Learning to Identify and Track Faces in Image Sequences

G.J. Edwards, C.J. Taylor and T.F. Cootes
Wolfson Image Analysis Unit,
Department of Medical BioPhysics,
University of Manchester, Manchester M13 9PT U.K.
gje@sv1.smb.man.ac.uk

Abstract

We address the problem of robust face identification in the presence of pose, lighting, and expression variation. Previous approaches to the problem have assumed similar models of variation for each individual, estimated from pooled training data. We describe a method of updating a first order global estimate of identity by learning the class-specific correlation between the estimate and the residual variation during a sequence. This is integrated with an optimal tracking scheme, in which identity variation is decoupled from pose, lighting and expression variation. The method results in robust tracking and a more stable estimate of facial identity under changing conditions.

1 Introduction

Locating and interpreting faces in images and image sequences is a difficult problem in machine vision, due to the inherent variability between and within individuals. The appearance of a face in an image varies with the identity of the individual, pose, lighting conditions, and deformations due to expression or speech.

Previous work has shown how the problem can be addressed by using statistical models which combine shape and intensity variation within a single framework. These *Combined Appearance Models* [4], account for all sources of variability in face images. We are interested in isolating the specific sources of variation present in face images, in order to improve identity recognition in the presence of pose, lighting and expression variation, and to allow more robust tracking, by modelling the dynamics of different sources of variability separately. We show how a discriminant analysis method [4] can be used to acheive this to a first-order approximation by assuming the sources of variation are orthogonal and identical for different individuals. This last assumption is necessary because it is unrealistically restrictive to assume a sufficiently large training set for every individual, to determine a class-specific model of variability. We describe how, using image sequences, the first-order approximation to the separation of sources of variability can be improved with a class-specific correction, to give a class-specific representation for particular individuals. This allows a more precise description of identity, and better decoupling of the sources of variation. The decoupling is used to provide separate dynamic models of variation for sequences which can be used in a Kalman filtering framework. We show an example of the method used to track a face in an image sequence, acheiving robust tracking, and yielding a more precise estimate of identity.

2 Background

In many face recognition applications the task is to locate faces in images, and identify them in a way which is robust with respect to changes in pose, expression, and lighting conditions. In this section we outline briefly an existing model-based approach to location and recognition, on which the current work is based.

2.1 Statistical Models

Statistical modelling of facial appearance has proved a successful approach to coding and interpreting face images, also providing a useful basis for locating faces in images. Kirby and Sirovich [7] describe a compact representation of facial appearance, where face images are decomposed into weighted sums of basis images using a Karhumen-Loeve expansion. The patch containing the face is coded using 50 expansion coefficients from which an approximation to the original can be reconstructed. Turk and Pentland [9] describe face identification using this 'Eigenface' representation. Lanitis et al. [8] describe the representation of both face shape and grey-level appearance; they use a *Point Distribution Model* (PDM) [3] to describe shape and an approach similar to Kirby and Sirovich [7] to represent shape-normalised grey-level appearance. More recently, Edwards et al. [4] have described the combination of shape and grey-level variation within a single statistical appearance model, which they call a *Combined Appearance Model*.

2.2 Face Appearance Models

In each of the approaches mentioned above, a feature vector $\mathbf{x}$ which describes the facial appearance - either in terms of shape, intensity or both - is represented by a combination of a small number of parameters, $\mathbf{b}$, which are assumed linearly independent. For example, in the Point Distribution Model [3](PDM) a face shape is coded using

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}\mathbf{b} \tag{1}$$

where $\mathbf{x}$ is an example of a shape, $\bar{\mathbf{x}}$ is the mean shape over the training set and $\mathbf{P}$ is a matrix of the first t eigenvectors of the covariance matrix of the training set. If the training set contains examples of different individuals obtained under varying lighting conditions and showing a range of poses and expressions, it is possible to approximate any plausible face shape by choosing values of $\mathbf{b}$, within limits derived from the training set. Since the eigenvectors which form $\mathbf{P}$ are linearly independent it is possible to rearrange equation 1 to extract the shape parameters, $\mathbf{b}$, for an example $\mathbf{x}$, according to

$$\mathbf{b} = \mathbf{P}^T(\mathbf{x} - \bar{\mathbf{x}}) \tag{2}$$

A face-shape PDM can be used to locate faces in new images by using *Active Shape Model* (ASM) search. The mean shape is projected into the image and iteratively modified to better fit the image evidence, subject to the shape constraints represented by the model. At each step, the region around each model point is searched for the best match to a local grey-level model learnt during training. This gives a new proposed shape. The model constraints are applied by using Equation 2, then Equation 1 to find the closest approximation to the proposed shape consistent with the model.

A Combined Appearance Model [4] can be generated from a set of examples as follows. First the shape parameters for each example are calculated using Equation 2. Next, a warping algorithm [2] is applied to each face patch to deform it to the mean shape; this allows a model of the shape-free grey-level appearance to be built in the form of Equation 1. Finally, the extracted shape and grey-level model parameters for each example are combined and a model is built, again in the form of Equation 1, in which the parameters describe both shape and grey-level variation. The final linear model accounts for correlations between shape and grey-level variation and is more compact than a model which treats the two separately. We have built such a model from a training set containing a wide variety of individuals for a range of poses, expressions and lighting conditions. Figure 1 shows the effect of varying the first few parameters of the Combined Appearance Model.

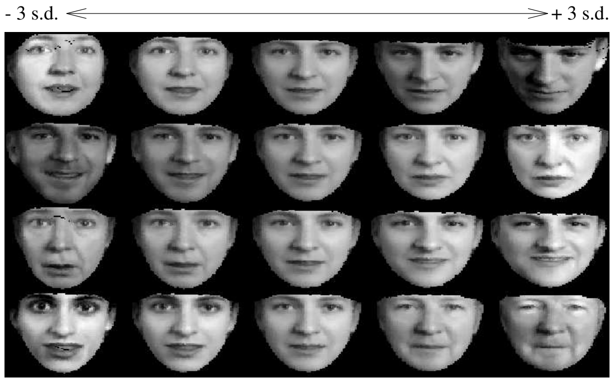

Figure 1. Effect of first few parameters of combined appearance model (+/- 3 standard deviations from mean)

Given an example face we can extract the shape and shape-free grey-level parameters, and approximate the combined model parameters using Equation 2. The reconstruction results obtained using the combined model parameters are shown in Figure 2.

Figure 2. Reconstructing faces using combined appearance parameters. For each face the original is shown on the left, the reconstruction on the right.

2.3 Identification Using Statistical Models

Lanitis et al. [8] describe face recognition using shape and grey-level parameters. In their approach the face is located in an image using Active Shape Model search, and the shape parameters extracted. The face patch is then deformed to the average shape, and the grey-level parameters extracted. The shape and grey-level parameters are used together for classification. As described above, we combine the shape and grey-level parameters and derive Combined

Appearance Model parameters, which can be used in a similar classifier, but providing a more compact model than by than considering shape and grey-level separately.

Given a new example of a face, and the extracted model parameters, the aim is to identify the individual in a way which is invariant to confounding factors such as lighting, pose and expression. If there exists a representative training set of face images, it is possible to do this using the Mahalonobis distance measure [6], which enhances the effect of inter-class variation (identity), whilst suppressing the effect of between class variation (pose,lighting,expression). This gives a scaled measure of the distance of an example from a particular class. The Mahalanobis distance D_i of the example from class i, is given by

$$D_i = (\mathbf{b} - \bar{\mathbf{b}}_i)\mathbf{C}^{-1}(\mathbf{b} - \bar{\mathbf{b}}_i) \tag{3}$$

where $\mathbf{b}$ is the vector of extracted appearance parameters, $\bar{\mathbf{b}}_i$ is the centroid of the multivariate distribution for class i, and $\mathbf{C}$ is the common within-class covariance matrix for all the training examples. Given sufficient training examples for each individual, the individual within-class covariance matrices $\mathbf{C_i}$ could be used - it is, however, restrictive to assume that such comprehensive training data can be obtained.

2.4 Isolating Sources of Variation

The classifier described above assumes that the within-class variation is very similar for each individual, and that the pooled covariance matrix provides a good overall estimate of this variation. Edwards et al. [4] used this assumption to linearly separate the inter-class variability from the intra-class variability using Linear Discriminant Analysis (LDA). The approach seeks to find a linear transformation of the appearance parameters which maximises inter-class variation, based on the pooled within-class and between-class covariance matrices. The identity of a face is given by a vector of *Discriminant Parameters*, $\mathbf{d}$, which ideally only code information important for identity. The transformation between appearance parameters, $\mathbf{b}$, and discriminant parameters, $\mathbf{d}$ is given by

$$\mathbf{b} = \mathbf{D}\mathbf{d} \tag{4}$$

where $\mathbf{D}$ is a matrix of orthogonal vectors describing the principal types of inter-class variation. Having calculated these inter-class *modes of variation*, Edwards et al. [4] showed that a subspace orthogonal to $\mathbf{D}$ could be constructed which modelled only intra-class variations due to change in pose, expression and lighting.The effect of this decomposition is to create a combined model which is still in the form of Equation 1, but where the parameters, $\mathbf{b}$, are partitioned into those that affect identity and those that describe within-class variation. Figure 3 shows the effect of varying the most significant identity parameter for such a model; also shown is the effect of applying the first mode of the residual (identity-removed) model to an example face. It can be seen that the linear separation is reasonably successful and that the identity remains unchanged.

Figure 3. Varying the most significant identity parameter(top), and manipulating residual variation without affecting identity(bottom)

3 Identification and Tracking from Sequences

Separating a combined appearance model into a part that deals with ID and a part that deals with residual variation allows classification of ID independently of confounding factors. It also has potential for applications in model-based tracking of faces. Intuitively, we can imagine different dynamic models for each separate source of variability. In particular, given a sequence of images of the same person we expect the identity to remain constant, whilst lighting, pose and expression vary each with its own dynamics.

In practise, the separation between the different types of variation which can be achieved using LDA is not perfect. The method provides a good first-order approximation, but, in reality, the within-class spread takes a different shape for each. When viewed *for each individual at a time*, there is typically correlation between the identity parameters and the residual parameters, even though for the data *as a whole*, the correlation is minimised.

For example, we can reason that the correlation between pose and identity must be class specific because of the 3D structure of the head; the way in which the appearance of the nose changes with pose, depends partly on its length - a person-specific quantity, not derivable from a frontal view. Ezzat and Poggio [5] describe class-specific normalisation of pose using multiple views of the same person, demonstrating the feasibility of a linear approach. They assume that different views of each individual are available in advance - here, we make no such assumption. We show that the estimation of class-specific variation can be integrated with tracking to make optimal use of both prior and new information in estimating ID and achieving robust tracking.

3.1 Class-Specific Refinement of Recognition from Sequences

We describe a class-specific linear correction to the result of the global LDA, given new examples of a face. To illustrate the problem, we consider a simplified synthetic situation in which appearance is described in some 2-dimensional space as shown in figure 4. We imagine a large number of representative training examples for two individuals, person X and person Y projected into this space. The optimum direction of group separation, $\mathbf{d}$, and the direction of residual variation $\mathbf{r}$, are shown. A perfect discriminant analysis of identity would allow two faces of different pose, lighting and expression to be normalised to a reference view, and thus the identity compared. It is clear from the diagram that an orthogonal projection onto the identity subspace is not ideal for either person X or person Y. Given a fully representative set of training images for X and Y, we could work out in advance the ideal projection. We do not however, wish (or need) to restrict ourselves to acquiring training data in advance. If we wish to identify an example of person Z, for whom we have only one example image, the best estimate possible is the orthogonal projection, A, since we cannot know from a single example whether Z behaves like X (in which case C would be the correct identity) or like Y (when B would be correct) or indeed, neither. The discriminant analysis produces only a first order approximation of class-specific variation.

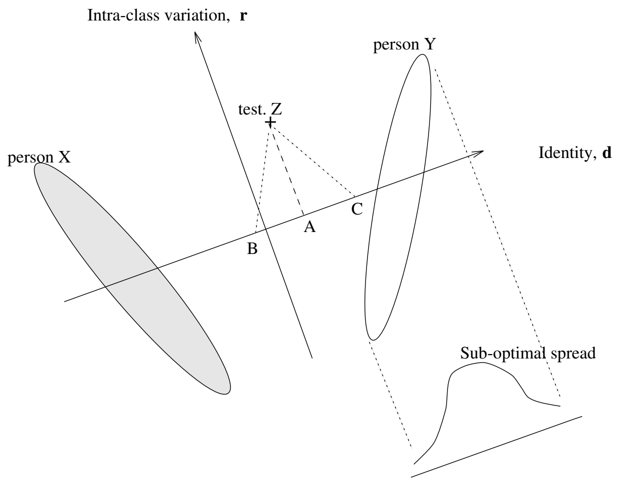

Figure 4. Limitation of Linear Discriminant Analysis: Best identification possible for single example, Z, is the projection, A. But if Z is an individual who behaves like X or Y, the optimum projections should be C or B respectively.

In our approach we seek to calculate class-specific corrections from image sequences. The framework used is the Combined Appearance Model, in which faces are represented by a parameter vector $\mathbf{b}$, as in Equation 1.

LDA is applied to obtain a first order global approximation of the linear variation describing identity, given by an identity vector, $\mathbf{d}$, and the residual linear variation, given by a vector $\mathbf{r}$. A vector of appearance parameters, $\mathbf{b}$ can thus be described by

$$\mathbf{b} = \bar{\mathbf{b}} + \mathbf{D}\mathbf{d} + \mathbf{R}\mathbf{r} \tag{5}$$

where $\mathbf{D}$ and $\mathbf{R}$ are matrices of orthogonal eigenvectors describing identity and residual variation respectively. $\mathbf{D}$ and $\mathbf{R}$ are orthogonal with respect to each other and the dimensions of $\mathbf{d}$ and $\mathbf{r}$ sum to the dimension of $\mathbf{b}$. The projection from a vector, $\mathbf{b}$ onto $\mathbf{d}$ and $\mathbf{r}$ is given by

$$\mathbf{d} = \mathbf{D}^T\mathbf{b} \tag{6}$$

and

$$\mathbf{r} = \mathbf{R}^T\mathbf{b} \tag{7}$$

Equation 6 gives the orthogonal projection onto the identity subspace, $\mathbf{d}$, the best classification available given a single example. We assume that this projection is not ideal, since it is not class-specific. Given further examples, in particular, from a sequence, we seek to apply a class-specific correction to this projection. It is assumed that the correction of identity required has a linear relationship with the residual parameters, but that this relationship is different for each individual.

Formally, if $\mathbf{d}_c$ is the true projection onto the identity subspace, $\mathbf{d}$ is the orthogonal projection, $\mathbf{r}$ is the projection onto the residual subspace, and $\bar{\mathbf{r}}$ is the mean of the residual subspace (average lighting,pose,expression) then,

$$\mathbf{d} - \mathbf{d_c} = \mathbf{A}(\mathbf{r} - \bar{\mathbf{r}}) \tag{8}$$

where $\mathbf{A}$ is a matrix giving the correction of the identity, given the residual parameters. If $\mathbf{d}$ is an p by 1 column vector, and $\mathbf{r}$ an q by 1 column vector, then the matrix $\mathbf{A}$ is p by q.

During a sequence, many examples *of the same face* are seen. We can use these examples to solve Equation 8 in a least-squares sense for the matrix $\mathbf{A}$, thus giving the class-specific correction required for the particular individual. The vector $\mathbf{d_c}$ is unknown, but if we assume that the residual correction is linear, then $\mathbf{A}$ can be found by normalising $\mathbf{d}$ and $\mathbf{r}$ about the local means of the sequence, $\bar{\mathbf{d}}_l$, and $\bar{\mathbf{r}}_l$, writing

$$\mathbf{d}' = \mathbf{d} - \bar{\mathbf{d}}_l \tag{9}$$

and

$$\mathbf{r}' = \mathbf{r} - \bar{\mathbf{r}}_l \tag{10}$$

$$\mathbf{d}' = \mathbf{A}(\mathbf{r}') \tag{11}$$

Let $A_{i,j}$ represent the elements of $\mathbf{A}$ The elements of $\mathbf{d}'$ and $\mathbf{r}'$ are independent and the value of the ith element of $\mathbf{d}'$ is given by

$$d'_i = \sum_{j=1}^{q} A_{i,j} r'_j \tag{12}$$

Thus, each row of $\mathbf{A}$ relates the residual variation, $\mathbf{r}'$, to one of the identity parameters, d'_i. If we have $N > q$ examples of the individual face, we can solve for each row, i, of the correction matrix separately. Let $\mathbf{d}'^{(1\ldots N)}_i$ be a vector of the examples of d'_i seen and $\mathbf{r}'^{(1\ldots N)}$ a matrix of the examples of $\mathbf{r}'$ seen. Let $\mathbf{a}_i$ be row i of the correction matrix, then we can write,

$$\mathbf{d}'^{(1\ldots N)}_i = \mathbf{r}'^{(1\ldots N)} \mathbf{a}_i^T \tag{13}$$

This is simply an overdetermined system of linear equations and can be solved for the elements of $\mathbf{a}_i$ by standard methods. Having found $\mathbf{A}$, we can, given a new example, with measured identity, $\mathbf{d}$, and residual variation, $\mathbf{r}$, solve Equation 8 to find $\mathbf{d_c}$, the corrected identity.

Each column of $\mathbf{A}$ describes the effect of each residual parameter on the correction of identity. The magnitude of the column is a measure of how much new information has been learnt about the corresponding residual parameter. For example, if there is very little lighting change in the sequence, those residual parameters corresponding to lighting will have little effect on the correction, and the estimate will revert to the orthogonal projection in that direction.

3.2 Tracking Face Sequences

In each frame of an image sequence, an Active Shape Model can be used to locate the face. The iterative search procedure returns a set of shape parameters describing the best match found of the model to the data. We can also extract the shape-free grey-level parameters from the extracted shape, and thence calculate the combined appearance model parameters.

Baumberg [1] has described a Kalman filter framework used as a optimal recursive estimator of shape from sequences using an Active Shape Model. In order to improve tracking robustness, we propose a similar scheme, based on the decoupling of identity variation from residual variation. The combined model parameters are projected into the the identity and residual subspaces by Equations 6 and 7. At each frame, t, the identity vector, $\mathbf{d}_t$, and residual vector $\mathbf{r}_t$ are recorded. Until enough frames have been recorded to allow Equation 13 to be solved, the correction matrix, $\mathbf{A}$ is set to contain all zeros, so that the corrected estimate of identity, $\mathbf{d_c}$ is the same as the orthogonally projected estimate, $\mathbf{d}$. Once Equation 13 can be solved, the identity estimate starts to be corrected.

Two sets of Kalman filters are used, one for the corrected identity parameters, in which the underlying model of motion is treated as a zeroth order, or constant position model, and another for the residual parameters, where the motion model is assumed to be first order, or constant velocity. This models the sequence realistically during tracking since the system model treats identity as fixed - something which is certainly true for sequences - and thus the tracking is robust to any noise in the tracking corresponding to apparent change of identity.

3.3 Example

We present an example of this system applied to a face sequence. Figure 5 shows frames selected from a sequence, together with the result of the Kalman filter-based Active Shape Model search overlayed on the image. The filter tracks identity as a zeroth order process and residual variation as a first order process. The subject talks and moves while varying expression. The amount of movement increases towards the end of the sequence.

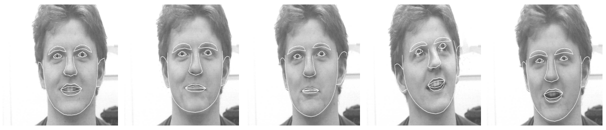

Figure 5. Tracking and identifying a face.

Figure 6 shows the values of the first 3 elements of the corrected identity vector, $\mathbf{d}_c$. Also shown are similar results without the class specific correction applied.

It can be seen that the corrected, filtered identity parameters are much more stable than the raw parameters.

3.4 Enhanced Visualization

After tracking many frames of a sequence the estimate of the corrected identity vector stabilises. A corresponding reconstruction of the person can be synthesized. The synthesized image is based on the evidence integrated over the sequence. This provides a means of generating high resolution reconstructions from lower resolution sequences. Figure 7 illustrates an example: The left hand image is a frame from a sequence of 95 images. In the centre image we show an example from the sequence after deliberate gaussian subsampling to synthesis a low-resolution source image. The reconstruction on the right shows the final estimate of the person based on evidence integrated over the low-resolution sequence.

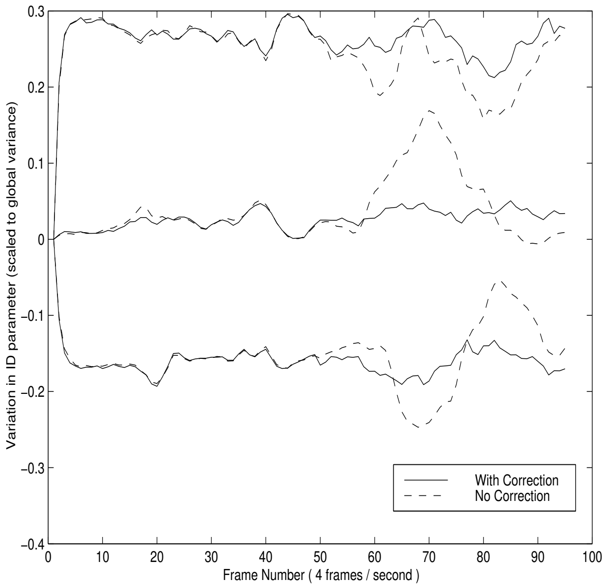

Figure 6. First 3 parameters of corrected and uncorrected identity vectors. Parameters are scaled by their respective variance over the training set.

4 Conclusion

We have outlined a technique for improving the stability of face identification and tracking when subject to variation in pose, expression and lighting conditions. The technique makes use of the observed effect of these types of variation in order to provide a better estimate of identity, and thus provides a method of using the extra information available in a sequence to improve classification. By correctly decoupling the individual sources of variation, it is possible to develop decoupled dynamic models for each. The technique we have described allows the initial approximate decoupling to be updated during a sequence, thus avoiding the need for large numbers of training examples for each individual.

5 Acknowledgements

We would like to thank EPSRC and British Telecom for supporting this research. We would also like to thank to Home Office(UK) and Dr.Jane Wittaker, Royal Manchester Children's hospital for the images used in these experiments. Helpful advice was also provided by Dr. Neil Thacker.

Figure 7. Synthesizing a high-res face from a low-res sequence. Left hand image: an original frame from sequence. Centre image: frame from deliberately blurred sequence. Right hand image: final reconstruction from low-res sequence

References

[1] A. M. Baumberg. *Learning Deformable Models for Tracking Human Motion*. PhD thesis, University of Leeds, 1995.

[2] F. L. Bookstein. Principal Warps: Thin-Plate Splines and the Decompostion of Deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6):567–585, 1989.

[3] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active Shape Models - Their Training and Application. *Computer Vision, Graphics and Image Understanding*, 61(1):38–59, 1995.

[4] G. J. Edwards, A. Lanitis, C. J. Taylor, and T. Cootes. Statistical Models of Face Images: Improving Specificity. In *British Machine Vision Conference 1996*, Edinburgh, UK, 1996.

[5] T. Ezzat and T. Poggio. Facial Analysis and Synthesis Using Image-Based Models. In *International Workshop on Automatic Face and Gesture Recognition 1997*, pages 116–121, Killington, Vermont, 1997.

[6] D. J. Hand. *Discrimination and Classification*. John Wiley and Sons, 1981.

[7] M. Kirby and L. Sirovich. Appliction of the Karhumen-Loeve Procedure for the Characterization of Human Faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1):103–108, 1990.

[8] A. Lanitis, C. Taylor, and T. Cootes. A Unified Approach to Coding and Interpreting Face Images. In 5^{th} *International Conference on Computer Vision*, pages 368–373, Cambridge, USA, 1995.

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