using an appearance model, learnt from a fixed training set. In comparison the CLM is unable to generate false templates (from false matches) and can be applied to search static images, not just video, but at the expense of requiring a manually labelled training set.

An elegant method of combining feature responses and shape constraints is due to Felzenswalb [7]. This Pictorial Structure Matching (PSM) approach is very efficient due to the use of pairwise constraints and a tree structure. However PSM is mainly a global search method and does not use a full shape model.

In this paper we use the Viola and Jones [12] face detector to find the face in the image. Within the detected face region we apply smaller Viola and Jones feature detectors constrained using the PSM algorithm [7], to compute initial feature points. We then refine these feature points using CLM local search and compare the results with the AAM [1], TST [4] and SOS [3] algorithms.

Figure 1: Training Examples

3 Methodology

3.1 Constrained Local Appearance Models

A joint shape and texture model is built from a training set of 1052 manually labelled faces (see Figure 1) using the method of Cootes *et al.* [2]. However in this new approach the texture sampling method is different. A training patch is sampled around each feature and normalised such that the pixel values have zero mean and unit variance [1]. The texture patches from a given training image are then concatenated to form a single grey value vector. The set of grey scale training vectors and normalised shape co-ordinates are used to construct linear models, as follows.

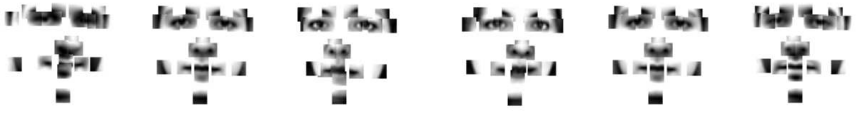
$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}_s \mathbf{b}_s \qquad \mathbf{g} = \bar{\mathbf{g}} + \mathbf{P}_g \mathbf{b}_g \tag{1}$$

Where $\bar{\mathbf{x}}$ is the mean shape, $\mathbf{P}_s$ is a set of orthogonal modes of variation and $\mathbf{b}_s$ is a set of shape parameters. Similarly $\bar{\mathbf{g}}$ is the mean normalised grey-level vector, $\mathbf{P}_g$ is a set of orthogonal modes of variation and $\mathbf{b}_g$ is a set of grey-level parameters. The shape and template texture models are combined using a further PCA to produce one joint model. The joint model has the following form.

$$\mathbf{b} = \mathbf{P}_c \mathbf{c} \qquad \text{where} \qquad \mathbf{P}_c = \begin{pmatrix} \mathbf{P}_{cs} \\ \mathbf{P}_{cg} \end{pmatrix} \quad \& \quad \mathbf{b} = \begin{pmatrix} \mathbf{W}_s \mathbf{b}_s \\ \mathbf{b}_g \end{pmatrix} \tag{2}$$

Here $\mathbf{b}$ is the concatenated shape and texture parameter vector, with a suitable weighting $\mathbf{W}_s$ to account for the difference between shape and texture units (see [1]). $\mathbf{c}$ is a set of joint appearance parameters. $\mathbf{P}_c$ is the orthogonal matrix computed using PCA, which partitions into two separate matrices $\mathbf{P}_{cs}$ and $\mathbf{P}_{cg}$ which together compute the shape and texture parameters given a joint parameter vector $\mathbf{c}$.

By varying the first two parameters of $\mathbf{c}$ the first two modes of variation for the joint appearance model can be computed, as shown in Figure 2.



(a) Mode 1 (b) Mode 2

Figure 2: Joint Modes of Shape and Texture Variation ($\pm$3std)

[1] The face regions from the training images are resampled to a fixed sized rectangle to allow for scale changes

Given the joint model and an unseen image with a set of initial feature points, the joint model can be fitted to the image by estimating the shape, texture and joint parameters as described in [1]. Given the joint parameter **c**, a set of templates with a shape approximating the current feature locations can be computed using Equations 1 and 2. For examples of feature templates generated using this method see Figure 8.

3.2 Shape Constrained Local Model Search

Given a set of initial feature points, the joint shape and texture model in in Section 3.1 is used to generate a set of grey value texture patches. The templates are applied to the search image and response images computed. Let (X_i, Y_i) be the position of feature point i and $I_i(X_i, Y_i)$ be the response of the i^{th} feature template at that point. The positions can be concatenated into a vector **X**,

$$\mathbf{X} = (X_1, \ldots, X_n, Y_1, \ldots, Y_n)^T \tag{3}$$

Where **X** is computed from the shape parameters $\mathbf{b}_s$ and a similarity transformation T_t from the shape model frame to the response image frame. **X** is calculated as follows.

$$\mathbf{X} \approx T_{\mathbf{t}}(\bar{\mathbf{x}} + \mathbf{P}_s \mathbf{b}_s) \tag{4}$$

The parameters of the similarity transform, $T_{\mathbf{t}}$ and, shape parameters $\mathbf{b}_s$ are concatenated into $\mathbf{p} = \left(\mathbf{t}^T | \mathbf{b}_s^T\right)^T$. Therefore **X** can be represented as a function of **p**. Given a starting value for **p** the search proceeds by optimising a function $f(\mathbf{p})$ based on the image response surfaces I_i and the statistical shape model learnt from the training set. The objective function we use is

$$f(\mathbf{p}) \quad = \quad \sum_{i=1}^{n} I_i(X_i, Y_i) + K \sum_{j=1}^{s} \frac{-b_j^2}{\lambda_j} \tag{5}$$

The second term is an estimate of the log-likelihood of the shape given shape parameters b_j and eigenvalues λ_j. This log-likelihood follows the approach of Dryden [6] in assuming the b_j are independent and Gaussian distributed. The parameter K is a weight determining the relative importance of good shape and high feature responses. The value of K can be determined by computing the ratio of $\sum_{i=1}^{n} I_i(X_i, Y_i)$ and $\sum_{j=1}^{s} \frac{b_j^2}{\lambda_j}$ when applied to a verification set with human labelled ground truth. The optimisation of $f(\mathbf{p})$ is performed using the Nelder-Meade simplex algorithm [10].

3.3 Search Algrorithm

The CLM search algorithm (see Figure 3) combines the methods described in Sections 3.1 and 3.2 and proceeds as follows:-

1. Input an initial set of feature points.
2. Repeat:-
   (a) Fit the joint model to the current set of feature points to generate a set of templates (see Section 3.1).

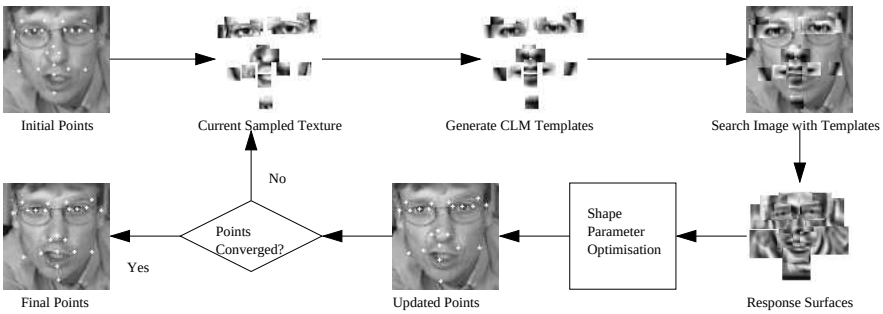


Figure 3: CLM search algorithm

(b) Use the shape constrained search method (see Section 3.2) to predict a new set of feature points.

Until Converged.

When tracking the initial points are propagated from the previous frame. On a new sequence (or if tracking failed) a global search can be used.

3.4 Differences between CLM, TST, SOS and AAM methods

The search procedure described in Section 3.2 is fundamentally different from the AAM [1]. However the joint model of appearance and texture (see Section 3.1) has the same form as the AAM. The CLM appearance model takes the form of rectangular regions around each feature, which allows the non-linear shape constrained search described in Section 3.2.

The CLM constrain local model search (see Section 3.2) is the same as described in the author's previous work [3] [4]. However the Template Selection Tracker (TST) method [4] uses nearest neighbour matching to select the most appropriate templates. The CLM uses a joint model fitted to the current feature points to generate templates. In contrast the Shape Optimised Search (SOS) method [3] uses fixed templates, which do not change during search.

4 Experiments

4.1 Test Criteria

The criteria for success is the distance of the points computed using automated methods compared to manually labelled ground truth. The distance metric is shown in Equation 6.

$$m_e = \frac{1}{ns}\sum_{i=1}^{i=n} d_i \tag{6}$$

Here d_i are the Euclidean point to point errors for each individual feature location and s is the ground truth inter-ocular distance between the left and right eye pupils. $n = 17$ as only the internal feature locations around the eyes, nose and mouth are used to compute

the distance measure. The five feature points on the edge of the face (see Figure 1) are ignored for evaluation purposes.

4.2 Detection Experiments

The localisation accuracy of the CLM, TST, SOS and AAM algorithms is tested by applying the methods to the publicly available BIOID [8] and XM2VTS [9] data sets. Note that these images are completely independent of the training images which contains different people imaged under different conditions (see Figure 1).

Our procedure for finding initial facial feature locations in a static image is to apply our implementation of the Viola and Jones face detector [12], then apply similar smaller region detectors within the face candidate region, which are constrained using the Pictorial Structure Matching (PSM) approach due to Felzenswalb [7]. This method produces a set of points from which to initialise the CLM, TST, SOS or AAM algorithms. Six different procedures are evaluated as follows:-

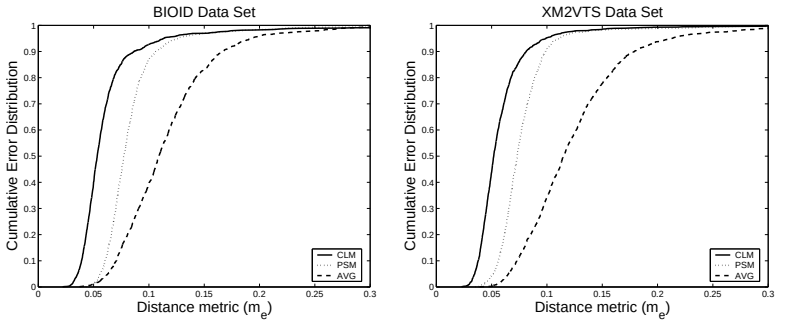
- AVG - Average points within the global Viola and Jones face detector.
- PSM - Pictorial Structure Matching points found within the Viola and Jones candidate face region.
- AAM - Active Appearance Modelalgorithm initialised with PSM points [2].
- SOS - Shape Optimised Search [3] initialised with the PSM points, using fixed templates (the mean of the texture model)
- TST - Template Selection Tracker [4] initialised with the PSM points and updating the templates.
- CLM - Constrained Local Model initialised with the PSM points and updating the templates.

Results of applying these methods to the BIOID and XM2VTS data sets are shown in Figure 4.

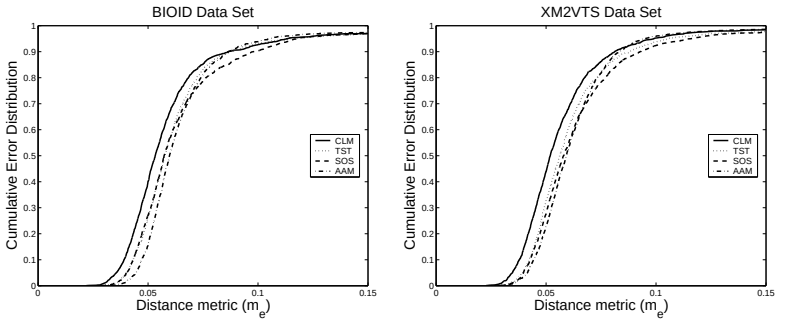
Figures 4(a) and 4(b) show that the least successful method is simply using the average points from the global face detector with no local search (AVG dashed line). However the global face detector alone is reasonably successful finding 95% of facial feature points within 20% of the inter-ocular separation on the BIOID data set and 92% on the XM2VTS image database. Given the detected face region the feature localisation accuracy is improved on both data sets by applying smaller feature detectors and using the PSM [7] constraint method (see the dotted line). Figure 4(a) and Figure 4(b) show a large improvement when using the CLM approach (solid line) to refine the PSM starting points (dotted line).

Figure 4(c) and Figure 4(d) compare the results of applying the CLM, TST, SOS and AAM methods to refine the PSM points. On both the BIOID and XM2VTS the CLM search method (solid line) gives the most accurate result relative to the human labelled ground truth. The SOS (dashed line) search is the worst performing method. The TST (dotted) and AAM (dot-dash) give comparable results.

[2] Note that the AAM formulation we use is actually the edge/corner AAM due to Scott *et al.* [11] which has been shown to be more effective than the basic texture method



(a) BIOID Results - CLM vs PSM vs AVG (b) XM2VTS Results - CLM vs PSM vs AVG



(c) BIOID Results - CLM vs TST vs SOS vs AAM (d) XM2VTS Results - CLM vs TST vs SOS vs AAM

Figure 4: Cumulative distribution of point to point measure (m_e) on BIOID and XM2VTS data sets

Figure 5 shows an example of the CLM search converging to a successful search solution on one example from the BIOID data set. The templates steadily change to resemble the image being searched.

4.3 Tracking Experiments

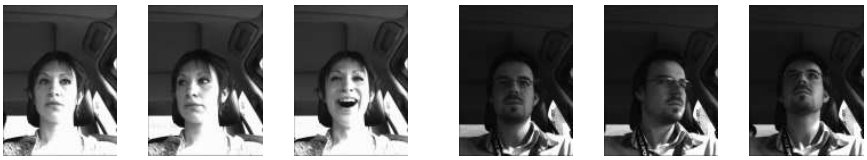
The CLM algorithm automatically adjusts the feature templates to match the current image. Therefore it is a natural tracking method in the sense that the templates learn to match the image, but are also constrained by joint shape and texture model to remain plausible feature templates.

We test the CLM method by applying it to two sequences of people driving in cars. The test sequence involves a large amount of lighting variation and head movement and is thus a challenging data set. Each sequence contains approximately 1000 frames (taken at 10fps). See Figure 6 for example frames from the two test sequences.

The face rotates out of plane at some point in all three sequences. Therefore we use a quality of fit measure to test when the face has been lost and re-initialise by searching subsequent frames with the global face detector. The quality of fit measure used for the CLM , TST and SOS method is the shape constrained response score (see Equation 5). The AAM fit quality is the sum of residuals of the texture model fitted to the image.

(a) Start (b) After It1 (c) After It2 (d) Final Pts

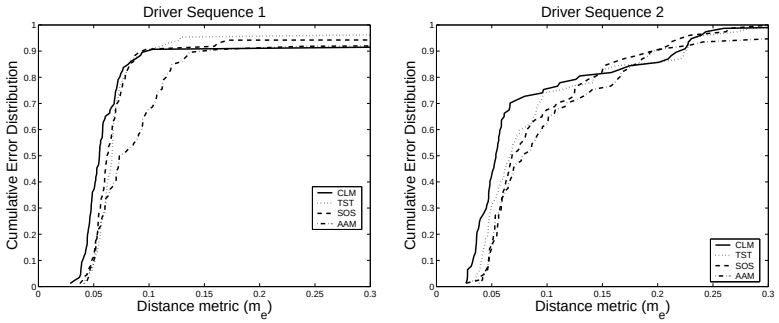
Figure 5: Evolution of CLM templates when searching a static image



(a) Sequence 1 (b) Sequence 2

Figure 6: Examples from car driver sequences

To provide ground truth for our experiments every 10^{th} frame (i.e once a second) is labelled by a human operator, provide all the facial features are visible. The distance measure for each labelled frame is Equation 6, unless the labelled face is undetected in the image, when the distance is recorded as infinite. The results of applying this detect/track scheme to the driver sequences are shown in Figure 7.



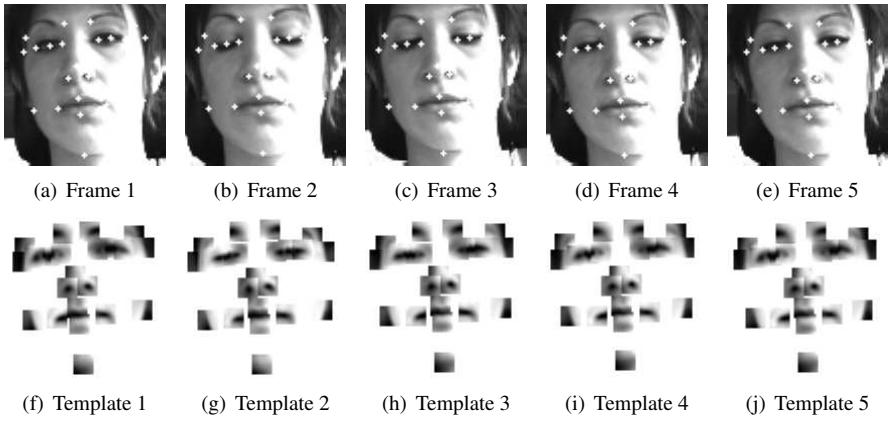
(a) Sequence 1 (b) Sequence 2

Figure 7: Cumulative distribution of point to point error measure

The graphs in Figure 7 show that the CLM (solid line), TST (dotted) and SOS (dashed) shape constrained search techniques (see Section 3.2) generally give better tracking performance than the AAM search algorithm (dot-dash). This is probably due to the shape constrained search being more robust to local minima compared to the AAM. Tracking results are also dependent on the ability of the models to detect tracking failure and re-initialise using the global search when required.

An example of the CLM templates used to track the face in Sequence 1 are shown in

Figure 4. The short series shows a period in the video where the driver is blinking. Note that the CLM successfully tracks the sequence during this period (see the white crosses on Figures 8(a) to 8(e)) and the template patches for the eyes are able to model the blinking texture of the underlying image, whilst the remaining templates remain unchanged (see Figures 8(f) to 8(j)).



(a) Frame 1 (b) Frame 2 (c) Frame 3 (d) Frame 4 (e) Frame 5

(f) Template 1 (g) Template 2 (h) Template 3 (i) Template 4 (j) Template 5

Figure 8: CLM templates during blinking

4.4 Timings

When searching static images the global search followed by local feature detection requires ∼120ms. The time to apply the CLM or AAM is less than ∼120ms, but in both cases depends on the number of iterations required to converge. The CLM requires ∼40ms per template generating iteration. When searching static images two or three iterations are usually required. However when tracking, for most frames only one iteration is usually required. Therefore when searching a static image from the BIOID data set (384x286 pixels), with a P4 3Ghz processor the full search time is ∼240ms or 4 frames per second, but when tracking only the CLM search time drops to ∼40ms or 25 frames per second. Note our face models are trained on many different people and can thus match to almost anyone. The tracked face is not included in the training set.

5 Summary and Conclusions

We have presented a novel algorithm to model a deformable object, which we refer to as the Constrained Local Model search (CLM). The method of building the CLM model is similar to the AAM [1] approach, but instead of modelling the whole object region we model a set of local feature templates. The feature templates are then matched to the image using an efficient shape constrained search of the template response surfaces. We show that when applied to faces the CLM is more accurate and more robust that the original AAM search.

We have shown that the CLM outperforms the TST, SOS and AAM method when applied to the the BIOID and XM2VTS static data sets using initial feature points found by a global search method (see Section 4.2). We have also shown that the CLM is more

robust than the AAM when used to track faces in a set of in-car driver sequences (see Section 4.3). The CLM is found to have similar computational efficiency to the AAM, able to track faces at approximately 25 frames per second (see Section 4.4).

Future work will involve extending our approach to model gradients rather than normalised pixel values, as this has been shown to improve the AAM search [11]. We may also investigate automatic model building methods, as presently the set of features and template region sizes are picked by hand, which may well be sub-optimal. Additionally the CLM local texture models and shape constraint search method may easily be extended to 3D for use in high dimensional medical data.

In conclusion the CLM method is a simple, efficient and robust alternative to the AAM algorithm, which models the appearance of a set of feature templates, instead of the image pixel values. We demonstrate that the new CLM approach outperforms the AAM when applied to human faces.

References

[1] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In 5^{th} *European Conference on Computer Vision 1998, Freiburg, Germany*, volume 2, pages 484–498, 1998.

[2] T. F. Cootes and C. J.Taylor. Active shape models. In 3^{rd} *British Machine Vision Conference 1992*, pages 266–275, 1992.

[3] 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.

[4] D. Cristinacce and T. Cootes. Facial feature detection and tracking with automatic template selection. In 7^{th} *International Conference on Automatic Face and Gesture Recognition 2006, Southampton, UK*, pages 429–434, 2006.

[5] N. Dowson and R. Bowden. Simultaneus modeling and tracking (smat) of feature sets. In 23^{rd} *Computer Vision and Pattern Recognition Conference 2005, San Diego, USA*, pages 99–105, 2005.

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

[7] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61:55–79, 2005.

[8] 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, Halmstad, Sweden*, pages 90–95, 2001.

[9] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. Xm2vtsdb: The extended m2vts database. In 2^{nd} *International Conference on Audio- and Video-Based Biometric Person Authentication 1999, Washington DC, USA*, pages 72–77, 1999.

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

[11] 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.

[12] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In 19^{th} *Computer Vision and Pattern Recognition Conference 2001, Hawaii, USA*, volume 1, pages 511–518, Kauai, Hawaii, 2001.