Face Recognition Using Active Appearance Models

G.J. Edwards, T.F. Cootes, and C.J. Taylor

Wolfson Image Analysis Unit,
Department of Medical Biophysics,
University of Manchester,
Manchester M13 9PT, U.K.
gje@sv1.smb.man.ac.uk
http://www.wiau.man.ac.uk

Abstract. We present a new framework for interpreting face images and image sequences using an Active Appearance Model (AAM). The AAM contains a statistical, photo-realistic model of the shape and grey-level appearance of faces. This paper demonstrates the use of the AAM's efficient iterative matching scheme for image interpretation. We use the AAM as a basis for face recognition, obtain good results for difficult images. We show how the AAM framework allows identity information to be decoupled from other variation, allowing evidence of identity to be integrated over a sequence. The AAM approach makes optimal use of the evidence from either a single image or image sequence. Since we derive a complete description of a given image our method can be used as the basis for a range of face image interpretation tasks.

1 Introduction

There is currently a great deal of interest in model-based approaches to the interpretation of images [17] [9] [15] [14][8]. The attractions are two-fold: robust interpretation is achieved by constraining solutions to be valid instances of the model example; and the ability to 'explain' an image in terms of a set of model parameters provides a basis for scene interpretation. In order to realise these benefits, the model of object appearance should be as complete as possible - able to synthesise a very close approximation to any image of the target object.

A model-based approach is particularly suited to the task of interpreting faces in images. Faces are highly variable, deformable objects, and manifest very different appearances in images depending on pose, lighting, expression, and the identity of the person. Interpretation of such images requires the ability to understand this variability in order to extract useful information. Currently, the most commonly required information is the identity of the face.

Although model-based methods have proved quite successful, none of the existing methods uses a full, photo-realistic model and attempts to match it directly by minimising the difference between model-synthesised example and the image under interpretation. Although suitable photo-realistic models exist, (e.g. Edwards *et al* [8]), they typically involve a large number of parameters (50-100) in order to deal with the variability due to differences between individuals, and changes in pose, expression, and lighting. Direct optimisation over such a high dimensional space seems daunting.

We show that a direct optimisation approach is feasible and leads to an algorithm which is rapid, accurate, and robust. We do not attempt to solve a general optimisation each time we wish to fit the model to a new image. Instead, we exploit the fact the optimisation problem is similar each time - we can learn these similarities off-line. This allows us to find directions of rapid convergence even though the search space has very high dimensionality. The main features of the approach are described here - full details and experimental validations have been presented elsewhere[4].

We apply this approach to face images and show first that, using the model parameters for classification we can obtain good results for person identification and expression recognition using a very difficult training and test set of still images. We also show how the method can be used in the interpretation of image sequences. The aim is to improve recognition performance by integrating evidence over many frames. Edwards et. al.[7] described how a face appearance model can be partitioned to give sets of parameters that independently vary identity, expression, pose and lighting. We exploit this idea to obtain an estimate of identity which is independent of other sources of variability and can be straightforwardly filtered to produce an optimal estimate of identity. We show that this leads to a stable estimate of ID, even in the presence of considerable noise. We also show how the approach can be used to produce high-resolution visualisation of poor quality sequences.

1.1 Background

Several model-based approaches to the interpretation of face images of have been described. The motivation is to achieve robust performance by using the model to constrain solutions to be valid examples of faces. A model also provides the basis for a broad range of applications by 'explaining' the appearance of a given image in terms of a compact set of model parameters, which may be used to characterise the pose, expression or identity of a face. In order to interpret a new image, an efficient method of finding the best match between image and model is required.

Turk and Pentland [17] use principal component analysis to describe face images in terms of a set of basis functions, or 'eigenfaces'. The eigenface representationis not robust to shape changes, and does not deal well with variability in pose and expression. However, the model can be fit to an image easily using correlation based methods. Ezzat and Poggio [9] synthesise new views of a face from a set of example views. They fit the model to an unseen view by a stochastic optimisation procedure. This is extremely slow, but can be robust because of the quality of the synthesised images. Cootes *et al* [3] describe a 3D model of the grey-level surface, allowing full synthesis of shape and appearance. However, they do not suggest a plausible search algorithm to match the model to a new image. Nastar *at al* [15] describe a related model of the 3D grey-level surface, combining physical and statistical modes of variation. Though they describe a search algorithm, it requires a very good initialisation. Lades *at al* [12] model shape and some grey level information using Gabor jets. However, they do not impose strong shape constraints and cannot easily synthesise a new instance. Cootes *et al* [5] model shape and local grey-level appearance, using Active Shape Models (ASMs) to locate flexible objects in new images. Lanitis *at al* [14] use this approach to interpret face images. Having found the shape using an ASM, the face is warped into a normalised

frame, in which a model of the intensities of the shape-free face are used to interpret the image. Edwards *at al* [8] extend this work to produce a combined model of shape and grey-level appearance, but again rely on the ASM to locate faces in new images. Our new approach can be seen as a further extension of this idea, using all the information in the combined appearance model to fit to the image. Covell [6] demonstrates that the parameters of an eigen-feature model can be used to drive shape model points to the correct place. We use a generalisation of this idea. Black and Yacoob [2] use local, hand-crafted models of image flow to track facial features, but do not attempt to model the whole face. Our active appearance model approach is a generalisation of this, in which the image difference patterns corresponding to changes in each model parameter are learnt and used to modify a model estimate.

2 Modelling Face Appearance

In this section we outline how our appearance models of faces were generated. The approach follows that described in Edwards *et al* [8] but includes extra grey-level normalisation steps. Some familiarity with the basic approach is required to understand the new Active Appearance Model algorithm.

The models were generated by combining a model of shape variation with a model of the appearance variations in a shape-normalised frame. We require a training set of labelled images, where landmark points are marked on each example face at key positions to outline the main features.

Given such a set we can generate a statistical model of shape variation (see [5] for details). The labelled points on a single face describe the shape of that face. We align all the sets of points into a common co-ordinate frame and represent each by a vector, $\mathbf{x}$. We then apply a principal component analysis (PCA) to the data. Any example can then be approximated using:

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

where $\bar{\mathbf{x}}$ is the mean shape, $\mathbf{P}_s$ is a set of orthogonal *modes of shape variation* and $\mathbf{b}_s$ is a set of shape parameters.

To build a statistical model of the grey-level appearance we warp each example image so that its control points match the mean shape (using a triangulation algorithm). We then sample the grey level information $\mathbf{g}_{im}$ from the *shape-normalised* image over the region covered by the mean shape. To minimise the effect of global lighting variation, we normalise this vector, obtaining $\mathbf{g}$. For details of this method see[4].

By applying PCA to this data we obtain a linear model:

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{P}_g \mathbf{b}_g \tag{2}$$

where $\bar{\mathbf{g}}$ is the mean normalised grey-level vector, $\mathbf{P}_g$ is a set of orthogonal *modes of grey-level variation* and $\mathbf{b}_g$ is a set of grey-level model parameters.

The shape and appearance of any example can thus be summarised by the vectors $\mathbf{b}_s$ and $\mathbf{b}_g$. Since there may be correlations between the shape and grey-level variations, we apply a further PCA to the data as follows. For each example we generate the concatenated vector

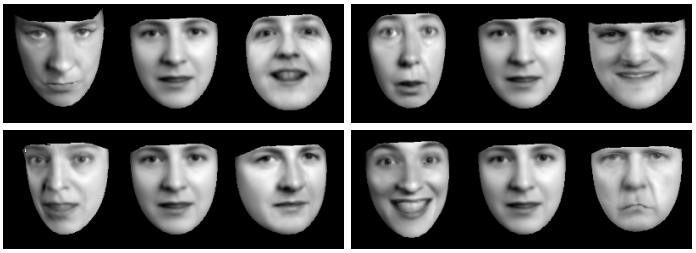

Fig. 1. First four modes of appearance variation (+/- 3 sd)

$$\mathbf{b} = \begin{pmatrix} \mathbf{W}_s \mathbf{b}_s \\ \mathbf{b}_g \end{pmatrix} = \begin{pmatrix} \mathbf{W}_s \mathbf{P}_s^T (\mathbf{x} - \bar{\mathbf{x}}) \\ \mathbf{P}_g^T (\mathbf{g} - \bar{\mathbf{g}}) \end{pmatrix} \tag{3}$$

where $\mathbf{W}_s$ is a diagonal matrix of weights for each shape parameter, allowing for the difference in units between the shape and grey models. We apply a PCA on these vectors, giving a further model

$$\mathbf{b} = \mathbf{Q}\mathbf{c} \tag{4}$$

where $\mathbf{Q}$ are the eigenvectors of $\mathbf{b}$ and $\mathbf{c}$ is a vector of *appearance* parameters controlling both the shape and grey-levels of the model. Since the shape and grey-model parameters have zero mean, $\mathbf{c}$ does too.

An example image can be synthesised for a given $\mathbf{c}$ by generating the shape-free grey-level image from the vector $\mathbf{g}$ and warping it using the control points described by $\mathbf{x}$. Full details of the modelling procedure can be found in [4].

We applied the method to build a model of facial appearance. Using a training set of 400 images of faces, each labelled with 122 points around the main features. From this we generated a shape model with 23 parameters, a shape-free grey model with 113 parameters and a combined appearance model which required only 80 parameters required to explain 98% of the observed variation. The model used about 10,000 pixel values to make up the face patch.

Figure 1 shows the effect of varying the first four appearance model parameters.

3 Active Appearance Model Search

Given the photo-realistic face model, we need a method of automatically matching the model to image data. Given a reasonable starting approximation, we require an efficient algorithm for adjusting the model parameters to match the image. In this section we give an overview of such an algorithm. Full technical details are given in [4].

3.1 Overview of AAM Search

Given an image containing a face and the photo-realistic face model, we seek the optimum set of model parameters (and location) that best describes the image data. One

metric we can use to describe the match between model and image is simply $\delta\mathbf{I}$, the vector of differences between the grey-level values in the image and a corresponding instance of the model. The quality of the match can be described by $\Delta = |\delta\mathbf{I}|^2$. As a general optimization problem, we would seek to vary the model parameters while minimizing Δ. This represents an enormous task, given that the model space has 80 dimensions. The Active Appearance Model method uses the full vector $\delta\mathbf{I}$ to drive the search, rather than a simple fitness score. We note that each attempt to match the model to a new face image is actually a similar optimisation problem. Solving a general optimization problem from scratch is unnecessary. The AAM attempts to learn something about how to solve this class of problems in advance. By providing a-priori knowledge of how to adjust the model parameters during during image search, an efficient run-time algorithm results. In particular, the AAM uses the spatial pattern in $\delta\mathbf{I}$, to encode information about how the model parameters should be changed in order to achieve a better fit. For example, if the largest differences between a face model and a face image occurred at the sides of the face, that would imply that a parameter that modified the width of the model face should be adjusted.

Cootes *et al.*[4] describe the training algorithm in detail. The method works by learning from an annotated set of training example in which the 'true' model parameters are known. For each example in the training set, a number of known model displacements are applied, and the corresponding difference vector recorded. Once enough training data has been generated, multivariate multiple regression is applied to model the relationship between the model displacement and image difference.

Image search then takes place by placing the model in the image and measuring the difference vector. The learnt regression model is then used to predict a movement of the face model likely to give a better match. The process is iterated to convergence. In our experiments, we implement a multi-resolution version of this algorithm, using lower resolution models in earlier stages of a search to give a wider location range. The model used contained 10,000 pixels at the highest level and 600 pixels at the lowest.

4 Face Recognition using AAM Search

Lanitis et al. [13] 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 Appearance Model parameters, which can be used in a similar classifier, but providing a more compact model than that obtained by 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 [11], which enhances the effect of inter-class variation (identity), whilst suppressing the effect of within class variation (pose,lighting,expression). This gives a scaled measure of the distance of an

Fig. 2. Varying the most significant identity parameter(top), and manipulating residual variation without affecting identity(bottom)

example from a particular class. The Mahalanobis distance D_i of the example from class i, is given by

$$D_i = (\mathbf{c} - \bar{\mathbf{c}}_i)\mathbf{C}^{-1}(\mathbf{c} - \bar{\mathbf{c}}_i) \tag{5}$$

where $\mathbf{c}$ is the vector of extracted appearance parameters, $\bar{\mathbf{c}}_i$ 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 is available.

4.1 Isolating Sources of Variation

The classifier described earlier 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. [7] use 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 to identity. The transformation between appearance parameters, $\mathbf{c}$, and discriminant parameters, $\mathbf{d}$ is given by

$$\mathbf{c} = \mathbf{D}\mathbf{d} \tag{6}$$

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. [7] 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, $\mathbf{c}$, are partitioned into those that affect identity and those that describe within-class variation. Figure 2 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.

The 'identity' subspace constructed gives a suitable frame of reference for classifica-

Fig. 3. Original image (right) and best fit (left) given landmark points

Fig. 4. Multi-Resolution search from displaced position

tion. The euclidean distance between images when projected onto this space is a measure of the similarity of ID between the images, since discriminant analysis ensures that the effect of confounding factors such as expression is minimised.

4.2 Search Results

A full analysis of the robustness and accuracy of AAM search is beyond the scope of this paper, but is described elsewhere[4]. In our experiments, we used the face AAM to search for faces in previously unseen images. Figure 3 shows the best fit of the model given the image points marked by hand for three faces. Figure 4 shows frames from a AAM search for each face, each starting with the mean model displaced from the true face centre.

4.3 Recognition Results

The model was used to perform two recognition experiments; recognition of identity, and recognition of expression. In both tests 400 faces were used - 200 for training and

Fig. 5. Typical examples from the experimental set

200 for testing. The set contained images of 20 different individuals captured under a range of conditions. This particular set of faces was chosen for its large range of expression changes as well as limited pose and lighting variation. These factors, the *within class* variability, serve to make the recognition tasks much harder than with controlled expression and pose. Figure 5 shows some typical examples from the set. The active appearance model was used to locate and interpret both the training and test images. In both cases the model was given the initial eye positions, and was then required to fit to the face image using the strategy described in section 3. Thus, for each face, a set of model parameters was extracted, and the results used for classification experiments.

4.4 Recognising Identity

The identity recognition was performed in the identity subspace as described in section 4.1. Each example vector of extracted model parameters was projected onto the ID-subspace. The training set was used to find the centroid, in the ID-subspace for each of the training faces. A test face was then classified according to the nearest centroid of the training set. In order to quantify the performance of the Active Appearance Model for location and interpretation, we compared the results with the best that could be achieved using this classifier with hand annotation. For each example (training and test) the 122 key-landmark points were placed by hand, and the model parameters extracted from the image as described in section 2. Using the above classifier, this method achieved 88% correct recognition. When the active appearance model was applied to the same images, the recognition rate remained at 88%. Although this represents equal performance with hand-annotation, a few of the failures were on different faces from the hand-annotated results. Thus we can conclude that the Active Appearance Model competes with hand annotation; any further improvement in classification rate requires addressing the classifier itself.

4.5 Recognising Expression

In order to test the performance of the Active Appearance Model for expression recognition, we tested the system against 25 human observers. Each observer was shown the set of 400 face images, and asked to classify the expression of each as one of: *happy, sad, afraid, angry, surprised, disgusted, neutral*. We then divided the results into two separate blocks of 200 images each, one used for training the expression classifier, and the other used for testing. Since there was considerable disagreement amongst the

human observers as to the correct expression, it was necessary to devise an objective measure of performance for both the humans and the model. A leave-one-out based scheme was devised thus: Taking the 200 test images, each human observer attached a label to each. This label was then compared with the label attached to that image by the 24 *other* observers. One point was scored for every agreement. In principle this could mean a maximum score of 24x200 = 4800 points, however, there were very few cases in which all the human observers agreed, so the actual maximum is much less. In order to give a performance baseline for this data, the score was calculated several times by making random choices alone. The other 200 images were used to train an expression classifier based on the model parameters. This classifier was then tested on the same 200 images as the human observers. The results were as follows:

Random choices score	660 +/- 150
Human observer score	2621 +/- 300
Machine score	1950

Although the machine does not perform as well as any of the human observers, the results encourage further exploration. The AAM search results are extremely accurate, and the ID recognition performance high. This suggests that expression recognition is limited by the simple linear classifier we have used. Further work will address a more sophisticated model of human expression characterisation.

5 Tracking and Identification from Sequences

In many recognition systems, the input data is actually a sequence of images of the same person. In principal, a greater amount of available is information than from a single image, even though any single frame of video may contain much less information than a good quality still image. We seek a principled way of interpreting the extra information available from a sequence. Since faces are deformable objects with highly variable appearance, this is a difficult problem. The task is to combine the image evidence whilst filtering noise, the difficulty is knowing the difference between real temporal changes to the data (eg. the person smiles) and changes simply due to systematic and/or random noise.

The model-based approach offers a potential solution - by projecting the image data into the model frame, we have a means of registering the data from frame to frame. 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 fact, most of the variation in the model is due to changes between individuals, variation which does not occur in a sequence. If this variation could be held constant we would expect more robust tracking, since the model would more specifically represent the input data.

Edwards et. al.[7] show that LDA can be used to partition the model into ID and non-ID subspaces as described in section 4.1. This provides the basis for a principled method of integrating evidence of identity over a sequence. If the model parameter for each frame are projected into the identity subspace, the expected variation over the

sequence is zero and we can apply an appropriate filter to achieve robust tracking and an optimal estimate of identity over the sequence.

Although useful, 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 person. 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.

Ezzat and Poggio [10] 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.

5.1 Class-Specific Refinement of Recognition from Sequences

In our approach, we reason that the imperfections of LDA when applied to a specific individual can be modelled by observing the behaviour of the model during a sequence. We describe a class-specific linear correction to the result of the global LDA, given a sequence 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 6. 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, **d**, and the direction of residual variation **r**, are shown. A perfect discriminant

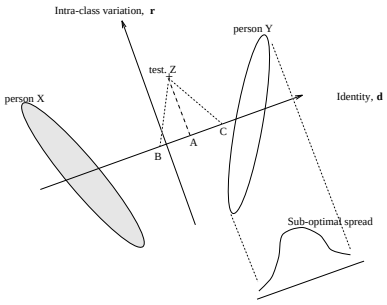

Fig. 6. Limitation of Linear Discriminant Analysis: Best identification possible for single example, Z, is the projection, A. But if Z is an individual who behaves like X or Y, the optimum projections should be C or B respectively.

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 to class-specific variation.
In our approach we seek to calculate class-specific corrections from image sequences. The framework used is the Appearance Model, in which faces are represented by a parameter vector $\mathbf{c}$, as in Equation 1.
LDA is applied to obtain a first order global approximation of the linear subspace 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, $\mathbf{c}$ can thus be described by

$$\mathbf{c} = \bar{\mathbf{c}} + \mathbf{D}\mathbf{c} + \mathbf{R}\mathbf{r} \tag{7}$$

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 $\mathbf{c}$. The projection from a vector, $\mathbf{b}$ onto $\mathbf{d}$ and $\mathbf{r}$ is given by

$$\mathbf{d} = \mathbf{D}^T\mathbf{c} \tag{8}$$

and

$$\mathbf{r} = \mathbf{R}^T\mathbf{c} \tag{9}$$

Equation 8 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,

$$\mathbf{d} - \mathbf{d_c} = \mathbf{A}(\mathbf{r} - \bar{\mathbf{r}}) \tag{10}$$

where $\mathbf{A}$ is a matrix giving the correction of the identity, given the residual parameters. During a sequence, many examples *of the same face* are seen. We can use these examples to solve Equation 10 in a least-squares sense for the matrix $\mathbf{A}$, by applying linear regression, thus giving the class-specific correction required for the particular individual.

5.2 Tracking Face Sequences

In each frame of an image sequence, an Active Appearance Model can be used to locate the face. The iterative search procedure returns a set of parameters describing the

best match found of the model to the data. Baumberg [1] and Rowe et. al. [16] 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, but using the full Appearance Model, and 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 8 and 9. 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 linear regression to be applied, 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 regression can be applied, the identity estimate starts to be corrected. Three sets of Kalman filters are used to track the face. Each track 2D-pose, $\mathbf{p}$, ID variation, $\mathbf{d_{id}}$, and non-ID, $\mathbf{d_{res}}$, variation respectively. The 2D-pose and non-ID variation are modelled as random-walk processes, the ID variation is modelled as a random constant, reflecting the expected dynamics of the system. The optimum parameters controlling the operation of Kalman filters can be estimated from the variation seen over the training set. For example, the ID filter is initialised on the mean face, with a estimated uncertainty covering the range of ID seen during training.

6 Tracking Results

In order to test this approach we took a short sequence of an individual reciting the alphabet whilst moving. We then successively degraded the sequence by adding Gaussian noise at 2.5,5,7.5,10,12.5 and 30% average displacement per pixel. Figure 7 shows frames selected from the uncorrupted sequence, together with the result of the Active Appearance Model search overlaid on the image. The subject talks and moves while varying expression. The amount of movement increases towards the end of the sequence.
After 40 frames the adaptive correction and Kalman filtering was switched on. We first show the results for the uncorrupted sequence. Figure 8 shows the value of the raw projection onto the first and second ID parameters. Considerable variation is observed over the sequence. The corrected, and the final, filtered estimates of the ID parameters are shown in figures 9 and 10 respectively. Figures 9 shows that, once the ID correction is switched on (at frame 40), a more stable estimate of ID results. Figure 10 shows that the combination of ID correction and temporal filtering results in an extremely stable estimate of ID. Figure 11 illustrates the stability of the ID estimate with image degradation. The value of the first ID parameter is shown on the y-axis . This is normalised over the total variation in ID-value over the training set. It is seen that the estimate remains reasonably consistent (within +/- 0.03% of the overall variation) at low levels of degradation, becoming unstable at a higher level.

7 Enhanced Visualisation

After tracking many frames of a sequence the estimate of the corrected identity vector stabilises. A corresponding reconstruction of the person can be synthesised. The syn-

Fig. 7. Tracking and identifying a face. Original frames are shown on the top row, reconstruction on the bottom.

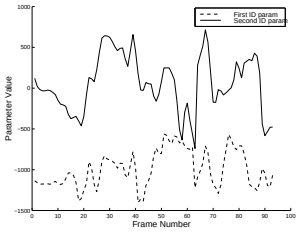

Fig. 8. Raw ID parameters

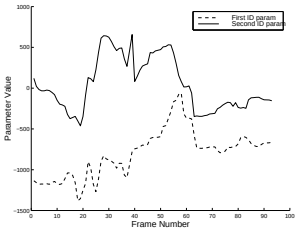

Fig. 9. Corrected ID parameters

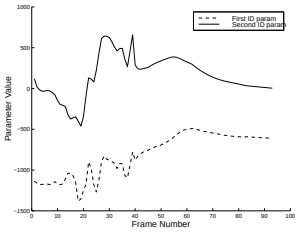

Fig. 10. Filtered, corrected, ID

thesised image is based on the evidence integrated over the sequence. This provides a means of generating high resolution reconstructions from lower resolution sequences. Figure 12 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.

8 Conclusions

We have described the use of an Active Appearance Model in face recognition. The model uses all the information available from the training data and facilitates the decoupling of model into ID and non-ID parts.

When used for static face identification the AAM proved as reliable as labelling the images by hand. A identification rate of 88% was achieved. When used for expression recognition the systems shows less agreement than human observers but nevertheless encourages further work in this area. A observation of the quality of model fit, and the excellent identity recognition performance suggests that the classifier itself rather than the AAM search limits the expression recognition performance.

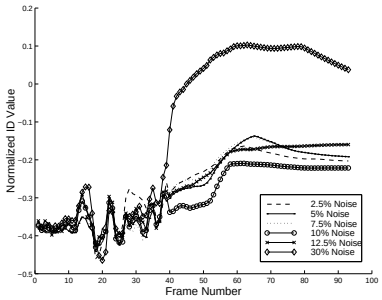

Fig. 11. Tracking Noisy Data. ID estimate remains consistent at increasing noise levels, becoming unstable at 30% noise level.

Fig. 12. Synthesising 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

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 tracking 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.

References

1. A. M. Baumberg. *Learning Deformable Models for Tracking Human Motion.* PhD thesis, University of Leeds, 1995.
2. M. J. Black and Y. Yacoob. Recognizing Facial Expressions under Rigid and Non-Rigid Facial Motions. In *International Workshop on Automatic Face and Gesture Recognition 1995*, pages 12–17, Zurich, 1995.

3. T. Cootes and C. Taylor. Modelling object appearance using the grey-level surface. In E. Hancock, editor, 5^{th} *British Machine Vison Conference*, pages 479–488, York, England, September 1994. BMVA Press.
4. T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *ECCV98 (to appear)*, Freiberg, Germany, 1998.
5. 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(1):38–59, 1995.
6. M. Covell. Eigen-points: Control-point Location using Principal Component Analysis. In *International Workshop on Automatic Face and Gesture Recognition 1996*, pages 122–127, Killington, USA, 1996.
7. 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.
8. G. J. Edwards, C. J. Taylor, and T. Cootes. Learning to Identify and Track Faces in Image Sequences. In *British Machine Vision Conference 1997*, Colchester, UK, 1997.
9. T. Ezzat and T. Poggio. Facial Analysis and Synthesis Using Image-Based Models. In *International Workshop on Automatic Face and Gesture Recognition 1996*, pages 116–121, Killington, Vermont, 1996.
10. T. Ezzat and T. Poggio. Facial Analysis and Synthesis Using Image-Based Models. In *International Workshop on Automatic Face and Gesture Recognition 1996*, pages 116–121, Killington, Vermont, 1996.
11. D. J. Hand. *Discrimination and Classification*. John Wiley and Sons, 1981.
12. M. Lades, J. Vorbruggen, J. Buhmann, J. Lange, C. von der Malsburt, R. Wurtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42:300–311, 1993.
13. 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.
14. A. Lanitis, C. Taylor, and T. Cootes. Automatic Interpretation and Coding of Face Images Using Flexible Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):743–756, 1997.
15. C. Nastar, B. Moghaddam, and A. Pentland. Generalized Image Matching: Statistical Learning of Physically-Based Deformations. In 4^{th} *European Conference on Computer Vision*, volume 1, pages 589–598, Cambridge, UK, 1996.
16. S. Rowe and A. Blake. Statistical Feature Modelling for Active Contours. In 4^{th} *European Conference on Computer Vision*, volume 2, pages 560–569, Cambridge, UK, 1996.
17. M. Turk and A. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.