Compensating for ensemble-specific effects when building facial models

N.P.Costen, T.F.Cootes and C.J.Taylor
University of Manchester
Division of Imaging Science and Biomedical Engineering,
Stopford Building, Oxford Road,
Manchester M13 9PT, U.K.

Abstract

When attempting to code faces for modeling or recognition, estimates of dimensions are typically obtained from an ensemble. These tend to be significantly sub-optimal. Firstly, ensembles are rarely balanced with regard to identity and expression. This can be overcome by dividing the ensemble by type of variation and rotating sub-spaces relative to one another. Secondly, each face contains both predictable and non-predictable qualities; only the predictable aspects are useful for defining coding systems for other faces. Variance-based methods of defining codes (PCA) will provide eigenvectors which are themselves potential faces. Predictable aspects will induce eigenvectors with comparable levels of spatial redundancy to the ensemble. We show that this gives relatively short and consistent codes, and allows fast and accurate fitting of codes to faces.

1 Introduction

A major objective with any face-processing system is to be able to code faces regardless of their relationship with our specific knowledge of faces, the ensemble. If this is possible, we can be confident that variation in codes between or within faces will reflect real differences, and these variations will be optimally consistent across faces. The most obvious method of improving the generality of codes is to increase the size of the ensemble. However, this usually has an effect of decreasing the codes' specificity by increasing the number of dimensions. Further, 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 [1].

Memorability information reflects small, discrete, easily verbalised features, for example skin blemishes or warts. Such information has essentially infinite dimensionality and will exhibit fortuitous correlations between faces. Thus it can disproportionately reduce both the specificity and the generality of a set of codes. Within a Principal Components setting, familiarity weighs on the early, high variance eigenvectors, while memorability correlates with the later, low-variance eigenvectors [2]. However, although this simple variance division does not offer an obvious cut-off point, the predictability of the general familiarity codes suggests a method. Given a full-reconstruction constraint (present when using Principal Components), they must themselves be acceptable as faces;

conversely the specific memorability codes must not be. The localized nature of the memorability information implies that the low-variance codes will be less spatially predictable than real faces. We thus use grey-level based codes, measuring the spatial redundancy in small, spatially-adjacent sub-samples. Probabilistic methods of distinguishing these from the samples found in the real faces of the ensemble are developed, and we show that this yields consistent estimates of the number of parameters, and allows the construction of both general and specific facial codes.

An additional problem is the selection of images used in the ensemble. Typically, it is impossible to obtain a completely balanced set of all the people in the ensemble in all possible pose, expression and lighting combinations. This is especially true if we wish to have average coverage of the identity space. However, if we can divide our complete ensemble into subsets which vary predominately on individual types of variation, we can overcome this problem by adopting a recoding strategy. This allows the construction of optimal non-orthogonal sub-spaces, which can then be combined with one another [5] to form a single, unbiased set of dimensions.

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. A number of previous studies [3, 4, 6] 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 $\pm 20^o$ [7]. 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 during training. 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[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. Given a set of N vectors $\mathbf{q_i}$ (either the pixel grey-levels, or the feature-point locations) sampled from the images, the covariance matrix $\mathbf{C}$ of the images is calculated,

$$\mathbf{C} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{q_i} - \bar{\mathbf{q}})(\mathbf{q_i} - \bar{\mathbf{q}})^T, \tag{1}$$

and orthogonal unit eigenvectors $\mathbf{\Phi}$ and a vector of eigenvalues λ are extracted from $\mathbf{C}$. This allows an estimate of the dimensions and range of the face-space. The weights $\mathbf{w_i}$ of a face can then be found,

$$\mathbf{w_i} = \mathbf{\Phi}^T(\mathbf{q_i} - \bar{\mathbf{q}}) \tag{2}$$

and the projected version $\mathbf{q}_\mathbf{i}'$ of the face,

$$\mathbf{q}_\mathbf{i}' = \mathbf{\Phi}\mathbf{w}_\mathbf{i} + \bar{\mathbf{q}}. \tag{3}$$

Since the columns of the matrix $\mathbf{\Phi}$ are orthogonal (and typically ordered by declining magnitude of λ_j) the similarity between $\mathbf{q}_\mathbf{i}$ and the projected version, $\mathbf{q}_\mathbf{i}'$ can be controlled by truncating $\mathbf{\Phi}$, and with it $\mathbf{w}$.

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 second PCA is performed [4]. This 'appearance model' allows the description of the face in terms of true, expected variation – the distortions needed to move from one to another [8]. An example of the first few dimensions of such a model is provided in Figure 1. However, it will potentially code the entire variation between the faces which form our ensemble, including both the general and specific variance. The followings studies aim to exclude the specific variance, leading to a smaller and more useful model.

3 Appearance Model Construction

For testing purposes, an ensemble of 314 facial images was used. This comprised 218 different individuals (the image to individual mapping was known), and was sub-divided into groups varying on facial pose, expression and lighting. Males and females were present in approximately equal proportions, and the individuals were drawn from a range of ages and ethnic groups. All the images had a uniform set of 68 landmarks found manually. A triangulation was applied to the points, bilinear interpolation used to warp the images to a standard shape and size which would yield a fixed number of pixels, which can be varied at the experimenter's will.

Since the images were gathered with a variety of cameras, it was necessary to normalise the lighting levels. The shape-free grey level patch g_i was sampled from the ith shape-normalised image. To minimise the effect of global lighting variation, this patch was normalised at each pixel j to give

$$g'_{ij} = (g_{ij} - \mu_j)/\sigma_j \tag{4}$$

where μ_j,σ_j are the mean and standard deviation for pixel j across the ensemble.

4 Dimensionality reduction

The number of parameters was controlled separately for the shape and region domains, before combining to form an appearance model describing all the relevant variance.

4.1 Shape approximation

In Equation 2, $\mathbf{q}_\mathbf{i}$ has n members, the x- and y-coordinates of the feature points, while $\mathbf{w}_\mathbf{i}$ is a vector of t *shape parameters*. The variance of the j^{th} parameter across the training set, w_j, is given by the λ_j. The number of eigenvectors used is chosen either to explain a given proportion of the variance in the aligned data (e.g. 98%), or, more appropriately, so

that the model can approximate the original data to a given accuracy. The smallest model giving an median root mean square error,

$$E_t = \mathbf{H}_{i=1}^{n} \left(\sqrt{\frac{1}{p} \sum_{j=1}^{p} (q_{ij} - q'_{tij})^2} \right) \tag{5}$$

of less than 1 pixel was selected. The $\mathbf{H}$ operator applies a fully-converged robust Huber M-estimator of the central tendency of the data. A robust estimate of central tendency was chosen to help exclude observations which were highly abnormal (and thus needed extra dimensions to allow modeling) either due to inaccuracies in markup, or memorable face shape.

A value of 1 pixel (in the un-aligned data) is chosen as this is the real maximum accuracy; ignoring experimenter error, there is maximum difference of 1 pixel between the 'real' and found locations. In practice, the ensemble requires 38 eigenvectors. Given the grey-level based nature of most memorability information, further reduction in dimensionality was not considered necessary.

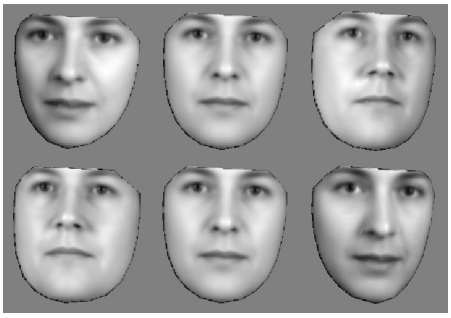

Figure 1: The first two dimensions of the face-space in the appearance model. From the left, $-2s.d.$, the mean $+2s.d.$.

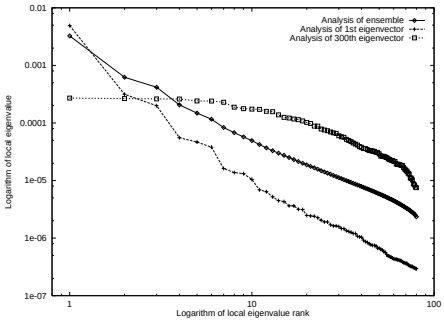

Figure 2: Eigenvalues of local analyses of the ensemble and the first and 300*th* eigenvectors.

4.2 Region approximation

The local consistency of the region eigenvectors was measured. 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. One possible method is to perform a local PCA on the eigenvector.

The eigenvectors were converted into an image of the shape implied by the mean of the shape-model, just as an approximated image must be 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 were then taken,sampling each pixel within the face an equal number of times. The samples are odd-numbered squares, i.e. 3×3, 5×5, 7×7 and so on.

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 eigenvalues. The sum of the eigenvalues equals the total variance in the ensemble. This will be constant across the region eigenvectors In addition, their rate of decline in magnitude will depend upon the redundancy of the samples from which they are drawn. A perfectly random eigenvector, where there is no predictability between adjacent pixels will have a set of constant eigenvalues. Conversely, an entirely predictable region eigenvector (say, a constant brightness-gradient from left to right) will generate a single non-zero 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 4. The difference from the mean grey-level image was then found and this difference image further normalised,

$$g'_{ij} = \frac{g_{ij} - \bar{g}_i}{\sqrt{\sum_{j=1}^{n}(g_{ij} - \bar{g}_i)^2}}, \tag{6}$$

setting the mean to zero and the sum for squares to one. A local analysis was then performed across each of the ensemble faces, and the mean $\bar{\mathbf{e}}$ and variance σ of the sets of eigenvalues found. With the eigenvalue curves of the first and 300*th* region eigenvectors of the ensemble, the means are shown in Figure 2. A 9×9 pixel window was used for the analysis.

The Mahalanobis distance between the eigenvalues derived from the ensemble and each set derived from an eigenvector can be calculated,

$$d_i^2 = \sum_{j=1}^{n} \frac{(\bar{\mathbf{e}}_j - \mathbf{w}_{ij})^2}{\sigma_j}, \tag{7}$$

with d_i^2 distributed as χ^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 could be a true face 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 eigenvector was examined. If this was higher than the first eigenvalue derived from the ensemble, the probability of the eigenvector being acceptable was assumed to be one. This required that d_i be measured on $n-2$ degrees of freedom.

5 Sub-space Calculation

The aim of the recoding algorithm is to take account of the multiple possible explanations of the coding of a given face. Considering the combination of different sub-spaces, if n_s are used, each described by eigenvectors $\mathbf{\Phi^{(j)}}$ with the associated eigenvalues $\lambda^{(\mathbf{j})}$, for a given face $\mathbf{q}$ the projection out of the combined subspaces is given by

$$\mathbf{q}' = \sum_{j=1}^{n_s} \mathbf{\Phi}^{(\mathbf{j})} \mathbf{w}^{(\mathbf{j})} + \bar{\mathbf{q}}. \tag{8}$$

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

$$E = \sum_{j=1}^{n_s} \sum_{i=1}^{N_j} \frac{(w_i^{(j)})^2}{\lambda_i^{(j)}} \tag{9}$$

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

$$\mathbf{w} = (\mathbf{D}\mathbf{M}^T\mathbf{M} + \mathbf{I})^{-1}\mathbf{D}\mathbf{M}^T(\mathbf{q} - \bar{\mathbf{q}}) \tag{10}$$

and this also gives a projected version of the face

$$\mathbf{q}' = (\mathbf{D}\mathbf{M}^T)^{-1}(\mathbf{D}\mathbf{M}^T\mathbf{M} + \mathbf{I})\mathbf{w} + \bar{\mathbf{q}} \tag{11}$$

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

The first stage was to obtain the appearance-model weights (using Equation 2) for each image used to build the truncated model. Separate PCAs were then performed upon the sets of the weights. The covariance matrices for the identity and lighting subspaces were calculated using Equation 1 while the pose and expression subspaces used

$$\mathbf{C}_W = \frac{1}{n_o n_p} \sum_{i=1}^{n_p} \sum_{k=1}^{n_o} (\mathbf{q}_{ki} - \bar{\mathbf{q}}_i)(\mathbf{q}_{ki} - \bar{\mathbf{q}}_i)^T \tag{12}$$

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.

The eigenvectors were combined to form $\mathbf{M}$ and Equations 10 and 11 used to give the projection $\mathbf{q}'_\mathbf{j}$ of face $\mathbf{q}$ for subspace j. This procedure looses useful variation. For example, the identity component of the expression and pose images was unlikely to be coded precisely by the identity set alone. Thus the full projection $\mathbf{q}'$ was calculated, and recoded image $\mathbf{r_j}$ included an apportioned error component:

$$\mathbf{r_j} = \mathbf{q}'_\mathbf{j} + \frac{(\mathbf{q}' - \mathbf{q}) \sum_{k=1}^{N_j} \lambda_k^{(j)}}{\sum_{j=1}^{n_s} \sum_{k=1}^{N_j} \lambda_k^{(j)}}. \tag{13}$$

This yielded four ensembles, each of 314 images. A further four PCAs were performed on the recoded ensembles (all using Equation 1), extracting the same number of components as on the previous PCA for the lighting, pose and expression subspaces, plus all the non-zero components for the identity sub-space. Combined, these formed a new estimate of $\mathbf{M}$, and Equations 10, 11 and 13 were applied to give a third-level estimate and so forth. Convergence was assessed by measuring the Mahalanobis distance between the projections of the images the various spaces. The algorithm continued until successive iterations produced the same pattern of distances; in practice this was almost always achieved by the third iteration.

6 Recoded Appearance Model

It is possible to add Principal Component spaces together[5]. This requires the mean of the space, the eigenvectors, eigenvalues and number of observations used in each of the spaces to construct a single, combined space which is will span all the examples in the sub-spaces, and will give proper emphasis to those spaces (most notably lighting and pose) which were under-represented in the complete ensemble. However, this recoding method will still tend to produce under-estimates of the variances of the under-represented spaces, as the majority of images will be unusually clustered around the mean. Thus, rather than the eigenvectors themselves, we use estimates of variance derived from the final recoded parameters of the sub-sets of images which we used to construct the original estimates of spaces. The images were sorted into the appropriate groups and Equations 10 and 11 used to give normalised versions. Equation 2, using the same space, then gave the weights and the variance for each eigenvalue was found.

It would have been possible to control the appearance model indirectly, through the space, but for ease of integration it was considered necessary to transform the recoded space into an appearance model. It was possible to calculate the eigenvectors for the new appearance model

$$\mathbf{\Phi}_{(n)} = (\mathbf{\Phi}_{(a)}^T \mathbf{\Phi}_{(r)}^T)^T \tag{14}$$

and then use the eigenvalues for the recoded space and mean of the old appearance model directly on the new appearance model eigenvectors. No alterations at the shape and texture level were necessary. The first two dimensions of this new recoded space are shown in Figure 3, they show a reduced effect of identity and expression and an increased effect of pose.

7 Results

An effective method of controlling the dimensions of an appearance model should be relatively unaffected by parameter variation, and should allow more accurate searching for faces than an un-controlled appearance model.

7.1 Effects of sub-sampling scheme

There are two major parameters which need to be determined for this algorithm; the size of the sampling window, and the minimum acceptable probability that a region eigenvector has less structure than the ensemble images. Taking the region eigenvectors derived from the full ensemble, the probability that each eigenvector was distinguishable from the ensemble was calculated. The size of the local sample was varied, from a 3×3 pixel square, up to 9×9 pixels. Since the probabilities do not rise very smoothly, a criterion probability, P_c was selected, and all the eigenvectors up to the highest ranked one with $P \leq P_c$ were accepted. The results are shown in Figure 4. As can be seen, the 5×5, 7×7 and 9×9 samples are almost identical, and the number of eigenvectors accepted are steady for a range of values of P_c in the 0.9 – 0.999 range, where a criterion is likely to be placed.

On this basis, the 7×7 pixel square was chosen, as was a maximum allowed probability of difference from the ensemble of $P_c = 0.99$. This yielded 89 eigenvectors; setting $P_c = 0.95$ would yield 86 eigenvectors, and $P_c = 0.999$, yields 99 eigenvectors.

7.2 Effects of ensemble and image size

The analyses with the parameters derived above were run on a range of ensembles of different sizes. All were quasi-random sub-sets of an enlarged ensemble, with 430 images. When the shape-free face is constructed, a decision must be made about the number of pixels to include in the vectors which are submitted to PCA. Thus the image-size is controlled, sampling by bilinear interpolation from the original, un-processed image. A set of models at different resolutions can be built; the lower dimensionality to be expected from the smaller models can be made use of when searching images for faces.

The results of this test are shown in Figure 5, showing the number of region eigenvectors accepted. There is relatively little redundancy between the shape and the region parameters; the number of appearance model parameters for a given ensemble–image-size combination can be obtained by adding the number of parameters for that ensemble.

Figure 3: The first two dimensions of the face-space as defined by the recoded appearance model.

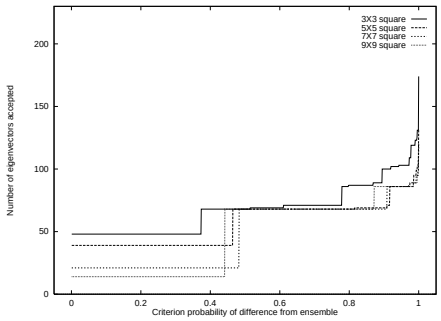

Figure 4: Effects of varying criterion probability on the number of region eigenvectors accepted, for a range of window-sizes.

Clearly, both the number of faces in the ensemble, and the number of pixels in each face affect the number of dimensions extracted. Increasing either will increase the variance to be explained and the ways the familiarity information in the ensemble faces may vary. However, the rate of increase declines with increasing number of faces in the ensemble, suggesting an asymptotic value of approximately 105 texture eigenvectors for 6000 pixels, and so about 140 appearance model parameters. This is considerably less than the figure suggested elsewhere [10], but does partially reflect the use of the appearance model. It should be noted however, that the 625 pixel line is actually higher than that for 6000 pixels. This reflects the major limitation of this method, that if the number of images and/or pixels is sufficiently small, all the texture eigenvectors will be of similar redundancy, and so cannot be discriminated.

7.3 Approximating Novel Faces

The aim of the recoding algorithm is to simulate the effects of using a larger more representative ensemble. In practice, this will take the form of a rotation within the larger vector-space in which the face-space is embedded, but should not involve a translation or alteration in the number of dimensions used. Thus, while the position of the face in the larger space should not change relative to the space mean, if the space is better able

to represent non-ensemble faces, they should be closer to the axes of the space. This can be observed by taking the absolute sum of appearance model weights; on average non-ensemble faces should have higher sums on the recoded space than on the original space.

A pair of appearance models were trained from the ensemble, using 6000 pixels per face. The variances were measured for a set of 157 images of people not in the ensemble, showing notable pose and expressive changes. The results are shown in Figure 6, where the x-axis is the sum of squares on the truncated, but un-recoded space, while the y-axis is the change in absolute sum for each image when the space is rotated. The mean change in sum is 0.1119; since positive values imply smaller absolute sums, the axes are moving closer to the non-ensemble images.

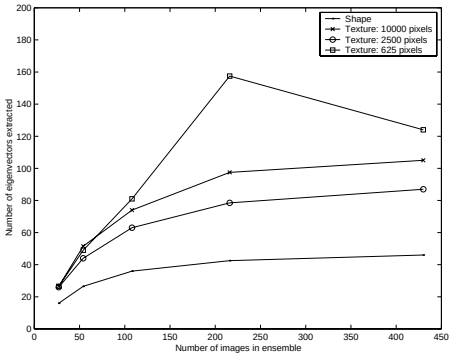

Figure 5: Effects of variation of the ensemble and image size on the number of eigenvectors extracted.

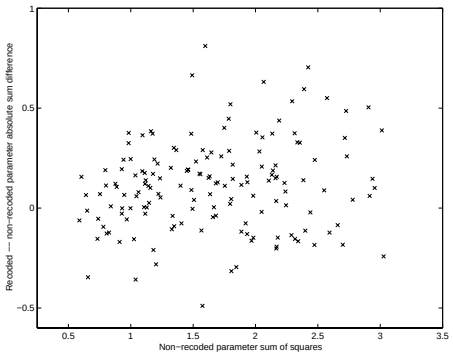

Figure 6: Effects of recoding on eigenvector parameter variation.

7.4 Fitting to Faces

Given an appearance model, it is possible to automatically find parameters which describe a face, without needing to first obtain a set of correspondence points. An Active Appearance Model [9] can be trained on the basis of coding errors, predicting the change in the appearance-model weights (and the pose and grey-level transformation) necessary to minimize the difference between a projected face $\mathbf{g}'$ and the actual image. Active appearance models were trained from each of the appearance models, and the accuracy with which they converged on the test-set measured. Since the actual point locations were known, these were used to supply starting positions, with the model being displaced known distances around this and allowed to converge.

To judge to efficiency of searching, the mean error between know and located grey-levels taken. While the mean error for the recoded model was 0.00804 pixels, that for the un-recoded model was 0.00578. It proved impossible to build an untruncated version of the active appearance model; the number of displacements necessary to estimate the weight-image relationship was too great.

8 Conclusions

Once faces have been accurately coded, the major problem is to ensure that only a useful sub-set of the codes are used for manipulations or measurement. A given set of codes will respond to both generic variation, useful when considering faces not in the ensemble, and variation specific to the ensemble. Only the former is truly 'facial'.

We have shown that this problem can be overcome by measuring the local structures in the eigenvectors, and comparing these with the ensemble. This yields smaller models, which are relatively independent of both the parameters of the local analysis, and the ensemble size. This then allows searching of images for novel faces. The relationship between the number of faces in the ensemble and the number of eigenvectors in the model suggests that, for a given image-resolution, there are a fixed number of 'real' modes of facial variation. Inadequacies in the ensemble can be overcome by modeling different causes of variation between and within faces, if we can classify the variation in a sub-ensemble. This makes the model more applicable to non-ensemble images, but causes slightly less accurate searching.

References

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

[2] A. J. O'Toole, K. A. Deffenbacher, D. Valentin and H. Abdi Structural aspects of face recognition and the other race effect. *Memory and Cogntion*,vol 22, pages 208–224, 1994.

[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] P. Hall and D. Marshall and R. Martin. Adding and Subtracting Eigenspaces. *British Machine Vision Conference, Vol 2*, pages 463–472, 1999.

[6] 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.

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

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

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

[10] P. S. Penev and L. Sirovich. The Global Dimensionality of Face Space. *4th Face and Gesture*, pages 264–270, 2000.