Simultaneous extraction of functional face subspaces

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. An improved method of measuring these was designed, within the space defined by an Appearance Model. Initial estimates of the subspaces (lighting, pose, identity, expression) were obtained by Principal Components Analysis on appropriate groups of faces. An iterative algorithm was applied to image codings to maximise the probability of coding across these non-orthogonal subspaces before obtaining the projection on each sub-space and recalculating the spaces. This procedure enhances identity recognition, reduces overall sub-space variance and produces Principal Components with greater span and less contamination.

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. For example, a security application will need to recognize individuals regardless of expression, pose and lighting, while a lip-reader will concentrate only on expression. In certain circumstances, accurate estimates of all the subspaces are needed, for example when 'transferring' face and head movements from a video-sequence of one individual to another to produce a synthetic sequence.

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 [7]. 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 iterative approach to a set of pure, but overlapping, spaces. These are more specific than the initial spaces, 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 [2, 3, 4] 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 $\pm 20^o$ [5]. 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 the ensemble. 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[2]. 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 λ 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 a face

$\mathbf{q}$ 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,

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

between faces $\mathbf{q}_1$ and $\mathbf{q}_2$, coding in terms of expected variation [6]. 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 [3].

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. These must be separated to allow detailed analysis of the face image. The following studies are performed embedded within this representation and will be directed towards deriving an explicit separation of the functional subspaces.

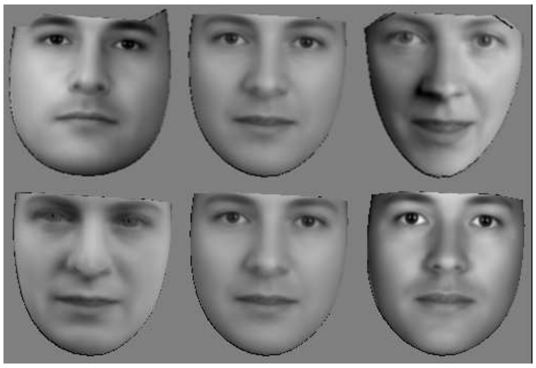

Figure 1. The first two dimensions of the face-space as defined by the appearance model. From the left, $-2s.d.$, the mean $+2s.d.$. The eigenfaces vary on identity, expression, pose and lighting.

3 Available Data

Although estimates of the subspaces might be gained from external codes of every ensemble face on each type of variation, these are typically not available. Rather, different sets, each showing major variation on one subspace alone were used. 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. An example of an ensemble image with landmarks is shown in Figure 2. 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.

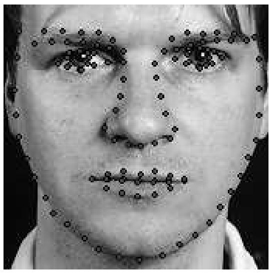

Figure 2. An example of an ensemble image (from the expression set), showing the correspondence points.

Since the images were gathered with a variety of cameras, it was necessary to normalise the lighting levels. For a given pixel, a grey-level of, say, $128/256$ has a different meaning in one shape-normalised image from another. 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

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

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

These operations allowed the construction of an appearance model [3] coding 99.5% of the variation in the 690 images, each with 19826 pixels in the face area. This required a total of 636 eigenvectors. For testing purposes, the feature points were found using a multi-resolution Active Appearance Model [9], constructed using the ensemble images, but without grey-level normalisation.

5 Sub-space Extraction

If n_s subspaces 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

$$\mathbf{q}' = \sum_{j=1}^{n_s} \mathbf{\Phi^{(j)}} \mathbf{w^{(j)}} + \bar{\mathbf{q}}. \tag{4}$$

The problem then becomes what constraints are needed to ensure $\mathbf{q}' = \mathbf{q}$ for fully-coded ensemble members.

5.1 Linear Methods

It was initially hoped that the different functional subspaces might be linearly separable. In that case, it would be possible to obtain the subspaces by successively projecting the faces though the spaces defined by the other categories of faces and the taking the coding error as the data for the subsequent PCA. However, tests showed that this was not practical. The fourth and final set of components consistently coded little but noise. A procedure where each sub-space removed only facial codes within it's own span (typically $\pm 2S.D.$) did produce a usable fourth set, but the application was essentially arbitrary, and only used a small sub-set to calculate each sub-space.

5.2 Non-linear Recoding

This strongly suggested that it might be possible to extract the relevant data in a more principled manner, using the relevant variation present in each image-set. The basic problem is that each of the subspaces specified by the ensembles will code both the desired, 'official' variance, and an unknown mixture of the other types. This contamination will stem mostly from a lack of control of the relevant facial factors, so for example, the 'neutral' expressions seen in the identity set actually contain a range of different, low-intensity expressions, and the ensemble will fail

Figure 3. The first two dimensions of the identity face-space. From the left, $-2s.d.$, the mean $+2s.d.$. The eigenfaces vary mostly on identity and lighting.

to span complete identity variation, as the starting identity eigenfaces show in Figure 3.

In addition, there is no guarantee that the desired, 'pure' principal components for one subspace will be orthogonal to the others. This stems from the ultimate linking factors, notably the three-dimensional face shape and the size and location of facial musculature. Significant improvements in tracking and recognition are possible by learning the path through face-space taken by sequence of face-images [8, 10]. This suggests that these relationships may be susceptible to second order modeling, and that the estimates of the modes of variation given by the ensembles will be biased by the selection of images.

These considerations suggest a scheme based on the differences in variance on the components extracted from the various ensembles. Assuming that the ensembles predominately code the types of variance they are intended to, the eigenvalues for the 'signal' components should be larger than those of the 'noise'. The 'signal' components should also be somewhat more orthogonal to one another, and should certainly be less affected by minor changes in the ensembles which create them.

The components can thus be improved by coding images on the over-exhaustive multiple subspaces in proportion to their variance, approximating the images on the separate subspaces and re-calculating the spaces. This implies the constraint on $\mathbf{w}$ in Equation 4 that

$$E = \sum_{j=1}^{n_s} \sum_{i=1}^{N_j} \frac{(w_i^{(j)})^2}{\lambda_i^{(j)}} \tag{5}$$

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

$$\mathbf{w} = (\mathbf{D}\mathbf{M}^T\mathbf{M} + \mathbf{I})^{-1}\mathbf{D}\mathbf{M}^T(\mathbf{q} - \bar{\mathbf{q}}) \tag{6}$$

and this also gives a projected version of the face

$$\mathbf{q}' = (\mathbf{D}\mathbf{M}^T)^{-1}(\mathbf{D}\mathbf{M}^T\mathbf{M} + \mathbf{I})\mathbf{w} + \bar{\mathbf{q}} \quad (7)$$

with $w_l = 0$ for those subspaces not required.

5.3 Implementation

The first stage was to subtract the overall mean from each face, so ensuring that the mean of each sub-space was as close to zero as possible. 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

$$\mathbf{C}_T = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{q}_i - \bar{\mathbf{q}})(\mathbf{q}_i - \bar{\mathbf{q}})^T \quad (8)$$

while the pose and expression subspaces used

$$\mathbf{C}_W = \frac{1}{n_o n_p}\sum_{i=1}^{n_p}\sum_{k=1}^{n_o}(\mathbf{q}_{ki} - \bar{\mathbf{q}}_i)(\mathbf{q}_{ki} - \bar{\mathbf{q}}_i)^T \quad (9)$$

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 6 and 7 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:

$$\mathbf{r_j} = \mathbf{q}'_{\mathbf{j}} + \frac{(\mathbf{q}' - \mathbf{q})\sum_{k=1}^{N_j}\lambda_k^{(j)}}{\sum_{j=1}^{n_s}\sum_{k=1}^{N_j}\lambda_k^{(j)}}. \quad (10)$$

This yielded four ensembles, each of 690 images. A further four PCAs were performed on the recoded ensembles (all using Equation 8), 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 6, 7 and 10 were applied to give a third-level estimate and so forth. The final result with regard to the identity image set is shown in Figure 4. In comparison with those in Figure 1 the facial dimensions appear to have the same identities, but are normalised for expression, pose and lighting.

Since the identity space was allowed to vary the number of eigenfaces, while the others were fixed, inevitably any noise present in the system will tend to accumulate in

Figure 4. The first two dimensions of the identity face-space. From the left, $-2s.d.$, the mean $+2s.d.$. The eigenfaces vary only on identity, the range of which has been increased.

the identity space, and will reduce recognition performance. Thus once the system had stabilized, a final PCA on

$$\mathbf{C}_B = \frac{1}{n_p}\sum_{i=1}^{n_p}(\bar{\mathbf{q}}_i - \bar{\mathbf{q}})(\bar{\mathbf{q}}_i - \bar{\mathbf{q}})^T \quad (11)$$

was applied to the identity projections of the complete set of images, coding 97% of the variance and reducing the identity eigenvectors from 497 to 153. This final rotation maximized between-person variance and was used only for recognition.

6 Results

Convergence of the system was estimated by taking the Mahalanobis distance d (using Equation 2), between all the images on each of the functional subspaces. A Pearson product-moment correlation,

$$r_i = \frac{\sum_{j=1}^{n_d}(d_{(i-1)j} - \bar{d}_{(i-1)})(d_{ij} - \bar{d}_i)}{\sqrt{\sum_{j=1}^{n_d}(d_{(i-1)j} - \bar{d}_{(i-1)})^2\sum_{j=1}^{n_d}(d_{ij} - \bar{d}_i)^2}} \quad (12)$$

between iterations i and $i-1$, where $n_d = 4 \times n(n-1)$, was calculated for successive iterations until $r_i = 1$.

The system gave a relatively smooth set of correlation coefficients as shown in Figure 5, converging in approximately seven iterations. Practical convergence was achieved by iteration 4, since to avoid numerical accuracy problems only 99.99% of the ensemble variance was coded.

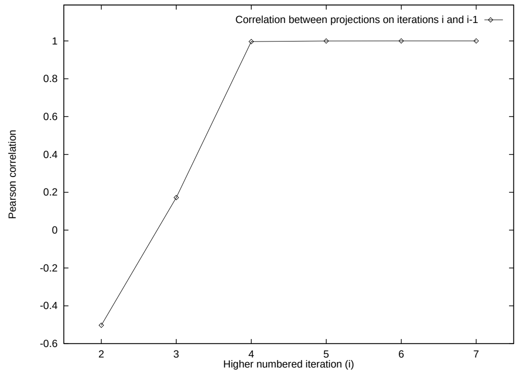

Figure 5. Correlations of the Mahalanobis distances separating all the images on the multiple space between iterations i and $i-1$.

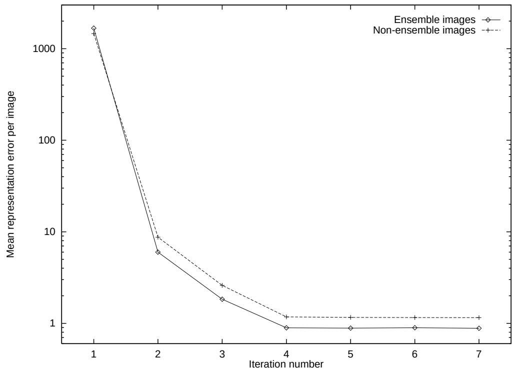

Figure 6. Mean coding errors for the ensemble and test images across iterations.

6.1 Coding errors

Since the iterations involve the inclusion of information which failed to be coded on the previous iteration, it should be expected that the difference between original and projected images should decline. This should apply to both ensemble and non-ensemble images as the eigenfaces become more representative.

This was tested by projecting the images through the combined subspaces (using Equations 6 and 7) and measuring the magnitude of the errors. This was performed for both the ensemble images and also for a large test set (referred to as 'Manchester', first used in [11]), 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. Figure 6 shows that in both cases, the errors drop to a negligible level by iteration 4. As a comparison, the two sets have mean magnitudes (total variance) of 11345 and 11807, measured on the appearance-model eigenweights. Errors on the individual subspaces remain high, approximately 4,000 to 11,000.

6.2 Sub-space specificity

The Manchester set was used to measure the level of normalisation, calculating the identity weights using Equation 6, and finding the person-mean $\bar{\mathbf{w}}_i$. Better removal of contaminating variance should reduce the variance for each individual relative to this mean. The variance,

$$V = \frac{1}{n_o n_p N} \sum_{i=1}^{n_p} \sum_{k=1}^{n_o} \sum_{j=1}^{N} (w_{kij} - \bar{w}_{ij})^2 \qquad (13)$$

was calculated. The results in Figure 7 show a steady decline in the identity sub-space variance. The only exception is the value for iteration one; this is unusual in having a large increase in the number of dimensions without an opportunity to re-distribute this variation into the other subspaces.

The results of projecting the faces into the other subspaces are shown, as is the variance in the appearance model. As might be expected, these are all higher than the identity sub-space value, and do not show marked declines as the iterations progress.

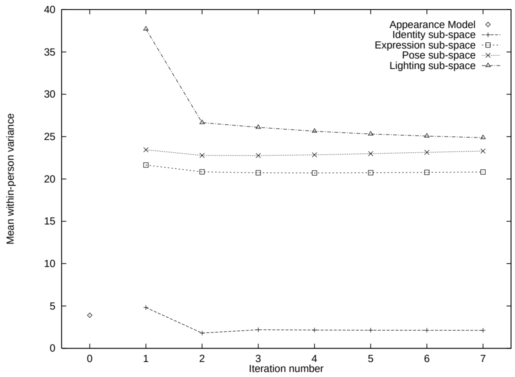

Figure 7. Mean within-person variances for the different subspaces as a function of iteration number.

6.3 Recognition

Recognition was also tested on the Manchester set, coding the images on the final rotated space. The Appearance Model used to provide correspondences, did not give completely accurate positions, degrading recognition accuracy. The pooled covariance matrix was found using Equation 9 on $\mathbf{w}_i$. This allowed

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

where $1 \leq k \leq (n_o \times n_p)$ to give Mahalanobis distances to the mean images. A recognition was scored when the smallest d had the same identity for i and k. The results in Figure 8 demonstrate that relative to the base condition, recognition improves by about one percent by iteration 4. Also shown are the effects of projecting the test images through the complete space to obtain the lighting - pose - expression normalised version, and then coding on the final rotated space. This does not produce an improvement in recognition. It should be noted here that there may well be contingent, non-functional correlations between parameters on different subspaces for individuals (for example, a consistent direction of gaze), whose omission may trade off against theoretically preferable eigenfaces.

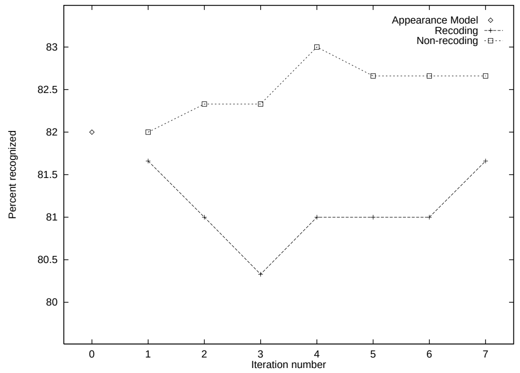

Figure 8. Recognition rates for Euclidean average-image matching.

7 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. In addition, it is 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.

We have shown that both of these problems can be overcome by using an iterative 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. This yields 'cleaner' eigenfaces, with lower within–appropriate–group variance and higher inappropriate–group variance. Both these facts reflect greater orthogonality between the subspaces. In addition, recognition on an entirely disjoint test set was improved, although marginally.

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] 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.

[3] 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.

[4] 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.

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

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

[7] 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.

[8] 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.

[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] 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.

[11] 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.