BMVC99

# Automatic extraction of the face identity-subspace

Nicholas Costen, Tim Cootes, Gareth Edwards and Chris Taylor
Wolfson Image Analysis Unit,
Department of Medical Biophysics, University of Manchester
Stopford Building, Oxford Road,
Manchester M13 9PT, U.K.

**Abstract**

Facial variation divides into a number of functional subspaces, and ensemble-specific variation. An improved method of measuring these is presented, within the space defined by an Appearance Model. Initial estimates of the subspaces (lighting, pose, identity and expression) are obtained by Principal Components Analysis on appropriate groups of faces. An expectation-maximization algorithm is applied to image codings to maximise the probability of coding across these non-orthogonal subspaces. Ensemble specific variation is then removed by measuring the spatial predictability of the eigenvectors excluding those which are less predictable than the ensemble. These procedures significantly enhance identity recognition for a disjoint test set.

## 1 Introduction

Facial variation can be conceptually divided into a number of 'functional' subspaces – types of variation which reflect useful facial dimensions [1]. A possible selection of these face-spaces is: identity, expression (here including all transient plastic deformations of the face), pose and lighting. Other spaces may be extracted, the most obvious being age. When designing a practical face-analysis system, one at least of these subspaces must be isolated and modeled, most notably identity information for recognition tasks. Although face-images can be fitted adequately using an appearance-model space which spans the images, it is not possible to linearly separate the different subspaces [8].Thus we simultaneously apportion image weights between initial overlapping estimates of these functional spaces in proportion with the sub-space variance. This divides the faces into a set of non-orthogonal projections, allowing an approach to a set of pure, but overlapping, spaces.

These spaces, particularly the identity-space, may well be over-specific to the ensemble used to generate them. If the psychology of facial variation is examined, it appears to divide into two aspects : *general familiarity* information, which is predictable from other faces and *memorability* information, which is not predictable [2], reflecting small, discrete, easily verbalised features, for example skin blemishes or warts. This latter information will greatly increase the dimensionality of the identity space, allowing fortuitous variation in apparent identity. Grey-level codes of such variation will display

lower levels of spatial predictability than real faces. We thus measure the spatial redundancy in small, spatially-adjacent sub-samples and compare these with the samples found in the ensemble. This allows a smaller space, improving identity recognition.

## 2 Background

Facial coding requires the approximation of a manifold, or high dimensional surface, on which any face can be said to lie. This allows accurate coding, recognition and reproduction of previously unseen examples. Previous studies [3, 4, 5] have suggested that using a *shape-free* coding provides a ready means of doing this, at least the when the range of pose-angle is relatively small, perhaps $\pm 20^o$ [6]. Here, the correspondence problem between faces is first solved by finding a pre-selected set of distinctive points (corners of eyes or mouths, for example) which are present in all faces. This is typically performed by hand for a training set. Those pixels thus defined as being part of the face can be warped to a standard shape by standard grey-level interpolation techniques, ensuring that the image-wise and face-wise coordinates of a given image are equivalent. If a rigid transformation to remove scale, location and orientation effects is performed on the point-locations, they can then be treated in the same way as the grey-levels, as again identical values for corresponding points on different faces will have the same meaning.

Although these operations will linearise the space, allowing interpolation between pairs of faces, they do not give an estimate of the dimensions. Thus, the acceptability as a face of an object cannot be measured; this reduces recognition accuracy[3]. In addition, redundancies between feature-point location and grey-level values cannot be described. Both these problems can be addressed by Principal Components Analysis. This extracts a set of orthogonal eigenvectors $\mathbf{\Phi}$ and eigenvalues $\lambda$ from the covariance matrix of the images (either the pixel grey-levels, or the feature-point locations). These provide an estimate of the dimensions and range of the face-space. The weights $\mathbf{w}$ of an image $\mathbf{q}$ can then be found,

$$\mathbf{w} = \mathbf{\Phi}^T(\mathbf{q} - \bar{\mathbf{q}}) \tag{1}$$

and this enables definition of the Mahalanobis distance between faces,

$$d^2_{1\to 2} = \sum_{i=1}^{N} \frac{(w_{1i} - w_{2i})^2}{\lambda_i} \tag{2}$$

between faces $\mathbf{q}_1$ and $\mathbf{q}_2$, coding in terms of expected variation [7]. Redundancies between shape and grey-levels are removed by performing separate PCAs upon the shape and grey-levels, before the weights of the ensemble are combined to form single vectors on which a second PCA is performed [4].

This 'appearance model' allows the description of the face in terms of true variation – the distortions needed to move from one to another. However, it will code the entire space as specified by our set of images, as can be seen in Figure 1. Thus, for example, the distance between the representations of two images will be a combination of the identity, facial expression, angle and lighting conditions, including both familiarity and memorability information present in the ensemble. These categories must be separated to allow detailed analysis of the face image.

# 3 Available Data

Four sets of images, each showing major variation on one subspace alone were used for training. The sets comprised :

1. A lighting set, consisting of 5 images of a single, male individual, all photographed fronto-parallel and with a fixed, neutral expression. The sitter was lit by a single lamp, moved around his face.
2. A pose set, comprising 100 images of 10 different sitters, 10 images per sitter. The sitters pointed their heads in a variety of two-dimensional directions, of relatively consistent angle. Expression and lighting changes were minimal.
3. An expression set, with 397 images of 19 different sitters, each making seven basic expressions: happy, sad, afraid, angry, surprised, neutral and disgusted. These images showed person-specific lighting variation, and some pose variation.
4. An identity set, with 188 different images, one per sitter. These were all fronto-parallel, in flat lighting and with neutral expressions. However, as is inevitable with any large group of individuals, there was variation in the apparent expression adopted as neutral.

# 4 Appearance Model Construction

All the images had a uniform set of 122 landmarks found manually. A triangulation was applied to the points and bilinear interpolation used to warp the faces to a standard shape and size which would yield a fixed number of pixels.

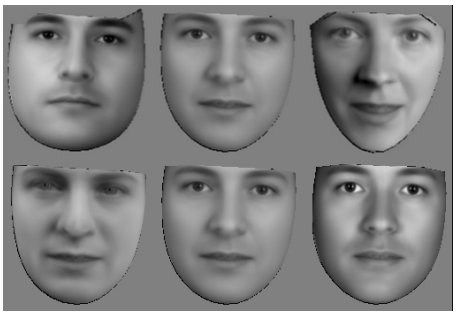

Figure 1: The first two dimensions of the face-space as defined by the appearance model. From the left, $-2s.d.$, the mean $+2s.d.$. The dimensions vary on identity, expression, pose and lighting.

Figure 2: The first two dimensions of the identity face-space as defined by the recoding algorithm. From the left, $-2s.d.$, the mean $+2s.d.$. The dimensions vary only on identity.

To minimise the effect of global lighting variation, the shape-free grey level patch $g_i$ was sampled from the $i$th shape-normalised image and normalised at each pixel $j$ to give

$$g'_{ij} = \frac{(g_{ij} - \mu_j)}{\sigma_j} \tag{3}$$

where $\mu_j$ and $\sigma_j$ are the mean and standard deviation for pixel $j$ across the ensemble.

These operations allowed the construction of an appearance model [4] coding 99.5% of the variation in the 690 images, each with 9943 pixels in the face area. This required 30 shape vectors, 455 region vectors and 374 appearance eigenvectors. For testing purposes, the feature points were found using a multi-resolution Active Appearance Model [10], constructed using the ensemble images, but without grey-level normalisation.

# 5 Sub-space Calculation

The aim of the recoding algorithm is to take account of the multiple possible explanations of the configuration of a given face. The first step is to thus to obtain approximations of these sub-spaces from the different groups of the ensemble. The overall mean was first subtracted from the appearance model parameters which described each image. Separate PCAs were then performed upon the image sets, discarding any further difference between the group and overall means. The covariance matrices for the identity and lighting subspaces were calculated as

$$\mathbf{C}_T = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{q}_i - \bar{\mathbf{q}})(\mathbf{q}_i - \bar{\mathbf{q}})^T \tag{4}$$

while the pose and expression subspaces used

$$\mathbf{C}_W = \frac{1}{n_o n_p}\sum_{i=1}^{n_p}\sum_{k=1}^{n_o}(\mathbf{q}_{ki} - \bar{\mathbf{q}}_i)(\mathbf{q}_{ki} - \bar{\mathbf{q}}_i)^T \tag{5}$$

where $n_o$ is the number of observations per individual, $n_p$ is the number of individuals, and $\bar{\mathbf{q}}_i$ is the mean of individual $i$. Although all the eigenvectors implied by the identity, lighting and expression sets were used, only the two most variable from the pose set were extracted.

## 5.1 Recoding

When considering the combination of different, probably non-orthogonal sub-spaces, if $n_s$ are used, each described by eigenvectors $\mathbf{\Phi}^{(\mathbf{j})}$ with the associated eigenvalues $\lambda^{(\mathbf{j})}$, for a given face $\mathbf{q}$ the projection out of the combined subspaces is given by

$$\mathbf{q}' = \sum_{j=1}^{n_s}\mathbf{\Phi}^{(\mathbf{j})}\mathbf{w}^{(\mathbf{j})} + \bar{\mathbf{q}}. \tag{6}$$

Assuming, as is reasonable in this case, that the different $\mathbf{\Phi}$ are not orthogonal and have more dimensions than are required to span the underlying space, there is a many-to-one relationship between $\mathbf{w}$ and $\mathbf{q}'$ and constraints must be imposed to ensure consistency of coding. One obvious constraint, used here, is that $\mathbf{w}$ be the most probable of the set producing $\mathbf{q}'$. This implies that

$$E = \sum_{j=1}^{n_s}\sum_{i=1}^{N_j}\frac{(w_i^{(j)})^2}{\lambda_i^{(j)}} \tag{7}$$

be minimised. Thus if $\mathbf{M}$ is the matrix formed by concatenating $\mathbf{\Phi}^{(\mathbf{j}=\mathbf{1},\mathbf{2}\ldots)}$ and $\mathbf{D}$ is the diagonal matrix of $\lambda^{(\mathbf{j}=\mathbf{1},\mathbf{2}\ldots)}$,

$$\mathbf{w} = (\mathbf{D}\mathbf{M}^T\mathbf{M} + \mathbf{I})^{-1}\mathbf{D}\mathbf{M}^T(\mathbf{q} - \bar{\mathbf{q}}) \tag{8}$$

and this also gives a projected version of the face

$$\mathbf{q}' = (\mathbf{D}\mathbf{M}^T)^{-1}(\mathbf{D}\mathbf{M}^T\mathbf{M} + \mathbf{I})\mathbf{w} + \bar{\mathbf{q}} \tag{9}$$

with $w_l = 0$ for those subspaces not required in the new version.

The eigenvectors were combined to form $\mathbf{M}$ and Equations 8 and 9 used to give the projection $\mathbf{q}'_{\mathbf{j}}$ of face $\mathbf{q}$ for subspace $j$. The dimensions which result with regard to the identity sub-space are shown in Figure 2. In comparison with those in Figure 1 the facial dimensions appear to have the same identities, but are normalised for expression, pose and lighting.

### 5.2 Recognition subspace extraction

It would be possible to perform recognition tests directly on the parameters $\mathbf{w}^{(\mathbf{i})}$ derived from Equation 8 (removing the non-identity portions of $\mathbf{w}$ after calculation), but this makes comparison with a non-recoded condition difficult as it is not clear what role should be given to the non-identity sets which are present in the Appearance Model. In addition, this ignores a further source of identity information, the additional 30 individuals who make up the pose, lighting and expression sets. Thus a final PCA on

$$\mathbf{C}_B = \frac{1}{n_p}\sum_{i=1}^{n_p}(\bar{\mathbf{q}}_i - \bar{\mathbf{q}})(\bar{\mathbf{q}}_i - \bar{\mathbf{q}})^T \tag{10}$$

was applied to the identity projections of the complete set of images, yielding a space with 217 dimensions.

This final rotation allowed direct control of the identity-space dimensionality, and also a sensible control condition. Equation 10 was applied to the Appearance Model parameters, obtaining a set of identity parameters (referred to a 'Not Recoded') which, while using the maximum amount of identity information and some contribution from non-identity information, did so in a entirely linear manner.

## 6 Dimensionality reduction

The dimensionality of the identity face spaces were controlled by implementing Equation 10 in the original image and landmark-location representation. The mean identity parameters $\bar{\mathbf{q}}_i$ (either recoded or not recoded) were projected out through the Appearance Model to give new images. A new appearance model which included all the relevant variance was then generated. This gave greater control than could be achieved working in the appearance domain alone, in particular allowing direct analysis of the type of information present in the region eigenvectors.

## 6.1 Shape approximation

Following Equation 1, a projected version of $\mathbf{q}'_\mathbf{t}$ of variable similarity to $\mathbf{q}$ can be found where

$$\mathbf{q_t}' = \mathbf{\Phi_t}\,\mathbf{w_t} + \bar{\mathbf{q}} \tag{11}$$

by truncating $\mathbf{w}$ to the shape parameters associated with the $t$ highest values in $\lambda$. $\mathbf{\Phi}_t$ only includes the first $t$ eigenvectors and $\mathbf{q}'$ has $p$ members, the x- and y-coordinates of the feature points. The number of parameters was chosen so that the model could approximate the original data to a given accuracy. The smallest model giving an average root mean square error,

$$E_t = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{1}{p}\sum_{j=1}^{p}(q_{ij} - q'_{tij})^2} \tag{12}$$

of less than 0.5 pixels was selected. This criterion was chosen as the real maximum accuracy when placing the landmark points on the images by hand. The 218 normalised ensemble images consistently required only 14 dimensions to code them to this accuracy. The relationship between ensemble-mean RMS and the number of eigenvectors is given in Figure 3, for the final recoded and rotated images. As would be expected, given the Appearance Model, the error is minimal if more than 30 shape parameters are used, although slight numerical inaccuracy in coding and projecting allows a maximum of 140 dimensions. Given the predominately grey-level based nature of memorability information, further reduction in dimensionality was not considered necessary; this model will only code useful identity information in disjoint images.

## 6.2 Region approximation

Memorability information can be measured by analysis of the local consistency of the region eigenvectors. Small-scale, memorability-type variations on faces should be significantly less predictable from adjacent pixels than large-scale familiarity-type variations. This should be measurable from the eigenvectors, since they will have constant sums of squares, irrespective of the variance associated with them (given by the eigenvalues). Local PCAs were performed on the eigenvectors, which were expressed as images for this purpose.

The individual region eigenvectors are one-dimensional objects, while the structure we wish to measure is two-dimensional. Thus the eigenvectors were first converted into an image of the shape implied by the mean of the shape-model, as are approximated images before distortion to the final shape. This formed an irregularly shaped patch, so a suitably sized border was added around it. All the pixel-values in the border were zero, so the mean and sum of squares were not affected. Square samples, typically $7 \times 7$ pixels were then taken centred upon each pixel with the face and border. Inevitably, a narrow band of zero values around the edge of the image was under sampled; only sampling pixels within the face-shape was tried, but it appears that the pixels close to the edge of the face are atypical with respect to the interior and have a larger effect on the result than might be expected. A PCA was then performed on the resultant set of samples. In effect, this produced a local Fourier decomposition of the eigenvector (the new, *local* eigenvectors) and a set of local eigenvalues.

There are two characteristics of eigenvalues which are important to this situation. The first is that their sum will equal the total variance in the local ensemble. This will be constant across the region eigenvectors. Second, their rate of decline in magnitude will depend upon the redundancy of the samples from which they are drawn. Consider a perfectly random region eigenvector, where there is no predictability between adjacent pixels. When the covariance matrix is constructed, only the major diagonal will on average have non-zero values in it. This would produce a set of constant local eigenvalues. Conversely, an entirely predictable region eigenvector (say, a constant brightness-gradient from left to right) will produce an covariance matrix with a single consistent pattern, and no other non-zero values. This will generate a single non-zero local eigenvalue.

The requirement is that we exclude any region eigenvectors which are too noisy, and so do not resemble faces sufficiently. This was assessed by treating the ensemble images as if they were eigenvectors themselves. Each image was sampled to become shape-free, and the grey-levels normalised using Equation 3. The difference from the mean grey-level image was then found and this difference image further normalised,

$$g'_{ij} = \frac{g_{ij} - \bar{g}_i}{\sqrt{\sum_{j=1}^{n}(g_{ij} - \bar{g}_i)^2}}, \tag{13}$$

seting the mean to zero and the sum for squares to one. A local analysis was then performed across each of the ensemble faces, using a $7 \times 7$ pixel window. The mean $\bar{\mathbf{e}}$ and variance $\sigma$ of the sets of eigenvalues were found. With the eigenvalue curves of the first and 200*th* region eigenvectors of the ensemble, the mean $\bar{\mathbf{e}}$ is shown in Figure 4; clearly $\bar{\mathbf{e}}$ lies between the two region eigenvectors.

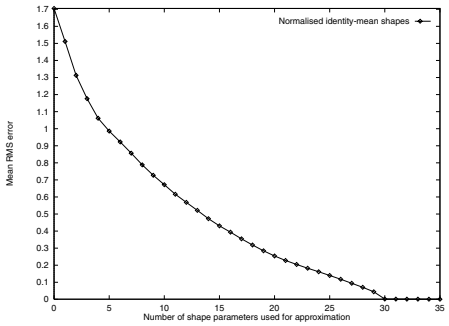


Figure 3: Relationship between the number of eigenvectors in the shape model and the Root Mean Error, averaged across the ensemble of the recoded identity space.

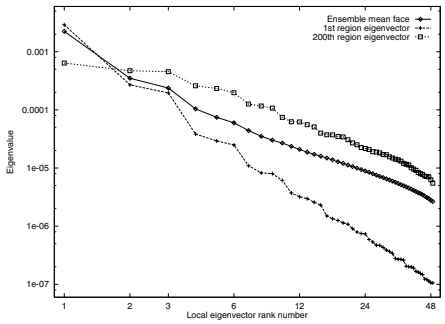


Figure 4: Eigenvalues of local analyses of the ensemble and the first and 200*th* eigenvectors of the recoded identity space.

The Mahalanobis distance between the mean eigenvalues $\bar{\mathbf{e}}$ and each set derived from a region eigenvector can be calculated using Equation 2, giving $d_{i \to \bar{e}}$ distributed as $\chi^2$ on $n-1$ degrees of freedom, as in this case, $\bar{\mathbf{e}}$ and $\mathbf{w}_i$ sum to the same value. Thus the probability that a given eigenvector might not be a true face, $P_{(\mathbf{w}_i \neq \bar{\mathbf{e}})}$ can be calculated. Since we are only interested in excluding eigenvectors which are less predictable than the ensemble, the value of the first eigenvalue for each region eigenvector was examined. If this was higher than the first eigenvalue in $\bar{\mathbf{e}}$, the probability of the eigenvector not being

acceptable was assumed to be zero. This required that $P_{(\mathbf{w}_i \neq \bar{\mathbf{e}})}$ be measured on $n-2$ degrees of freedom.

This operation was performed using the entire set of un-normalised ensemble using a $7 \times 7$-pixel texture patch. The un-normalised ensemble was used since this best resembled the non-ensemble test images, which were also not normalised. The pixel-patch size was chosen as being in the middle of a stable range, and region eigenvectors with lower variance than the lowest variance one with $P_{(\mathbf{w}_i \neq \bar{\mathbf{e}})} < 1$ were excluded. This retained approximately 90 eigenvectors. The distances from which $P_{(\mathbf{w}_i \neq \bar{\mathbf{e}})}$ for the final recoded identity space are shown in Figure 5.

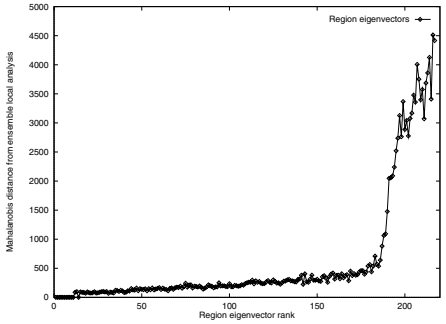


Figure 5: Mahlanobis distances between local PCAs of region eigenvectors and the ensemble for the final recoded identity space.

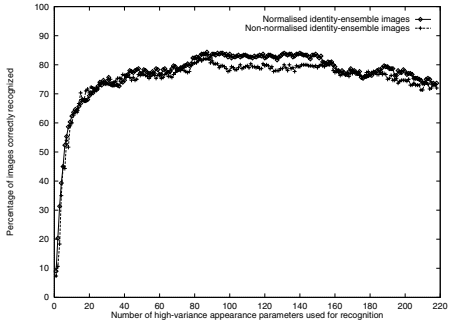


Figure 6: Recognition rates for Euclidean average-image matching as a function of the number of dimensions in the final identity appearance space.

# 7 Recognition Results

To assess the accuracy with which the identity space had been measured, recognition was measured on a large test set (first used in [12]), consisting of 600 images of 30 individuals, divided in half: a gallery of 10 images per person and a set of 10 probes per person. The Active Appearance Model used to provide correspondences did not give completely accurate positions, as information on the hue of the faces was not avaiable. The degraded recognition accuracy overall, but the relatively high error rates do allow confidence that performance is improved.

The appearance model parameters $\mathbf{w}$ for the identity models were found by Equation 1 and the cross-individual pooled covariance matrix was found using Equation 5 for the gallery images. This allowed

$$d^2_{i \to k} = (\mathbf{w_k} - \bar{\mathbf{w}}_\mathbf{i})^T \mathbf{C}_W^{-1} (\mathbf{w_k} - \bar{\mathbf{w}}_\mathbf{i}), \tag{14}$$

where $1 \leq k \leq (n_o \times n_p)$ to give Mahalanobis distances from the probes to the mean images of the gallery. A recognition was scored when the smallest $d$ had the same identity for $i$ and $k$.

The results in Figure 6 show the effects of varying the number of dimensions used to determine identity for the final recoded and rotated space, compared with the equivalent

non-recoded space. Clearly there is an advantage for the recoding, but this interacts with the number of dimensions, which suggests that misleading specific, memorability-type variance is being coded.

Table 1 gives the effects of the recoding and dimensionality reduction algorithms in more detail. Both operations increase recognition performance, truncating rather more dramatically than recoding. In addition, the smaller, truncated model also offers a speed advantage. In comparison, performance on the complete appearance model, which codes expression, pose and lighting as well as both familarity and memorability identity information, is very bad.

| Space Type | | Dimensions | | | Percent Recognized |
|---|---|---|---|---|---|
| | | Shape | Region | Appearance | |
| Appearance Model | | 30 | 455 | 374 | 17.33 |
| Not recoding | All dimensions | 140 | 217 | 217 | 68.66 |
| Not recoding | Reduced dimensions | 14 | 92 | 106 | 80.33 |
| Recoding | All dimensions | 139 | 218 | 217 | 73.00 |
| Recoding | Reduced dimensions | 14 | 91 | 103 | 83.66 |

Table 1: Effects of subspace type and dimensionality determination on face recognition. While 'Appearance Model' referes to distances measured on the complete face-space, 'Not recoding' and 'Recoding' refer to different methods of calculating identity spaces.

# 8 Conclusions

Once an accurate coding system for faces has been achieved, the major problem is to ensure that only a useful sub-set of the codes are used for any given manipulation or measurement. This is a notably difficult task, as there are multiple, non-orthogonal explanations of any given facial configuration. It is also typically the case that only a relatively small portion of the very large data-base required will be present in the full range of conditions and with the labels needed for a simple linear extraction. Further, an analysis of the ensemble which seeks to allow full reconstruction of images, as does PCA, will extract information which, while useful in describing the ensemble, will not be applicable to other images which need to be analysed.

We have shown that both of these problems can be overcome by using a recoding scheme which takes into account both the variance of and covariance between the functional subspaces which can be extracted to span sets of faces which vary in different ways, coupled with the analysis of the importance of the shape and region eigenvectors which can be derived from these spaces. In particular, the number of region eigenvectors can be controlled by measuring their degree of local predictability; un-predictable eigenvectors are not useful for coding non-ensemble faces. Together, when tested on images which were deliberately hard to code, being both previously unseen, and with low-quality correspondence matches these significantly raise recognition rates, halving the number of errors.

This method also allows simultaneous extraction of pose, lighting and expression codes. Correlational analysis of the residuals following coding on the identity spaces might allow further improved recognition; this would depend upon the use of the recoding

algorithm to produce an identity-only image which can then be separated into familiarity and memorability aspects. Further advances may also be possible by taking explict steps to minimize the variance in the recoded identity space of parameters for a single individual and possibly expression, lighting and pose across individuals, rather than allowing them to vary freely, as here. Such a image-sequence based algorithm could be applied at both sub-space construction and gallery (and possibly probe) encoding time.

# References

[1] M. J. Black, D. J. Fleet and Y. Yacoob. A framework for modeling appearance change in image sequences. *6th ICCV*, pages 660–667, 1998.

[2] J. R. Vokey and J. D. Read Familiarity, memorability, and the effect of typicality on the recognition of faces. *Memory and Cognition*,Vol 20, pages 291–302, 1992.

[3] N. P. Costen, I. G. Craw, G. J. Robertson and S. Akamatsu. Automatic face recognition: What representation ? *European Conference on Computer Vision, Vol 1*, pages 504–513, 1996.

[4] G. J. Edwards, A. Lanitis, C. J. Taylor and T. F. Cootes. Modelling the variability in face images. *2nd Face and Gesture*, pages 328–333, 1996.

[5] N. P. Costen, I. G. Craw, T. Kato, G. Robertson and S. Akamatsu. Manifold caricatures: On the psychological consistency of computer face recognition. *2nd Face and Gesture*, pages 4–10, 1996.

[6] T. Poggio and D. Beymer. Learning networks for face analysis and synthesis. *Face and Gesture*, pages 160–165, 1995.

[7] B. Moghaddam, W. Wahid and A. Pentland. Beyond eigenfaces: Probabilistic matching for face recognition. *3rd Face and Gesture*, pages 30–35, 1998.

[8] S. Duvdevani-Bar, S. Edelman, A. J. Howell and H. Buxton. A similarity-based method for the generalization of face recognition over pose and expression. *3rd Face and Gesture*, pages 118–123, 1998.

[9] D. B. Graham and N. M. Allinson. Face recognition from unfamiliar views: Subspace methods and pose dependency. *3rd Face and Gesture*, pages 348–353, 1998.

[10] T. F. Cootes, G. J. Edwards and C. J. Taylor. Active Appearance Models. *European Conference on Computer Vision, Vol 2*, pages 484–498, 1998.

[11] G. J. Edwards, C. J. Taylor and T. F. Cootes. Learning to Identify and Track Faces in Image Sequences. *British Machine Vision Conference*, pages 130–139, 1997.

[12] A. Lanitis, C. J. Taylor and T. F. Cootes. An automatic face identification system using flexible appearance models. *British Machine Vision Conference*, pages 65–74, 1994.