Coupled-View Active Appearance Models

T.F. Cootes, G.V. Wheeler, K.N. Walker, C.J. Taylor
Dept. Imaging Science and Biomedical Engineering
University of Manchester, Manchester M13 9PT U.K.
t.cootes@man.ac.uk

Abstract

This paper [1] describes building models which represent the appearance of an object (in particular, a face) as seen from two or more different viewpoints simultaneously. A small number of 2D linear statistical models are sufficient to capture the shape and appearance of a face from a wide range of viewpoints. Given multiple images of the same face we can learn a coupled model describing the relationship between the frontal appearance and the profile of a face. This relationship can be used to predict new views of a face seen from one view. Such a coupled model can be used to constrain search algorithms which seek to locate a face in multiple views simultaneously, leading to more robust results than searching each view independently.

1 Introduction

This paper describes building coupled models which represent the appearance of a face as seen from two different view-points. The majority of work on face tracking and recognition assumes near fronto-parallel views, and tends to break down when presented with large rotations or profile views. Three general approaches have been used to deal with this; a) use a full 3D model [18, 4, 13], b) introduce non-linearities into a 2D model [6, 14, 16] and c) use a set of models to represent appearance from different view-points [12, 2]. In this paper we explore the last approach, using statistical models of shape and appearance to represent the variations in appearance from a particular view-point and the correlations between models of different view-points.

One potential application is in face recognition systems. If a sytem relies on a single image, it can easily be fooled with a photograph. This can be prevented either by using sequences, or by using two cameras with a sufficiently different view-point. The methods described below can be used to determine the expected relationship between the different views, and should be able to distinguish a photograph from a true 3D structure.

The appearance models are trained on example images labelled with sets of landmarks to define the correspondences between images [1]. Lanitis *et.al.* [9] showed that a linear model was sufficient to simulate considerable changes in viewpoint, as long as all the modelled features (the landmarks) remained visible.

[1] This paper appears in Proc. British Machine Vision Conference, 2000. Vol.1,pp52-61

A model trained on near fronto-parallel face images can cope with pose variations of up to 45^o either side. For much larger angle displacements, some features become occluded, and the assumptions of the model break down.

It has been demonstrated [2] that to deal with full 180^o rotation (from left profile to right profile), one needs only 5 models, roughly centred on viewpoints at -90^o,-45^o,0^o,45^o,90^o (where 0^o corresponds to fronto-parallel). The pairs of models at $\pm 90^o$ (full profile) and $\pm 45^o$(half profile) are simply reflections of each other, so only 3 distinct models are required. These models can be used for estimating head pose, for tracking faces through wide changes in orientation and for synthesizing new views of a subject given a single view.

Each model is trained on labelled images of a variety of people with a range of head orientations, chosen so none of the features for that model become occluded. The different models use different sets of features (see Figure 2). Each example view can then be approximated using the appropriate appearance model with a vector of parameters, $\mathbf{c}$. We can learn the relationship between $\mathbf{c}$ and head orientation, allowing us to both estimate the orientation of any head and to be able to synthesize a face at any orientation.

There are clearly correlations between the parameters of one view model and those of a different view model. In order to learn these, we need images taken from two views simultaneously. For our experiments we achieved this using a judiciously placed mirror, giving a frontal and a profile view (Figure 1).

Figure 1: Using a mirror we capture frontal and profile appearance simultaneously

By annotating such images and matching frontal and profile models, we obtain corresponding sets of parameters. These can be analysed to produce a joint model which controls both frontal and profile appearance. Such a joint model can be used to synthesize new views given a single view. Though this can perhaps be done most effectively with a full 3D model [18], we demonstrate that good results can be achieved just with a set of 2D models. The joint model can also be used to constrain an Active Appearance Model search [3, 1], allowing simultaneous matching of frontal and profile models to pairs of images.

In the following we describe the techniques in more detail and give examples of the models, their ability to synthesize new views and to search unseen images.

2 Background

Statistical models of shape and texture have been widely used for recognition, tracking and synthesis [7, 9, 3, 17], but have tended to only be used with near

fronto-parallel images.

Moghaddam and Pentland [12] describe using view-based eigenface models to represent a wide variety of viewpoints. Our work is similar to this, but by including shape variation (rather than the rigid eigen-patches), we require fewer models and can obtain better reconstructions with fewer model modes.

Maurer and von der Malsburg [10] demonstrated tracking heads through wide angles by tracking graphs whose nodes are facial features, located with Gabor jets. The system is effective for tracking, but is not able to synthesize the appearance of the face being tracked.

Murase and Nayar [6] showed that the projections of multiple views of a rigid object into an eigenspace fell on a 2D manifold in that space. By modelling this manifold they could recognise objects from arbitrary views. A similar approach has been taken by Gong *et.al.* [15, 8] who use non-linear representations of the projections into an eigen-face space for tracking and pose estimation, and by Graham and Allinson [5] who use it for recognition from unfamiliar viewpoints.

Romdhani *et.al.* [14] have extended the Active Shape Model to deal with full 180^o rotation of a face using a non-linear model. However, the non-linearities mean the method is slow to match to a new image. They have also extended the AAM [16] using a kernel PCA. A non-linear 2D shape model is combined with a non-linear texture model on a 3D texture template. The approach is promising, but considerably more complex than using a small set of linear 2D models.

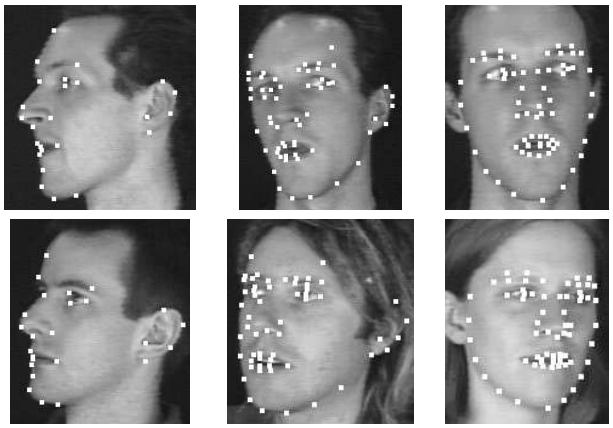
Vetter [18] has demonstrated how a 3D statistical model of face shape and texture can be used to generate new views given a single view. The model can be matched to a new image from more or less any viewpoint using a general optimisation scheme, though this is slow. Similar work has been described by Fua and Miccio [4] and Pighin *et.al.* [13]. By explicitly taking into account the 3D nature of the problem, this approach is likely to yield better reconstructions than the purely 2D method described below. However, the view based models we propose could be used to drive the parameters of the 3D head model, speeding up matching times.

3 View-Based Models of Appearance

An appearance model can represent both the shape and texture variability seen in a training set. The training set consists of labelled images, where key landmark points are marked on each example object. The training set is usually labelled manually, though automatic methods are being developed. For instance, Figure 2 shows examples of labelled images used to train the view-based face models.

Given such a set we can generate a statistical models of shape and texture variation (see [1, 3] for details). The shape of an object can be represented as a vector $\mathbf{x}$ and the texture (grey-levels or colour values) represented as a vector $\mathbf{g}$. The appearance model has parameters, $\mathbf{c}$, controlling the shape and texture according to

$$\begin{aligned} \mathbf{x} &= \bar{\mathbf{x}} + \mathbf{Q}_s \mathbf{c} \\ \mathbf{g} &= \bar{\mathbf{g}} + \mathbf{Q}_g \mathbf{c} \end{aligned} \tag{1}$$



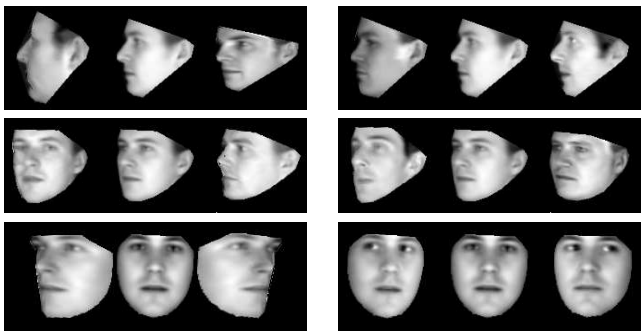
Profile Half Profile Frontal

Figure 2: Examples from the training sets for the models

where $\bar{\mathbf{x}}$ is the mean shape, $\bar{\mathbf{g}}$ the mean texture and $\mathbf{Q}_s$,$\mathbf{Q}_g$ are matrices describing the modes of variation derived from the training set.

We trained three distinct models on data similar to that shown in Figure 2. The profile model was trained on about 450 landmarked images taken of 70 different individuals from a variety of orientations. The half-profile model was trained on 82 images of 15 individual, and the frontal model on about 450 images of 70 individuals.

An example image can be synthesised for a given $\mathbf{c}$ by generating a texture image from the vector $\mathbf{g}$ and warping it using the control points described by $\mathbf{x}$. For instance, Figure 3 shows the effects of varying the first two appearance model parameters, c_1, c_2, of models trained on sets of face images, labelled as shown in Figure 2. These change both the shape and the texture component of the synthesised image.



c_1 varies ± 2 s.d.s c_2 varies ± 2 s.d.s

Figure 3: First two modes of the face models (top to bottom: profile, half-profile and frontal)

4 Predicting Pose

We assume that the model parameters are related to the viewing angle, θ, approximately as

$$\mathbf{c} = \mathbf{c}_0 + \mathbf{c}_c \cos(\theta) + \mathbf{c}_s \sin(\theta) \tag{2}$$

where $\mathbf{c}_0$, $\mathbf{c}_c$ and $\mathbf{c}_s$ are vectors estimated from training data. Here we consider only rotation about a vertical axis - head turning. Nodding can be dealt with in a similar way. We estimate the head orientation in each of our training examples, θ_i, accurate to about $\pm 10^o$. For each such image we find the best fitting model parameters, $\mathbf{c}_i$. We then perform regression between $\{\mathbf{c}_i\}$ and the vectors $\{(1, \cos(\theta_i), \sin(\theta_i))'\}$ to learn $\mathbf{c}_0$,$\mathbf{c}_c$ and $\mathbf{c}_s$.

Figure 4 shows reconstructions in which the orientation, θ, is varied in Equation 2.

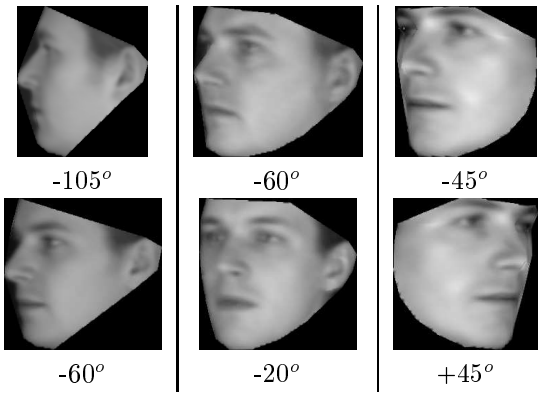


Figure 4: Rotation modes of three face models

Given a new example with parameters $\mathbf{c}$, we can estimate its orientation as follows. Let $\mathbf{R}_c^{-1}$ be the left pseudo-inverse of the matrix $(\mathbf{c}_c|\mathbf{c}_s)$ (thus $\mathbf{R}_c^{-1}(\mathbf{c}_c|\mathbf{c}_s) = \mathbf{I}_2$). Let

$$(x_a, y_a)' = \mathbf{R}_c^{-1}(\mathbf{c} - \mathbf{c}_0) \tag{3}$$

then the best estimate of the orientation is $\tan^{-1}(y_a/x_a)$. Experiments suggest the estimate is accurate to about $\pm 5^o$[2].

5 Synthesizing Rotation

Given a single view of a new person, we can find the best model match and determine their head orientation. We can then use the best model to synthesize new views at any orientation that can be represented by the model. If the best matching parameters are $\mathbf{c}$, we use Equation 3 to estimate the angle, θ. Let $\mathbf{r}$ be the residual vector not explained by the rotation model, ie

$$\mathbf{r} = \mathbf{c} - (\mathbf{c}_0 + \mathbf{c}_c \cos(\theta) + \mathbf{c}_s \sin(\theta)) \tag{4}$$

To reconstruct at a new angle, α, we simply use the parameters

$$\mathbf{c}(\alpha) = \mathbf{c}_0 + \mathbf{c}_c \cos(\alpha) + \mathbf{c}_s \sin(\alpha) + \mathbf{r} \tag{5}$$

For instance, Figure 5 shows fitting a model to a roughly frontal image and rotating it. The top example uses a new view of someone in the training set. The lower example is a previously unseen person from the Surrey face database [11].

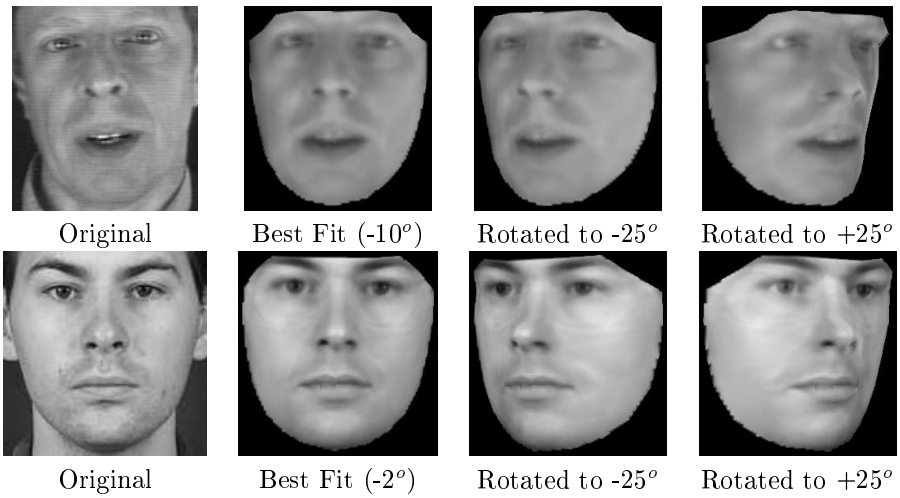


Figure 5: By fitting a model we can estimate the orientation, then synthesize new views

This only allows us to vary the angle in the range defined by the current view model. To generate significantly different views we must learn the relationship between parameters for one view model and another.

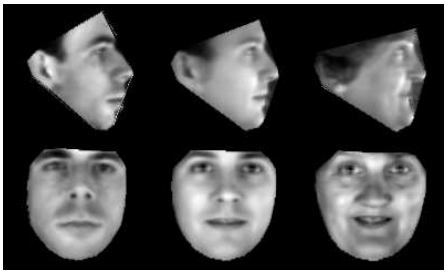
6 Coupled-View Appearance Models

Given enough pairs of images taken from different view points, we can build a model of the relationship between the model parameters in one view and those in another. Ideally the images should be taken simultaneously, allowing correllations between changes in expression to be learnt. We have achieved this using a single video camera and a mirror (see Figure 1). A looser model can be built from images taken at different times, assuming a similar expression (typically neutral) is adopted in both.

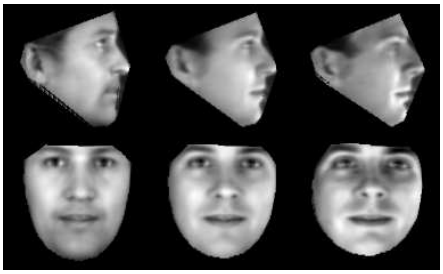
Let $\mathbf{r}_{ij}$ be the residual model parameters for the object in the i^{th} image in view j, formed from the best fitting parameters by removing the contribution from the angle model (Equation 4). We form the combined parameter vectors $\mathbf{j}_i = (\mathbf{r}_{i1}^T, \mathbf{r}_{i2}^T)^T$. We can then perform a principal component analysis on the set of $\{\mathbf{j}_i\}$ to obtain the main modes of variation of a combined model,

$$\mathbf{j} = \hat{\mathbf{j}} + \mathbf{P}\mathbf{b} \tag{6}$$

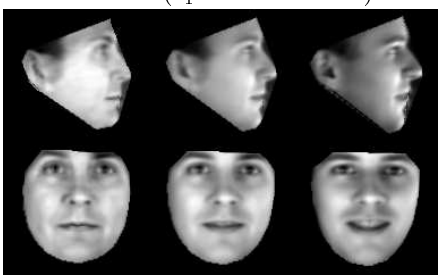
Figure 6 shows the effect of varying the first four of the parameters controlling such a model representing both frontal and profile face appearance. The modes mix changes in identity and changes in expression. For instance mode 3 appears to demonstrate the relationship between frontal and profile views during a smile.



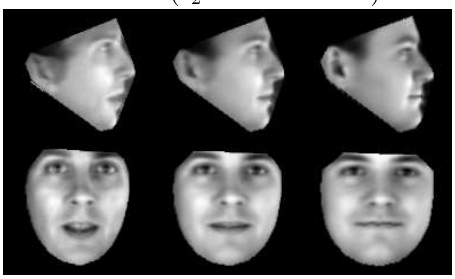
Mode 1 (b_1 varies $\pm$2s.d.s)



Mode 2 (b_2 varies $\pm$2s.d.s)



Mode 3 (b_3 varies $\pm$2s.d.s)

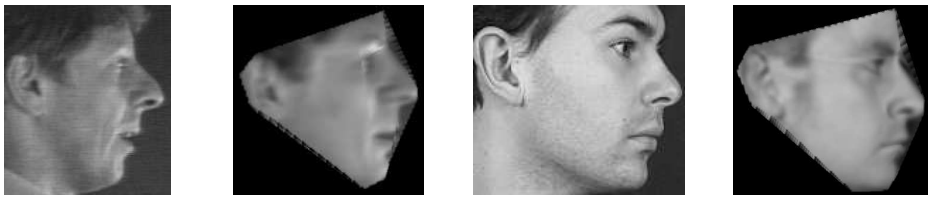


Mode 4 (b_4 varies $\pm$2s.d.s)

Figure 6: Modes of joint model, controlling frontal and profile appearance

6.1 Predicting New Views

We can use the joint model to generate different views of a subject. We find the joint parameters which generate a frontal view best matching the current target, then use the model to generate the corresponding profile view. Figures 7(a,b) show the actual profile and profile predicted from a new view of someone in the training set. In this case the model is able to estimate the expression (a half smile). Because we only have a limited set of images in which we have non-neutral expressions, the joint model built with them is not good at generalising to new people. To deal with this, we built a second joint model, trained on about 100 frontal and profile images taken from the Surrey XM2VTS face database [11]. These have neutral expressions, but the image pairs are not taken simultaneously, and the head orientation can vary significantly. However, the rotation effects can be removed using the approach described above, and the model can be used to predict unseen views of neutral expressions. Figures 7(c,d) show the actual profile and profile predicted from a new person (the frontal image is shown in Figure 5). With a large enough training set we would be able to deal with both expression changes and a wide variety of people.



a) Actual profile b) Predicted profile c) Actual profile d) Predicted profile

Figure 7: The joint model can be used to predict appearance across views (see Fig 5 for frontal view from which the predictions are made)

7 Coupled Model Matching

Given two different views of a target, and corresponding models, we can exploit the correlations to improve the robustness of matching algorithms. One approach would be to modify the Active Appearance Model search algorithm to drive the parameters, $\mathbf{b}$, of the joint model, together with the current estimates of pose, texture transformation and 3D orientation parameters. However, the approach we have implemented is to train two independent AAMs (one for the frontal model, one for the profile), and to run the search in parallel, constraining the parameters with the joint model at each step. In particular, each iteration of the matching algorithm proceeds as follows:

- Perform one iteration of the AAM on the frontal model, and one on the profile model, to update the current estimate of $\mathbf{c}_1$, $\mathbf{c}_2$ and the associated pose and texture transformation parameters.
- Estimate the relative head orientation with the frontal and profile models, θ_1, θ_2
- Use Equation 4 to estimate the residuals $\mathbf{r}_1$, $\mathbf{r}_2$
- Form the combined vector $\mathbf{j} = (\mathbf{r}_1^T, \mathbf{r}_2^T)^T$
- Compute the best joint parameters, $\mathbf{b} = \mathbf{P}^T(\mathbf{j} - \hat{\mathbf{j}})$ and apply limits to taste.
- Compute the revised residuals using $(\mathbf{r}_1'^T, \mathbf{r}_2'^T)^T = \hat{\mathbf{j}} + \mathbf{P}\mathbf{b}$
- Use Equation 2 to add the effect of head orientation back in

Note that this approach makes no assumptions about the exact relative viewing angles. If appropriate we can learn the relationship between θ_1 and θ_2 ($\theta_1 = \theta_2 + const$). This could be used as a further constraint. Similarly the relative positions and scales could be learnt.

To explore whether these constraints actually improve robustness, we performed the following experiment. We manually labelled 50 images (not in the original training set), then performed multi-resolution search, starting with the mean model parameters in the correct pose. We ran the experiment twice, once using the joint model constraints described above, once without any constraints (treating the two models as completely independent).

Table 1 summarises the results. After each search we measure the RMS distance between found points and hand labelled points, and the RMS error per pixel

Measure	Frontal Model		Profile Model	
	Coupled	Independent	Coupled	Independent
RMS Point Error (pixels)	4.8 ± 0.5	5.1 ± 0.5	3.3 ± 0.15	3.8 ± 0.3
RMS Texture Error (grey-levels)	7.9 ± 0.25	7.9 ± 0.25	8.3 ± 0.25	8.8 ± 0.4

Table 1: Comparison between coupled search and independent search

between the model reconstruction and the image (the intensity values are in the range [0,255]). The results demonstrates that in this case the use of the constraints between images improved the performance, but not by a great deal. We would expect that adding stronger constraints, such as that between the angles θ_1,θ_2, and the relative scales and positions, would lead to further improvements.

8 Discussion and Conclusions

We have demonstrated that a small number of view-based statistical models of appearance can represent the face from a wide range of viewing angles. Although we have concentrated on rotation about a vertical axis, rotation about a horizontal axis (nodding) could easily be included (and probably wouldn't require any extra models for modest rotations). We have shown that a linear model can represent the correlations between appearance in two views and that such a model can be used to predict appearance from new viewpoints given a single image of a person.

Such models can be used to constrain search when matching models to two views of an object taken simultaneously. We have treated the parameters of each model as equally important, which in this case is a reasonable approximation. However, if one view is significantly less useful, if the search algorithm fails in that view it can have a deleterious effect on the match in a different view. We will consider weighting the parameters to take this into account.

The joint model has implicitly captured the 3D structure of the object. We could use uncalibrated stereo techniques to determine the actual 3D structure it represents, though this isn't necessary for many applications.

We anticipate the approach will be useful in many applications, including driving animated avatars, calculating head pose and making face recognition systems more invariant to viewing angle.

References

[1] 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.

[2] T. F. Cootes, K. N. Walker, and C. J. Taylor. View-based active appearance models. In 4^{th} *International Conference on Automatic Face and Gesture Recognition 2000*, pages 227–232, Grenoble,France, 2000.

[3] G. Edwards, C. J. Taylor, and T. F. Cootes. Interpreting face images using active appearance models. In 3^{rd} *International Conference on Automatic Face and Gesture Recognition 1998*, pages 300–305, Japan, 1998.

[4] P. Fua and C. Miccio. From regular images to animated heads: A least squares approach. In H.Burkhardt and B. Neumann, editors, 5^{th} *European Conference on Computer Vision*, volume 1, pages 188–202. Springer, Berlin, 1998.

[5] D. Graham and N. Allinson. Face recognition from unfamiliar views: Subspace methods and pose dependency. In 3^{rd} *International Conference on Automatic Face and Gesture Recognition 1998*, pages 348–353, Japan, 1998.

[6] H.Murase and S. Nayar. Learning and recognition of 3d objects from appearance. *International Journal of Computer Vision*, pages 5–25, Jan. 1995.

[7] M. J. Jones and T. Poggio. Multidimensional morphable models : A framework for representing and matching object classes. *International Journal of Computer Vision*, 2(29):107–131, 1998.

[8] J. Kwong and S. Gong. Learning support vector machines for a multi-view face model. In T. Pridmore and D. Elliman, editors, 10^{th} *British Machine Vison Conference*, volume 2, pages 503–512, Nottingham, UK, Sept. 1999. BMVA Press.

[9] A. Lanitis, C. J. Taylor, and T. F. Cootes. Automatic interpretation and coding of face images using flexible models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):743–756, 1997.

[10] T. Maurer and C. von der Malsburg. Tracking and learning graphs and pose on image sequences of faces. In 2^{nd} *International Conference on Automatic Face and Gesture Recognition 1997*, pages 176–181, Los Alamitos, California, Oct. 1996. IEEE Computer Society Press.

[11] 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.

[12] B. Moghaddam and A. Pentland. Face recognition using view-based and modular eigenspaces. In *SPIE*, volume 2277, pages 12–21, 1994.

[13] F. Pighin, R. Szeliski, and D. Salesin. Resynthesizing facial animation through 3d model-based tracking. In 7^{th} *International Conference on Computer Vision*, pages 137–142, 1999.

[14] S. Romdhani, S. Gong, and A. Psarrou. A multi-view non-linear active shape model using kernel pca. In T. Pridmore and D. Elliman, editors, 10^{th} *British Machine Vison Conference*, volume 2, pages 483–492, Nottingham, UK, Sept. 1999. BMVA Press.

[15] J. Sherrah, S. Gong, and E. Ong. Understanding pose discrimination in similarity space. In T. Pridmore and D. Elliman, editors, 10^{th} *British Machine Vison Conference*, volume 2, pages 523–532, Nottingham, UK, Sept. 1999. BMVA Press.

[16] S.Romdhani, A.Psarrou, and S. Gong. On utilising template and feature-based correspondence in multi-view appearance models. In 6^{th} *European Conference on Computer Vision*, volume 1, pages 799–813. Springer, 2000.

[17] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[18] T. Vetter. Learning novel views to a single face image. In 2^{nd} *International Conference on Automatic Face and Gesture Recognition 1997*, pages 22–27, Los Alamitos, California, Oct. 1996. IEEE Computer Society Press.