Statistical Models of Face Images - Improving Specificity

G.J. Edwards, A. Lanitis, C.J. Taylor, T. F. Cootes
Department of Medical Biophysics
University of Manchester
Oxford Road, Manchester, M13 9PT, UK
email: {gje,lan,ctaylor,bim}@sv1.smb.man.ac.uk
Tel: (0161) 275 5130 Fax: (0161) 275 5146

Abstract

Model based approaches to the interpretation of face images have proved very successful. We have previously described statistically based models of face shape and grey-level appearance and shown how they can be used to perform various coding and interpretation tasks. In the paper we describe improved methods of modelling which couple shape and grey-level information more directly than our existing methods, isolate the changes in appearance due to different sources of variability (person, expression, pose, lighting), and deal with non-linear shape variation. We show that the new methods are better suited to interpretation and tracking tasks.

1 Introduction

Model-based approaches to the interpretation and coding of face images have proved very successful. Methods described so far include: Modelling grey-level variation using eigenfaces [1, 2], models based on class specific linear projection [3], combined shape and grey level models[4, 5], models based on the physical and anatomical structure of faces [6], 3D models [7], hand-crafted shape models [8], local non-linear shape manifolds[9] and models based on elastic meshes coubled with local intensity pattern descriptions[10]. Compehensive literature reviews of these techniques and other techniques related to face interpretation can be found in [11,12,13].

The success of a model-based approach relies on the quality of the face model used. In general the models must fulfill two main criteria: generality and specificity. General models are those that account for all possible sources of appearance variation in face images. Specific models constrain the variability allowed so that only 'legal' examples can be generated. In addition to these criteria successful models should be compact and also have the potential to be used in image search algorithms. In the past we have achieved promising results using a model-based approach [14, 15]. In this paper we describe further developments of our models of facial appearance; by using the improved models we aim to improve the performance of our system. In particular we describe how shape and global grey-level variation can be modeled using a single

rather than separate models. We also describe how the different sources of variability can be isolated, given a suitable training set of images. Isolating the sources of variation can be useful in image synthesis and in tracking; where the dynamics of the different sources of variation will differ. The models we have previously described are based on a linear formulation. We present the results of shape modeling and image search experiments using a non-linear formulation for modeling shape. These show that more accurate results are obtained using the non-linear approach.

2 Overview of Our Previous Work

Our approach can be divided into two main phases: modeling, in which flexible models of facial appearance are generated, and interpretation, in which the models are used for coding and interpreting face images. Flexible models [16] are generated from a set of training examples, by statistical analysis. As a result of the analysis training examples can be reconstructed/ parameterized using:

$$X = \overline{X} + Pb$$

where **X** is a training example, $\overline{\mathbf{X}}$ is the mean example, **P** is the matrix of eigenvectors and **b** is a vector of weights , or model parameters.
Flexible models can be used for modeling shape and/or grey-level variation. We model the shapes of facial features and their spatial relationships using a single flexible shape model (a Point Distribution Model) [16]. The effect of the most significant shape parameters is shown in figure 1[1] . Our shape model is augmented with flexible grey-level models using two complementary approaches. In the first we generate a flexible grey-level model of `shape-free' appearance by deforming each face in the training set to have the same shape as the mean face (the effect of the main parameters of this model is shown in figure 2). In the second approach we use a large number of local profile models, one at each landmark point of the shape model. The first approach is more complete but the second is more robust to partial occlusion[14].

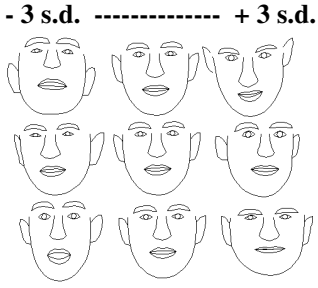

Fig 1. The first 3 modes of shape variation

[1] For the experiments described in this section the Manchester Face Database [17] was used

- 3 s.d. ---------- + 3 s.d.

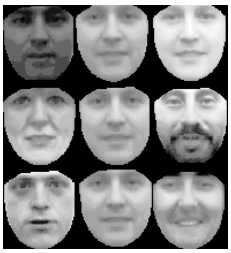

Fig 2. First 3 modes of shape-free grey-level variation

Shape and grey-level models are used together to describe the overall appearance of each face; collectively we refer to the model parameters as appearance parameters. It is important to note that the coding we achieve is reversible - a given face image can be reconstructed from its appearance parameters. When a new image is presented to our system, facial features are located automatically using Active Shape Model (ASM) search [16,18] based on the flexible shape model obtained during training. The resulting automatically located model points are transformed into shape model parameters. Grey-level information at each model point is collected and transformed to local grey-level model parameters. Then the face is deformed to the mean face shape and the grey-level appearance is transformed into the parameters of the shape-free grey-level model. We have presented [14,15] results showing that this representation can be used for image reconstruction, person identification (including gender recognition), expression recognition and pose recovery from static images.

3 Training Combined Shape and Grey-level models

In our previous work we used separate shape and grey-level models to represent facial appearance. However, shape and grey-level variations may be correlated; certain combinations of shape and grey-level modes may correspond to illegal facial reconstructions, thus the overall model is not specific enough. For example the shape mode of variation responsible for opening and closing the mouth is correlated with the grey level mode responsible for the appearance of teeth. We have generated a combined shape and grey-level model in order to overcome this problem. We first train individual shape and shape-free grey-level models [14,15] and convert all training examples to the corresponding model parameters. This results in the representation of training examples by a vector containing both shape and grey-level parameters. Principal component analysis is applied to the new training vectors in order to extract the combined shape and grey-level modes of variation. Before applying the final PCA we scale the shape parameters so that their variance within the training set is equal to the variance of the grey-level parameters. Figure 3 shows the first few modes of the combined shape / grey-level model trained using images from the Home Office Database [19]. Figure 4 shows parametric reconstructions of original images using a combined shape and grey-level model. (The model shown included hair.)

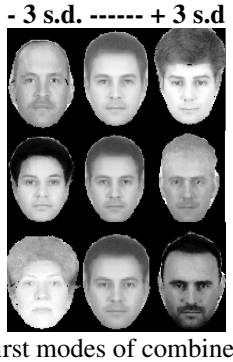

Fig 3. First modes of combined model

Fig 4. Original images with reconstructions

4 Isolating sources of Variation

There are four main sources of appearance variation in face images:

- Pose changes.
- Lighting changes.
- Changes due to difference in individual appearance.
- Changes due to expression, or other face movement, e.g. speaking.

In the models we have presented so far, individual modes of variation tend to be associated with a particular source of variation. However, this is not guaranteed to be the case. If modes corresponding to the different sources of variation could be isolated reliably there would be several benefits. Firstly, for image synthesis applications we could manipulate chosen characteristics without changing others, for example, expression without ID. Secondly, for tracking, we could model the variation of the different components independently, for example, ID modes would be expected to remain constant whilst others would vary over time. Finally we hope that visualization of the modes, for example, the expression modes, will provide insight into the factors involved in recognition.

4.1 Discriminant Analysis

To achieve the desired isolation of the variation arising from the different sources, we employ canonical discriminant analysis over the discrete range of classes of interest. The classification is clear for person ID, where a class corresponds to a particular individual. For expression, we choose a classification based on seven basic expressions; happy, sad, afraid, angry, disgusted, surprised, neutral. The models shown in this paper were trained using the pooled results from experiments involving 30 observers assigning one of seven expressions to each image[2] . The goal of canonical discriminant analysis is to define linear combinations of a set of variables, which separate the classes as well as possible. If there are p variables, in a vector $\mathbf{X}$, the ith discriminant function Z_i is given by:

$$Z_i = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p$$

Finding the coefficients a_{ij} is an eigenvalue problem. The within class covariance matrix **W,** and the total sample covariance matrix **T** are found, and from these the between class covariance is computed (see [21]):

$$\mathbf{B = T - W}$$

The discriminant functions are the eigenvectors of the matrix $\mathbf{W^{-1}B}$, with the corresponding eigenvalues describing the amount of separation, the first function

[2] For the experiments described in this section the Manchester Expression Database[20] was used.

reflecting as much class difference as possible, and so on. For computational simplicity, we perform this analysis on the **b**-vectors for shape and grey-level appearance, obtained using our conventional principle component analysis. The **b**-vector for a particular example is given by:

$$\mathbf{b} = \mathbf{Dd}$$

where **d** is a vector of discriminant parameter weights and **D** is the matrix of unit eigenvectors defining the discriminant functions. The original shape/grey-level vector is therefore given by:

$$\mathbf{X} = \overline{\mathbf{X}} + \mathbf{PDd}$$

The maximum rank of **D** is (no. classes - 1) which is normally less than the number of b -values. Examples can be parameterized and reconstructed using the coefficients of the d-vector (which we call *Discriminant Model Parameters.)* Reconstructions of three canonical discriminant modes for expression are shown in figure 5. Figure 6 shows the equivalent modes for changes between individual (ID modes).

- 2 s.d. ---------- + 2 s.d.

Fig 5 Three major discriminant modes for expression.

- 2 s.d. ---------- + 2 s.d.

Fig 6. Three major discriminant modes for individual.

4.2 Removing Discriminant Modes from the Principal Component Model

Having found a set of discriminant modes, we can project an example to the best least squares approximation in discriminant space. The difference between the actual example and it's approximation is then used as a new training example. The new set of examples should contain no variation due to the modes defined by the discriminant vectors. A principle component model can be constructed using these examples, which should now only display modes of variation orthogonal to the discriminant space. The example in figure 7 shows the first three modes of the principle component model after first removing variation due to change of individual. Encouragingly, there is only slight change in the perceived individual, which indicates a good degree of separability between individual and expression modes of variation. There remains, of course, variation due to pose and lighting conditions.

In figure 8 both expression and ID have been removed which ought to leave only variation due to pose and lighting. It appears to make very little difference in which order the two sources of variation are removed. The training set used in this example did not feature a great deal of pose variation; this accounts for the relatively small amount of variation shown in figure 8.
The idea that the different sources of variation are completely separable is, of course, a simplification. For example, different individuals have characteristic smiles. The results shown in figures 5,6,7 and 8 suggest that the assumption of separability is, however, a useful approximation.

- 2 s.d. ---------- **+ 2 s.d.**

Fig 7. First 3 principle components after ID variation removed

- 2 s.d. ---------- **+ 2 s.d.**

Fig 8. First 2 principle components after removing expression and ID

It is possible to construct the best least-squares fit to an image after removing sources of variation. Figure 9 shows example of this fitting, in which the model has had expression variation removed about the mean happy face. The model is unconstrained in all modes orthogonal to 'expression' but cannot move away from 'happy'. Reconstructing faces in this manner may be very useful in the sythesis of virtual faces in, for example, forensic applications.

Figure 9. Model best-fit with expression variaton removed around 'happy'. Original image on left. Best fit to face patch on right.

5 Using Non-Linear Shape Models

The shape model described in our earlier work[14,15] is based on a linear formulation which may fail when we attempt to model extreme pose variation face images. In this

section we briefly describe how a non-linear shape model can be built using a Multi-Layer Perceptron (MLP) and describe experiments for assessing the goodness of the non-linear model in locating facial features when compared with the results obtained using the linear model.

5.1 Non-Linear PDM's using Multi-Layer Perceptrons

The use of multilayer perceptrons (MLPs) for carrying out non-linear principal component analysis has been described by Kramer [22]. His approach involves training an MLP to give a set of outputs which are as close as possible to the inputs, over a training set of examples.

Recently we have described [23,24] how non-linear PDM's can be formulated using a similar approach. During the training procedure we perform an initial linear PCA on the training shape data. Using the basis functions calculated during this procedure we convert all training examples to principal components and feed them into an MLP of the form shown in figure 10.

During the training phase the weights of the network are adjusted so that the inputs of the network are faithfully reconstructed at the output nodes. The key feature is a "bottleneck" layer with a small number of neurons. In order to achieve outputs equal to the inputs, the MLP is forced to code the data into a number of components equal to the number of neurons in the bottleneck layer thus effecting a non-linear dimension reduction. Once the MLP has been trained using the conjugate gradient decent algorithm, we split it into an encoder and a decoder. We use the encoder to obtain the coded representations of the training set and the decoder to reconstruct training examples given the coded representation. Figure 11 shows schematically the parametrization and reconstruction of training shapes using the non-linear PDM. Non-linear PDM's can be used in image search in a similar scheme as a linear PDM [24]. A combined image search strategy which used the non-linear model in the initial stages of the search followed by refinement using the linear model was also investigated[24].

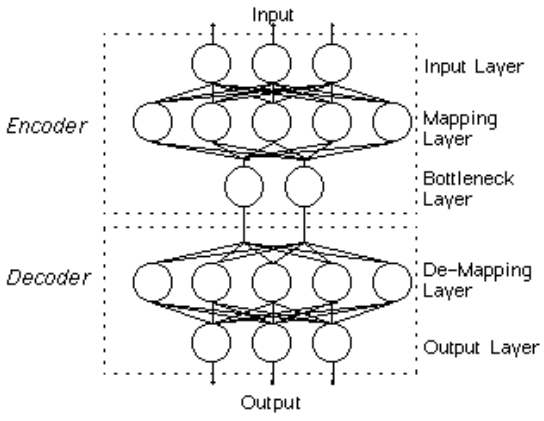

Fig 10. Structure of the MLP

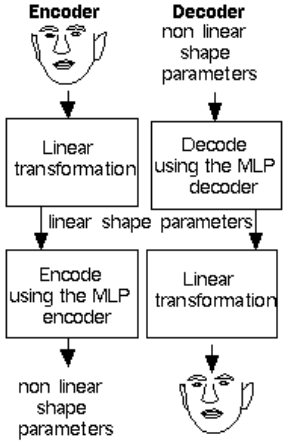

Fig 11. Shape Space / Parameter Space Projections

5.2 Performance of Non-Linear Face Models

We trained a linear PDM and a non-linear PDM (or MLP-PDM) using the same training data as that used previously (see section 2). The linear PDM needed 16 shape parameters to explain 99% of the variability in the training set whilst the MLP-PDM needed only eight. This implies that there where non-linear dependencies between the modes of variation in the linear model and that, as a result, it must be capable of generating illegal solutions. The MLP-PDM model was substantially more compact and thus more specific. The first few significant modes of variation for both the linear PDM and MLP-PDM are similar. We have carried out systematic experiments to compare the performance of linear PDMs and MLP-PDMs in the context of their ability to locate facial features using ASM search [16, 18]. We have tested the fitting procedure by fitting the model to 40 face images. We performed two main experiments. For the first experiment the initial pose was chosen randomly within the following limits: rotation of +/-20 degrees, displacement from the correct position by +/- 30 pixels, and starting scale of 0.6 to 1.4 of the mean scale. These limits usually resulted in a very poor starting point for the iterative search procedure. For the second set of experiments the initial pose was defined within narrower limits: rotation of +/-10 degrees, displacement from the correct position by +/- 10 pixels, and starting scale of 0.8 to 1.2 of the mean scale. For both experiments, the model was initialized to the mean shape. For each test image we fitted the models using three different initial poses giving a total of 120 number of trials. The correct positions for all 144 model points were marked manually on all the test images. At each iteration of the ASM search the goodness of fit, defined as the mean Euclidean distance, d, between the positions of the model points and their correct positions, was calculated. Figure 12 summarizes the results of the experiments; the graphs show the average value of d against the iteration number, over all 120 model fitting trials. In experiment 1 ASM search using an MLP-PDM performs better than the linear PDM. For experiment 2 where the starting position of the model is on average closer to the target, image search using a linear PDM performs better. The performance of the combined method is better in both experiments than search using either a linear PDM or an MLP-PDM alone.

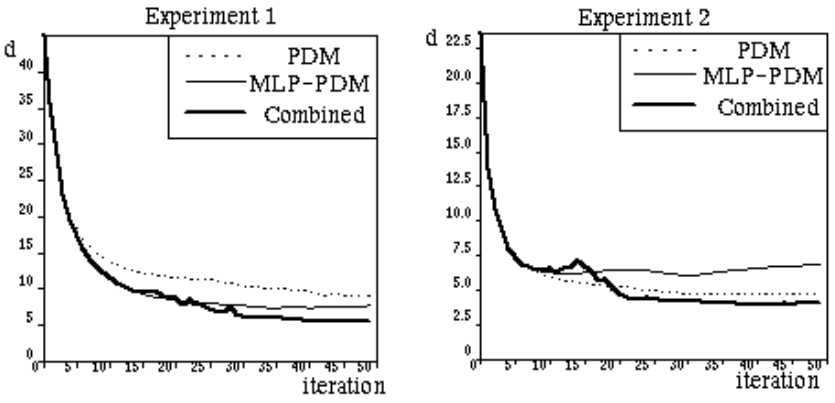

Fig 12. Results for the image search experiments

6. Conclusions

We have described our work in progress on developing improved models of facial appearance. We have shown that it is feasible to model both grey-level and shape variation using a single model resulting in a more specific overall model. By using discriminant analysis techniques we have shown that modes corresponding to different sources of variation can be isolated. The next step is to use these decoupled modes to track faces in image sequences. We have also described how a non-linear PDM can be built and we have presented results for locating facial features. These results show that, when the initial placement of the model is bad, image search using the non-linear shape model performs better than image search using a linear PDM. The reason for this is the ability of the MLP-PDM to be more specific to the class exemplified by the training set. The domain of possible solutions is reduced since only plausible solutions are allowed, resulting in an increased chance of locating image objects successfully. However, when the starting position is good, image search using a linear PDM performs better, since in this case models are unlikely to be driven to illegal shapes. Image search using a combination of linear and non-linear models proved to be the most robust and accurate in our experiments.

7. Acknowledgments

We would like to thank EPSRC and British Telecom for supporting this research. We would also like to thank the Home Office (UK) who provided the Home Office Database, Dr Jane Whittaker from Royal Manchester Children's Hospital and The Manchester Actors Center for providing the Manchester Expression Database, the volunteers who gave face images for the Manchester Face Database, Mr John Connell and Mr Ian Wilks.

8. References

[1] M. Turk and A. Pentland. Eigenfaces for Recognition. Journal of Cognitive Neuroscience, **3**, no 1, pp 71-86, 1991.

[2] I. Craw and P. Cameron. Face Recognition by Computer. Procs of British Machine Vision Conference 1992, pp 489-507, eds. David Hogg and Roger Boyle, Springer Verlag, 1992.

[3] P. Belhumeur, J.Hespanha, D.Kriegman. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. Procs ECCV96 (eds. B.Buxton and R.Cipolla), **1**, pp 45-58, Springer, 1996.

[4] T.F.Cootes and C.J.Taylor. Modelling Object Appearance Using the Grey-Level Surface, Procs of BMVC94 (ed E. Hancock) , **2**, pp 479-488, BMVA Press, 1994.

[5] C.Nastar, B.Moghaddam and A.Pentland. Generalized Image Matching: Statistical Learning of Physically-Based Deformations. Procs ECCV96 (eds. B.Buxton and R.Cipolla), **1**, pp 589-598, Springer, 1996.

[6] D. Terzopoulos, K. Waters. Analysis and Synthesis of Facial Image Sequences Using Physical and Anatomical Models. IEEE Trans. of PAMI, **15**, no 6, pp 569-579, 1993.

[7] H. Li, P. Roivainen, R. Forchheimer. 3-D Motion Estimation in Model-Based Facial Coding. IEEE Trans. of PAMI, **15**, no 6, pp 545-555, 1993.

[8] A.L. Yuille, D.S. Cohen and P. Halliman. Feature Extraction From Faces Using Deformable Templates. International Journal of Computer Vision **8**, pp 104-109, 1992
[9]C. Bregler, S.Omohundro. Nonlinear Manifold Learning for Visual Speech Recognition. Procs. of the 5th International Conference on Computer Vision, pp 494-499, IEEE Computer Society Press, Cambridge, USA, 1995.
[10] M. Lades, J.C. Vorbruggen, J. Buhmann, J. Lange, C. Malsburg, R.P. Wurtz and W. Konen. Distortion Invariant Object Recognition in the Dynamic Link Architecture. IEEE Transactions on Computers, **42**, no 3, pp 300-311, 1993.
[11] R. Chellapa, C.L. Wilson and S. Sirohey. Human and machine Recognition of Faces: A Survey. Procs of the IEEE, **83**, no 5, 1995.
[12] A. Samal and P. Iyengar. Automatic Recognition and Analysis of Human Faces and Facial Expressions: A Survey. Pattern Recognition, **25**, no. 1, pp. 65-77, 1992.
[13] D. Valentin, H.Abdi, A. O'Toole and G.W. Cotrell. Connectionist Models of Face Processing: A Survey. Pattern Recognition, **27**, no. 9, pp. 1209-1230,1994.
[14] A.Lanitis, C.J. Taylor, T.F.Cootes. Automatic Face Indentification System using Flexible Appearance Models, Image and Vision Computing, **13**, no 5, pp 393-401, 1995.
[15] A.Lanitis, C.J. Taylor, T.F.Cootes. A Unified Approach To Coding and Interpreting Face Images. Procs. of the 5th International Conference on Computer Vision, pp 368-373, IEEE Computer Society Press, Cambridge, USA, 1995.
[16] T.F. Cootes, C.J. Taylor, D.H. Cooper and J. Graham. Active Shape Models - Their Training And Application. Computer Vision Graphics and Image Understanding, **61**, no 1, pp 38-59, 1995.
[17] The Manchester Face Database contains 23 images of each of 30 individuals. (10 training, 10 test and three difficult test images per individual). Significant individual variation, pose, expression, and lighting variation exist in this database. The database is publicly available at: http://peipa.essex.ac.uk/ftp/ipa/pix/faces/manchester
[18] T.F. Cootes, C.J. Taylor and A. Lanitis. Active Shape Models: Evaluation of a Multi-Resolution Method For Improving Image Search. Procs. of the 5th British Machine Vision Conference 1994, **1**, pp 327-336, ed. Edwin Hancock, BMVA Press, 1994.
[19] The Home Office Database contains about 500 images, one per individual. Most of the variation in the database is due to inter-individual appearance. The individuals in the database cover a wide age range (20 -60 years), both genders and different ethnic origins.
[20] The Manchester Expression Database contains 400 images (about 22 images from 16 individuals). All 16 individuals were actors who posed realistic facial expressions while listening to descriptions of situations, used to assist their acting. The database contains appearance variations due to individual appearance, lighting, expression and pose.
[21] B.J.F Manly, *Multivariate Statistical Methods, a Primer,* Chapman and Hall, London (1986)
[22] M.A. Kramer, Nonlinear Principal Component Analysis Using Autoassociative Neural Networks, AIChE Journal pp 233-243, 1991.
[23] P.D. Souza, T.F. Cootes, C.J. Taylor, E.C. Di-Mauro. Non-Linear Point Distribution Modelling Using a Multi-Layer Perceptron, Procs of the British Machine Vision Conference, ed. David Pycock, BMVA Press, **1**, pp 107-116, 1995.
[24] A.Lanitis, P.D.Sozou, C.J.Taylor, T.F.Cootes and E.C. Di-Mauro. A General Non-Linear Method For Modelling Shape Variation and Locating Image Objects. To Appear in the Procs of the International Conference of Pattern Recognition, 1996.