Modelling Faces Dynamically Across Views and Over Time

Yongmin Li, Shaogang Gong and Heather Liddell
Department of Computer Science, Queen Mary, University of London
London, E1 4NS, UK {yongmin, sgg, heather}@dcs.qmw.ac.uk

Abstract

A comprehensive novel multi-view dynamic face model is presented in this paper to address two challenging problems in face recognition and facial analysis: modelling faces with large pose variation and modelling faces dynamically in video sequences. The model consists of a sparse 3D shape model learnt from 2D images, a shape-and-pose-free texture model, and an affine geometrical model. Model fitting is performed by optimising (1) a global fitting criterion on the overall face appearance whilst it changes across views and over time, (2) a local fitting criterion on a set of landmarks, and (3) a temporal fitting criterion between successive frames in a video sequence. By temporally estimating the model parameters over a sequence input, the identity and geometrical information of a face is extracted separately. The former is crucial to face recognition and facial analysis. The latter is used to aid tracking and aligning faces. We demonstrate the results of successfully applying this model on faces with large variation of pose and expression over time.

1. Introduction

The issue of face recognition and facial analysis (facial expression, ageing, and caricature) has been extensively addressed in recent years. Various approaches including Eigenfaces [16], Elastic Graph model [11], Linear Object Classes [18], Active Shape Models (ASMs) [3] and Active Appearance Models (AAMs) [2] have shown to be promising under different assumptions.

1.1. Modelling Faces with Large Pose Variation

In particular, modelling faces across views is one of the most challenging problems because of self-occlusion, self-shading, and the consequent non-linearity in both shape and texture. Both ASMs and AAMs are unfortunately restricted to a narrow view due to the linear assumption of the 2D appearance. To address this problem, Romdhani et al [15] developed a multi-view appearance model using Kernel Principal Component Analysis (KPCA). The non-linearity of KPCA enables the model to deal with large pose variation, but has a price of intensive computation. Cootes *et al.* [4] proposed the view-based Active Appearance Models which employ three models for profile, half-profile and frontal views. On the other hand, Moghaddam and Pentland [13] presented a view-based and modular eigenspace method. Li *et al.* [12] introduced a view-based piece-wise SVM (Support Vector Machine) model of the face space. But the division of the face space in these methods is rather arbitrary and often coarse, therefore ad hoc.

An alternative approach to alleviate the multi-view problem is to use 3D models. DeCarlo and Metaxas [5] presented a 3D deformable face model in which optical flow and edge information are combined. Their model successfully tracked faces in sequences with significant expression change and pose change. Jebara and Pentland [10] proposed an approach to recover the 3D face structure using *Structure from Motion*. The estimation of the 3D structure is further constrained for reliable feature tracking by a 3D range data model of an average human face. Vetter and Blanz [17] introduced a flexible 3D face model learnt from examples of 3D range face data. A novel 2D face image can be matched to the 3D model in an *analysis-by-synthesis* manner. Then images of the novel face in different views, illumination, and expression can be synthesised by changing the parameters of the matched model.

1.2. Modelling Faces Dynamically on Sequences

In parallel to modelling faces across views, the issue of exploiting the face dynamics using spatial-temporal information from video sequences has also received great interest. From video sequences, not only can more information about the visual objects be acquired, but also the temporal continuity and subject constancy can provide a more robust representation [8]. Gong et al. [9] introduced an approach that uses Partially Recurrent Neural Networks to recognise temporal signatures of faces. Edwards *et al.* [6] proposed an integrated approach to decouple the identity variance from the residual variance of pose, lighting and expression. By learning the correlation between the two parts of variance online, a class-specific refinement for the identity covariance can be achieved. Yamaguchi et al. [19] presented a method for face recognition from sequences by building a subspace for the detected faces from a given sequence and then matching the subspace with prototype subspaces.

1.3. Overview of this Work

To comprehensively address the two problems stated above, we present an integrated multi-view dynamic face model. It consists of three parts: a sparse 3D shape model trained from 2D images labelled with pose and landmarks, a *shape-and-pose-free* texture model, and an affine geometrical model. Section 2 gives the details of model components and model construction. A model fitting algorithm is presented in Section 3 formulated by optimising the global fitting criterion of the overall face appearance, the local fitting criterion on a set of 2D landmarks, and the temporal fitting criterion between the information on successive frames of a sequence. Section 4 describes the issue of temporal model fitting, i.e. obtaining a robust estimation of model parameters dynamically from sequences where faces are undergoing large pose and expression changes. Conclusions are drawn in Section 5.

2. Multi-View Dynamic Model

Our multi-view dynamic face model consists of a sparse 3D Point Distribution Model (PDM) [3] learnt from 2D images in different views, a *shape-and-pose-free* texture model, and an affine geometrical model which controls the rotation, scale and translation of faces. The first two parts of the model aim to represent the identities of faces to be analysed, while the latter is used for alignment and tracking.

2.1. Constructing 3D Shape from 2D Images

As the 2D appearance of different people from the same view can be more similar than that of one person at different views, the problem of modelling the appearance of faces with large pose variation is non-trivial for 2D models. But if 3D geometrical information is available, this situation can be alleviated to some extent. A straight forward way to collect 3D information about faces is using sensors such as a 3D laser scanner. However, the huge amount of 3D range data may bring a heavy burden to the computation. Another difficulty comes from establishing the correspondence between the dense 3D data. In this work, we learn a 3D face shape model containing only a sparse set of feature points from 2D face images in different views.

2.1.1. Database of 2D Multi-view Faces

The database used in this work includes 2D face images from 31 subjects, 133 poses of each subject (see [8] for more details of the data acquisition process). The pose of a face is defined by two parameters: tilt and yaw (α, β), the rotation angles about horizontal and vertical axes respectively. The rotation in image plane is not taken into account on the basis that human heads are assumed mostly upright.

A sparse set of 44 landmarks locating the mouth, nose, eyes, and face contour were semi-automatically labelled on each face image.

2.1.2. Estimating the 3D Shape

Given a set of 2D face images with known positions of the landmarks and pose, the 3D positions of the landmarks can be estimated using linear regression. The rotation centre used to measure the pose angles is assumed to be the centre point of the eye centres and the mouth centre. We set this point as the origin of the object coordinate system.

Orthographic projection is adopted for simplicity. Suppose the 3D coordinates of a landmark in the object coordinate system is (X, Y, Z), the position of this landmark in the 2D image with pose (α, β) is given by:

$$(x, y)^{\mathrm{T}} = \mathbf{R}(X, Y, Z)^{\mathrm{T}} \tag{1}$$

where $\mathbf{R}$ is the rotation matrix for pose (α, β) obtained by rotating about the horizontal axis first by α and then about the vertical axis by β.

$$\mathbf{R} = \left[\begin{array}{lll} cos(\beta) & 0 & sin(\beta) \\ sin(\alpha)sin(\beta) & cos(\alpha) & -sin(\alpha)cos(\beta) \end{array} \right] \tag{2}$$

Note that the results are only slightly different if rotating in the reverse order, i.e. first β, then α.

If $M(M \geq 2)$ face images in different poses are available, one can estimate the 3D coordinates (X, Y, Z) of a landmark using linear regression by minimising

$$\sum_{i=1}^{M} \left((x - x_i)^2 + (y - y_i)^2 \right) \tag{3}$$

where (x_i, y_i) is the known 2D position of the landmark. Then the 3D shape vector $\mathbf{p}$ is obtained as:

$$\mathbf{p} = (X_1, Y_1, Z_1, X_2, Y_2, Z_2, ..., X_{N_l}, Y_{N_l}, Z_{N_l})^{\mathrm{T}} \tag{4}$$

where N_l is the number of landmarks,

Ideally, the larger the range of poses covered by the training images, the more accurate the 3D position. However, when a face rotates to nearly profile view, some of the landmarks are invisible in the image. Therefore, for each subject, 45 of the 133 face images with poses between $[-20° , 20°]$ in tilt and $[-40° , 40°]$ in yaw are selected for training. Also, the training set M should be big enough. In our experiments, a random selection of 20 out of 45 face images from each subject is used to learn the 3D shape vector of all landmarks. For each subject, 50 shape vectors are estimated in this manner to learn the statistical 3D PDM of faces.

2.2. A Sparse 3D PDM of Faces

Although only a sparse set of 44 landmarks are chosen to represent the 3D shape of faces, the dimensionality is still too high to fit the shape model. However, human faces can be represented in an abstract low dimensional shape space since they are actually share a similar structure. The PDM is adopted to construct this low dimensional shape space.

Performing Principal Component Analysis (PCA) on N given 3D face shape vectors $\{\mathbf{p}_i, i = 1, 2, ..., N\}$ which are estimated using the method described in Section 2.1.2, one obtains the mean shape $\bar{\mathbf{p}}$ and the eigen matrix $\mathbf{U}$ which is comprised of the first N_s significant eigen vectors

$$\mathbf{U} = [\mathbf{u}_1 \mathbf{u}_2 ... \mathbf{u}_{N_s}] \tag{5}$$

Then a shape pattern $\mathbf{p}$ can be represented by a vector in the PDM space

$$\mathbf{s} = \mathbf{U}^{\mathrm{T}}(\mathbf{p} - \bar{\mathbf{p}}) \tag{6}$$

whose dimension is N_s. The reconstructed 3D shape from $\mathbf{s}$ is obtained from

$$\mathbf{p}_r = \mathbf{U}\mathbf{s} + \bar{\mathbf{p}} \tag{7}$$

We trained the PDM on a set of 600 3D shape patterns from 12 different subjects (50 of each subject) with pose changes between $[-20°, 20°]$ in tilt and $[-40°, 40°]$ in yaw. Each 3D shape pattern was estimated from a random selection of 20 of 45 face images of the same subject as stated in Section 2.1.2.

It is important to point out that the reason for using the small range of pose *in the training stage* is to make sure all landmarks are visible in the image. Otherwise, if some landmarks are invisible, it would be difficult to label the positions of those landmarks. However, this constraint is not imposed when fitting the model onto a novel image or sequence. It will be shown later that the model can be fitted successfully even when part of a face is invisible in a 2D image.

Figure 1 shows the projection, on $[-40°, 40°]$ in yaw (from left to right), of the first shape mode changing from the mean shape by $\{-3, 0, 3\}$ of standard deviation (from top to bottom). The first 10 eigenshapes take 95.5% of all variance.

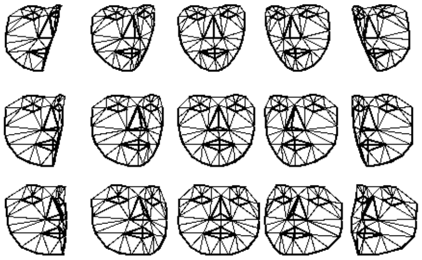

Figure 1. The first mode of the 3D PDM.

2.3. A Shape-and-Pose-Free Texture Model

There is no doubt that texture carries as important representative information as shape. However, accurately modelling face texture is nontrivial since it is quite sensitive to change of illumination, pose, and expression. In this work, we mainly focus on the problem of modelling facial texture variation arising from pose change. Explicitly modelling surface reflection and shading properties provides a solution to this problem. As an alternative, we present here a statistical approach to model face textures by extracting *shape-and-pose-free* texture information.

To decouple the covariance between shape and texture, a face image fitted by the shape model (Section 2.2) is warped to the mean shape at frontal view with $0°$ in both tilt and yaw. This is implemented by forming a triangulation from the landmarks and employing a piece-wise affine transformation between each triangle pair (see left in Figure 2). By warping to the mean shape, one obtains the shape-free texture of the given face image. Furthermore, by warping to the frontal view, a pose-free texture representation is achieved. Figure 2 illustrates the triangulation mesh of the mean shape in frontal view, a face image, the face fitted by the shape model, and the warped texture pattern to the mean shape in frontal view.

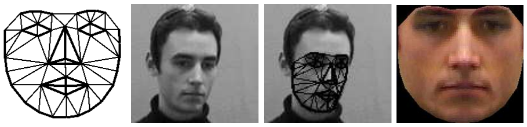

Figure 2. Extract the shape-and-pose-free texture of a face image.

We applied PCA to a set of 540 *shape-and-pose-free* face textures from 12 subjects with pose changes between $[-20°, 20°]$ in tilt and $[-40°, 40°]$ in yaw (45 from each subject). The first 12 eigen modes take 96.4% of all variance.

During the fitting process, a *shape-and-pose-free* texture pattern $\mathbf{q}$ of a face image, which is already warped to the mean shape in the frontal view, can be represented by

$$\mathbf{t} = \mathbf{V}^{\mathrm{T}}(\mathbf{q} - \bar{\mathbf{q}}) \tag{8}$$

where $\bar{\mathbf{q}}$ is the mean texture, and $\mathbf{V}$ is constructed by the first N_t significant eigen vectors of the texture PCA

$$\mathbf{V} = [\mathbf{v}_1 \mathbf{v}_2 ... \mathbf{v}_{N_s}] \tag{9}$$

The reconstruction of the texture pattern is

$$\mathbf{q}_r = \mathbf{V}\mathbf{t} + \bar{\mathbf{q}} \tag{10}$$

2.4. Representing Face Patterns

Based on the analysis above, a face pattern can be represented in the following way. First, a 3D shape model is fitted to the given image or video sequence containing faces. The shape parameters of the fitted face is given by Equation (6). The face texture is warped onto the mean shape of the 3D PDM model in the frontal view. Then the texture parameters of the face can be obtained using Equation (8). Finally, by adding parameters controlling pose, shift and scale, the complete parameter set of the dynamic model for a given face pattern is

$$\mathbf{c} = (\mathbf{s}, \mathbf{t}, \alpha, \beta, dx, dy, r)^{\mathrm{T}} \tag{11}$$

where (α, β) is pose in tilt and yaw, (dx, dy) is the translation of the centroid of the face, and r is its scale.

The parameter set consists of two parts: the identity information $(\mathbf{s}, \mathbf{t})$ which is crucial to face recognition and facial analysis, and the geometrical information $(\alpha, \beta, dx, dy, r)$ which is important for face alignment and tracking.

3. Model Fitting Algorithm

Model fitting in this context is to search for the optimal parameters of the model for an unknown face image to be interpreted. The parameters are given by:

$$\mathbf{c}^* = argmin(L(\mathbf{c})) \tag{12}$$

where $L(\mathbf{c})$ is a loss function which evaluates how well the model fits onto the image.

We formulate the loss function as

$$\begin{aligned} L(\mathbf{c}) \;=\; & \|\mathbf{q}_r(\mathbf{c}) - \mathbf{q}\| + \\ & \xi \sum_{i=1}^{N_l} w_i \mathcal{M}(\hat{\mathbf{F}}_i(\mathbf{c}), \mathbf{F}_{i0}) + \\ & \eta \sum_{i=1}^{N_l} w_i \mathcal{M}(\hat{\mathbf{F}}_i(\mathbf{c}), \hat{\mathbf{F}}_i(\mathbf{c}_{-1})) \end{aligned} \qquad (13)$$

The first term on the right-hand side evaluates the difference between the image appearance and the model synthesised appearance, where $\mathbf{q}_r(\mathbf{c})$ is the reconstructed texture given by (10), and $\mathbf{q}$ is the original texture warped onto the mean shape in frontal view. This is based on the principle of *analysis-by-synthesis* [7, 2, 17]. The better the model fits, the smaller the difference.

The second term, which is measured in Mahalanobis distance, describes the local texture similarity of each landmark to the template of this specific landmark estimated from training images, where $\hat{\mathbf{F}}_i(\mathbf{c})$ is the response of Gabor wavelet filters [11] or derivatives of Gaussian, on the current position of the ith landmark. The same filters have been applied to the training face images. A set of templates, one for each landmark, is obtained using PCA. $\mathbf{F}_{i0}$ denotes the template centroid. The Mahalanobis distance $\mathcal{M}(\hat{\mathbf{F}}_i(\mathbf{c}), \mathbf{F}_{i0})$ is calculated using distance-in-feature-space (DIFS) [14]. Each $\mathcal{M}(\hat{\mathbf{F}}_i(\mathbf{c}), \mathbf{F}_{i0})$ is weighted by w_i, which measures the visibility of the ith landmark. The value of w_i is computed from the normal of the landmark on the 3D shape. ξ is a normalisation coefficient, and N_l is the number of landmarks. It was noted in our experiments that the Gabor wavelet filter does not outperform simpler derivatives of Gaussian.

The last term, which is only enabled when the input is a video sequence, compares the difference between the filtered local texture around each landmark $\hat{\mathbf{F}}_i(\mathbf{c})$ and that in the previous frame $\hat{\mathbf{F}}_i(\mathbf{c}_{-1})$. The Mahalanobis distance $\mathcal{M}(\hat{\mathbf{F}}_i(\mathbf{c}), \hat{\mathbf{F}}_i(\mathbf{c}_{-1}))$ is also calculated using DIFS. η is a normalisation coefficient.

The loss function defined in (13) can be interpreted as follows: it is a weighted summation of the fitting criterion of the *global* appearance to the model synthesised appearance, the *local* fitting criterion around each landmark, and the *temporal* fitting criterion to the previous pattern.

Based on stochastic search, the fitting algorithm of the multi-view face model is implemented as in Table 1. The evaluation of the loss function used in step 4 is carried out as in Table 2. A Support Vector Machine based method was used for real-time pose estimation [12] in Step 1. Figure 3 illustrates the process of applying the above algorithm to a face image.

```
1  assume initial parameter $\mathbf{c}_0 = (\mathbf{s}, \alpha, \beta, r, dx, dy)$
2  randomly sample $n$ parameter points around initial $\mathbf{c}_0$
3  randomly sample $m$ parameter points around each of
   the $n$ points
4  evaluate the values of the loss function $L(\mathbf{c})$ for each
   of the $m \times n$ parameters
5  sort the loss function values in ascending order
6  if no improvement from the top value, stop
7  otherwise, save the first $n$ parameters, then go to 3
```

Table 1. Fitting algorithm

```
1  perform pose estimation using $(\mathbf{s}, r, dx, dy)$
2  restore 2D shape using $(\mathbf{s}, \alpha, \beta, r, dx, dy)$
     • reconstruct 3D shape $\mathbf{p}_r$ from $\mathbf{s}$ using (7)
     • project $\mathbf{p}_r$ to $(\alpha, \beta)$
     • scale to $r$ and translate to $(dx, dy)$
3  evaluate the global appearance fitting criterion given
   as the first term in (13)
     • warp the texture enclosed by the 2D shape to the
       mean shape in frontal view to obtain the shape-
       and-pose-free texture $\mathbf{q}$
     • compute the texture parameter $\mathbf{t}$ by projecting $\mathbf{q}$
       using (8)
     • reconstruct $\mathbf{q}_r$ using (10)
     • calculate the similarity
4  sample and filter the local texture around each land-
   mark
5  evaluate the local fitting criterion of landmarks given
   by the second term in (13)
6  evaluate the temporal fitting criterion of landmarks, if
   necessary, given by the third term in (13)
7  compute the overall loss in (13)
```

Table 2. Evaluation of $L(\mathbf{c})$

4. Fitting the Model to Sequences

By fitting the multi-view face model to face images, one extracts and separates the identity parameters and geometrical parameters from the raw images. A solution to this problem can be greatly improved when a continuous video input is available. From video sequences, not only can different views and various textures be used for model fitting, but also the temporal continuity provides the possibility to exploit the facial dynamics encoded in the input stream.

4.1. Temporal Estimation of Model Parameters

We assume an input sequence contains only one subject whose identity is unchanged throughout the sequence. Fitting the model onto a sequence frame by frame independently may receive a fluctuant estimation of the model parameters since there is no identity constancy constraint imposed on the fitting process. Instead, in each frame, it only tries to minimise the loss function given in (13). Other reasons for the fluctuation include local optima and image noise. When face recognition or facial analysis is performed under continuous video stream input, the model fitting problem should be regarded as dynamic parameter estimation of an underlying stochastical process where the identity parameters $(\mathbf{s}, \mathbf{t})$ are kept constant and the geometrical param-

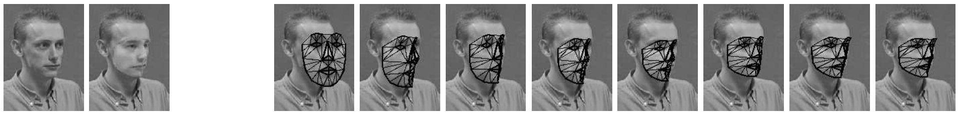

Figure 3. Fit the multi-view face model to a face image. The first two images shows the original face image and the fitted pattern warped on the original image. The others are the fitting results in 8 iterations.

eters change freely. In this paper, we only discuss the issue of temporal estimation of the identity parameters because of its importance to face recognition and facial analysis.

A straightforward method to estimate the identity parameters temporally is performing Gaussian estimation [1] which is based on the least squares principle. However, this method computes all the information accumulated in a batch way which is not appropriate for tracking. Alternatively, Kalman filters [1] provide a recursive solution to this problem.

The problem of estimating the identity parameters of the model using a Kalman filter can be formulated as follows. For a shape vector, the state transition equation is

$$\mathbf{s}(k) = \mathbf{s}(k-1) \tag{14}$$

The observation is taken from the 2D projection of the 3D shape since this is the only visible part of the 3D shape.

$$\mathbf{o}'(k) = \mathbf{R_t}(k)(\mathbf{U}\mathbf{s}(k) + \bar{\mathbf{p}}) + \mathbf{w}(k) \tag{15}$$

where $\mathbf{w}(k)$ denotes a zero-mean, white observation noise, and $\mathbf{R_t}(k)$ is the rotation and projection matrix extended by $\mathbf{R}$ in (2),

$$\mathbf{R_t}(k) = \begin{bmatrix} \mathbf{R} & 0 & ... & 0 \\ 0 & \mathbf{R} & ... & 0 \\ & & ... & \\ 0 & 0 & ... & \mathbf{R} \end{bmatrix} \tag{16}$$

Defining

$$\mathbf{H}(k) = \mathbf{R_t}(k)\mathbf{U} \tag{17}$$

$$\mathbf{o}(k) = \mathbf{o}'(k) - \mathbf{R_t}(k)\bar{\mathbf{p}} \tag{18}$$

the observation equation is then given by

$$\mathbf{o}(k) = \mathbf{H}(k)\mathbf{s}(k) + \mathbf{w}(k) \tag{19}$$

Therefore, temporal estimation of the model identity parameters can be performed by a Kalman filter:

$$\hat{\mathbf{s}}(k) = \hat{\mathbf{s}}(k-1) + \mathbf{K}(k)[\mathbf{o}(k) - \mathbf{H}(k)\hat{\mathbf{s}}(k-1)] \tag{20}$$

$$\mathbf{P}(k) = \mathbf{P}(k-1) - \mathbf{K}(k)\mathbf{H}(k)\mathbf{P}(k-1) \tag{21}$$

$$\mathbf{K}(k) = \mathbf{P}(k-1)\mathbf{H}^{\mathrm{T}}(k)[\mathbf{H}(k)\mathbf{P}(k-1)\mathbf{H}^{\mathrm{T}}(k) + \mathbf{Q}]^{-1} \tag{22}$$

where $\mathbf{K}$ is Kalman gain, $\mathbf{P}$ is the error covariance matrix, and $\mathbf{Q}$ is the covariance matrix of $\mathbf{w}(k)$ which can be estimated from the training data,.

A Kalman filter can also be designed for the texture vector in a similar way. However, unlike the one for the shape vector, where the observation vector is formulated from the 2D projection of the 3D shape, the state vector, i.e. the texture parameter $\mathbf{t}$, is fully observable, thus the observation vector and the state vector can be identical.

4.2. Tracking Out-of-Range Poses

As stated in Section 2.2, the 3D PDM shape model is trained from 2D images with limited pose range where all landmarks are visible. To verify if the model generalises well on out-of-range poses, we applied the model on sequences where faces are undergoing large pose change. The pose range in those sequences are normally profile to profile.

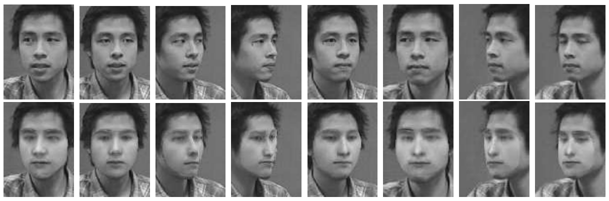

Figure 4. Tracking faces undergoing large pose change. The first row is original images from sample frames, and the second row shows the reconstructed face patterns overlapped on the original images.

The results depict that the model is capable of coping with large variation of pose even though it is trained on a limited range of views. This can be explained for two reasons. First, the shape information is represented in 3D, so the model can be rotated and projected to 2D for any given pose. Second, in the loss function (13), the local and temporal criteria are defined in a pose-specific way since they are weighted by a visibility measure which depends on pose. In all the experiments, the model has demonstrated a reliable performance between $[-70°, 70°]$ in yaw. However, when the pose is nearly $\pm 90°$, tracking may fail since little information is available in this view.

4.3. Tracking Faces with Expression Changes

To verify the robustness of the model, we also fitted it on sequences containing faces undergoing significant expression changes. The results from one of those sequences is shown in Figure 5. It is noted that the fitting is less well in some frames due to significant expression change. The main reason is that all the face images used for training are taken in neutral expression. However, due to the averaging and smoothing effect of Kalman filter, the fitting process still converged to a stable estimation of the subject identity and shown to be very robust over time despite errors in individual frames.

It is important to point out that the aim of this experiment is to estimate the identity parameters and is not to recognise the expressions of the subject. In other words, a state-invariant model defined by (14) is used on the basis of subject constancy.

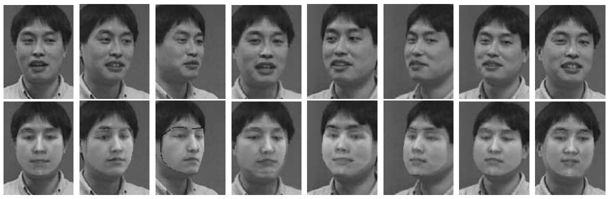

Figure 5. Tracking faces with significant expression change.

5. Conclusions

In this work, we focus on two important issues of face recognition and facial analysis, modelling face appearance with large pose variation and modelling faces dynamically over time. To address the problems, we present an integrated multi-view dynamic face model which includes a sparse 3D PDM shape model, a *shape-and-pose-free* texture model and an affine geometrical model. The contributions of this work are summarised as follows:

1. A 3D PDM shape model is learned from 2D images labelled with poses and landmarks. Instead of using dense 3D range data, this model consists of a sparse set of landmarks only.
2. A *shape-and-pose-free* texture model is built to decouple the covariance between shape and texture.
3. Although only face images from limited pose range are used in the training stage to ensure all landmarks are visible in the images, this limitation of pose range is never imposed when applying the model for tracking. Experimental results indicate that it is able to cope with pose variation from profile to profile.
4. By applying the model, two sets of information, the identity parameters and geometrical parameters, are obtained. The former is crucial to face recognition and facial analysis, and the latter is important for alignment and tracking.
5. Fitting criteria are formulated from the global fitting criterion of the entire face, the local fitting criterion of the landmarks and the temporal fitting criterion to previous patterns.
6. Temporal estimation of model parameters is employed to provide a more robust and stable fit over time.

References

[1] K. Brammer and G. Siffling. *Kalman-Bucy Filters*. Artech House, Norwood, USA, 1989.

[2] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. In *European Conference on Computer Vision*, volume 2, pages 484–498, Freiburg, Germany, 1998.

[3] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models - their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[4] T. Cootes, K. Walker, and C. Taylor. View-based active appearance models. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 227–232, Grenoble, France, 2000.

[5] D. DeCarlo and D. Metaxas. Deformable model-based face shape and motion estimation. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 146–150, Vermont, US, 1996.

[6] G. Edwards, C. Taylor, and T. Cootes. Interpreting face images using active appearance models. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 300–305, Nara, Japan, 1998.

[7] T. Ezzat and T. Poggio. Facial analysis and synthesis using image-based methods. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 116–121, Vermont, US, 1996.

[8] S. Gong, S. McKenna, and A. Psarrou. *Dynamic Vision: From Images to Face Recognition*. World Scientific Publishing and Imperial College Press, April 2000.

[9] S. Gong, A. Psarrou, I. Katsouli, and P. Palavouzis. Tracking and recognition of face sequences. In *European Workshop on Combined Real and Synthetic Image Processing for Broadcast and Video Production*, pages 96–112, Hamburg, Germany, 1994.

[10] T. Jebara and A. Pentland. Parametrized structure from motion for 3d adaptive feedback tracking of faces. In *IEEE Conference on Computer Vision and Patter Recognition*, 1997.

[11] M. Lades, J. Vorbruggen, J. Buhmann, J. Lange, C. Malsburg, R. Wurtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42(3):300–311, 1993.

[12] Y. Li, S. Gong, and H. Liddell. Support vector regression and classification based multi-view face detection and recognition. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 300–305, Grenoble, France, March 2000.

[13] B. Moghaddam and A. Pentland. Face recognition using view-based and modular eigenspaces. In *Automatic Systems for the Identification and Inspection of Humans, SPIE*, volume 2277, 1994.

[14] B. Moghaddam and A. Pentland. Probalilistic visual learning for object representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):137–143, 1997.

[15] S. Romdhani, S. Gong, and A. Psarrou. A multi-view nonlinear active shape model using kernel pca. In *British Machine Vision Conference*, pages 483–492, Nottingham, UK, 1999.

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

[17] T. Vetter and V. Blanz. Generalization to novel views from a single face image. In Wechsler, Philips, Bruce, Fogelman-Soulie, and Huang, editors, *Face Recognition: From Theory to Applications*, pages 310–326. Springer-Verlag, 1998.

[18] T. Vetter and T. Poggio. Linear object classes and image synthesis from a single example image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):733–742, 1997.

[19] O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 318–323, Nara, Japan, 1998.