International Journal of Computer Vision 53(1), 71–92, 2003

Constructing Facial Identity Surfaces for Recognition

YONGMIN LI
Content and Coding Lab, BTexact Technologies, pp2 Ross Building, Adastral Park, Ipswich IP5 3RE, UK
Yongmin.Li@bt.com

SHAOGANG GONG AND HEATHER LIDDELL
Department of Computer Science, Queen Mary, University of London, London E1 4NS, UK
sgg@dcs.qmul.ac.uk
heather@dcs.qmul.ac.uk

Received March 1, 2001; Revised December 17, 2002; Accepted December 18, 2002

Abstract. We present a novel approach to face recognition by constructing facial identity structures across views and over time, referred to as identity surfaces, in a Kernel Discriminant Analysis (KDA) feature space. This approach is aimed at addressing three challenging problems in face recognition: modelling faces across multiple views, extracting non-linear discriminatory features, and recognising faces over time. First, a multi-view face model is designed which can be automatically fitted to face images and sequences to extract the normalised facial texture patterns. This model is capable of dealing with faces with large pose variation. Second, KDA is developed to compute the most significant non-linear basis vectors with the intention of maximising the between-class variance and minimising the within-class variance. We applied KDA to the problem of multi-view face recognition, and a significant improvement has been achieved in reliability and accuracy. Third, identity surfaces are constructed in a pose-parameterised discriminatory feature space. Dynamic face recognition is then performed by matching the object trajectory computed from a video input and model trajectories constructed on the identity surfaces. These two types of trajectories encode the spatio-temporal dynamics of moving faces.

Keywords: face recognition, Kernel Discriminant Analysis, dynamic face models, facial identity surfaces

1. Introduction

Face recognition, an important visual perceptual task, has been of great interest in recent years both theoretically and practically. Applications of face recognition cover various areas including integrated surveillance, visually mediated interaction, human-machine interface, multimedia and teleconferencing. Various approaches have been proposed to address the problem under different assumptions and conditions. In the rest of this section, we first review the previous studies in this area in Section 1.1, then discuss the limitations of the previous work in Section 1.2 and introduce our approach to dynamic face recognition using identity surfaces in Section 1.3.

1.1. Previous Work

Shape based methods for face recognition have been developed in numerical studies. The Active Contour Model (Snake) (Kass et al., 1987) is one of the most popular models adopted in this area (Waite and Welsh, 1990; Wu et al., 1996; Okubo and Watanabe, 1998; Yokoyama et al., 1998). However, the Active Contour Model imposes *soft* rather than *problem-specific* constraints to the favoured shapes. By using a parametric shape vector with few degrees of freedom, one can achieve the so-called "Deformable Model", where *hard* constraints and default shapes of more specific classes of shapes are explicitly defined. Examples of using Deformable Models for face recognition include (Yuille et al., 1992; Craw et al., 1992; Bennett and Craw, 1991; Brunelli and Poggio, 1993).

Cootes et al. (1995) developed the Active Shape Models (ASMs) where the characteristics of a class of objects are learnt from a training set of correctly annotated images. New shapes of the class can be represented by a weighted sum of a small number of significant basis shape vectors. These models can be used for image search through an iterative refinement algorithm analogous to that employed by Active Contour Models. The key difference is that ASMs can only deform to fit the data in ways consistent with the training set (Cootes et al., 1994, 1995). The ASMs have been successfully applied to face recognition (Lanitis et al., 1997). Compared with the classical geometric deformable shape model, the significant advantage of ASMs lies in the fact that the class-specific shape constraints are learnt directly from training examples rather than handcrafted by geometric shapes.

The ASMs were originally developed as linear models. However, non-linear extensions have been proposed, for example, Edwards et al. (1998a) have described a non-linear ASM built from a Multi-Layer Perceptron, and Romdhani et al. (1999) introduced a non-linear shape model to address the problem of corresponding faces with large pose variation using Kernel Principal Component Analysis (Scholkopf, 1997; Scholkopf et al., 1997).

In contrast to shape, modelling faces by their 2D appearance seeks to capture the holistic characteristics of faces rather than a set of individual features. One of the straightforward approaches is to represent faces by a set of generic templates, which can be either natural face images or synthesised images. Brunelli and Poggio (1993) compared a geometrical-feature-based algorithm with a template-based algorithm. They claimed that the results obtained for the testing sets show about 90% correct recognition using geometric features and perfect recognition using template matching. Sung and Poggio (1994) generated 6 face prototypes and 6 near-face-nonface prototypes as templates to match a new image pattern. A well-tuned Neural Network is employed to synthesise these matching results. Another approach using Support Vector Machines is presented by Osuna and Poggio, where the most representative examples, known as Support Vectors, are extracted automatically (Osuna et al., 1997a, 1997b).

Statistical techniques like Principal Component Analysis (PCA) (Bishop, 1995; Sirovich and Kirby, 1987), Linear Discriminant Analysis (LDA) (Fisher, 1938; Fukunaga, 1972) and Kernel Principal Component Analysis (KPCA) (Scholkopf, 1997; Scholkopf et al., 1997) have been widely adopted to extract the abstract features of face patterns. Sirovich and Kirby (1987) first introduced PCA, also known as the Karhunen-Loeve transform, to face recognition. This approach is commonly referred to as the "eigenface" method. Turk and Pentland (1989) used a similar method to code face images and capture face features. Moghaddam and Pentland (1997) extended this approach to view-based and modular eigenspaces with an intention of recognising faces under varying views and locating facial features, such as eyes and mouth. They also modelled two mutually exclusive classes of variation between facial images: intra-personal and extra-personal variations using eigenspace density estimation (Moghaddam et al., 1998). Romdhani et al. (1999) have applied KPCA to model the non-linearity of facial appearance caused from large pose change.

LDA proved to be an appropriate linear technique to extract the most discriminatory features. Swets and Weng (1996) applied LDA to face recognition and compared the performance with that of PCA. Their experiments showed an improved performance with the LDA method when a large number of training images of each face class are available. To avoid the singular problem which may occur when eigen-decomposing the scatter matrix in LDA, a PCA is usually performed prior to LDA to provide an adequately small dimensionality for further analysis (Swets and Weng, 1996; Zhao et al., 1998). Edwards et al. (1996) adopted LDA to select discriminant parameters based on Active Appearance Models. They claimed that these parameters can be used effectively to decouple identity variance from pose, lighting and expression variance.

Shape and texture are often employed complementarily to represent faces. Vetter and Poggio (Vetter, 1998) proposed to synthesise faces in different views using linear combination of shape and texture prototype patterns. Lanitis et al. (1997) developed a face identification system using both facial shape and texture information where the shape information is obtained through an ASM and texture patterns are sampled around landmarks or computed as shape-free grey-level. Cootes et al. (1998) introduced the Active Appearance Model which combines both the shape and grey-level variation within a single statistical model.

Faces can also be modelled by a 3D mesh describing the geometric configuration and a texture map expressing the surface properties. DeCarlo and Metaxas (2000) presented a 3D deformable face model with a polygon mesh. The model is formed from ten component parts, each with its own set of deformation. Jebara and Pentland (1997) 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 generic face model which is formed offline from a database of range face data. Vetter and Blanz (1998) introduced a flexible 3D face model learnt from examples of individual 3D face data. 3D face models have also been used for person-independent face tracking and feature detection (Li et al., 1993; Shakunaga et al., 1998).

Apart from the research on static images, video based face recognition has attracted great interest recently. Gong et al. (1994) have addressed the issue of encoding and recognising moving faces using temporal signatures in a multi-view eigenspace. Howell and Buxton (1996) reported a preliminary system for face recognition from image sequences based on Radial Basis Function networks. McKenna and Gong (1998) described an integrated system for recognising moving faces in poorly constrained dynamic scenes. Steffens et al. (1998) presented a real-time face recognition system which is able to capture, track and recognise a person walking toward or passing a pair of stereo cameras. Choudhury et al. (1999) proposed a person identification system to recognise and verify people from unconstrained video and audio. Edwards et al. (1998b, 1999) proposed an approach to learning the class-specific correction of identity parameters from image sequences.

1.2. Limitations of Previous Work

It is important to point out that most of the previous work in face recognition is mainly concerned with frontal-view. Recognising faces across views is more challenging than that at a fixed view, e.g. frontal view, because of the severe non-linearity caused by rotation in depth, self-occlusion, self-shading and illumination change. The eigenface method has been extended to view-based and modular eigenspaces with the intention of recognising faces under varying views by Moghaddam and Pentland (1994). Li et al. (2000b) presented a view-based piece-wise SVM model of the face space. Cootes et al. (2000) proposed the view-based Active Appearance Models which employ three models for profile, half-profile and frontal views. However the division of the face space in these methods is rather arbitrary, ad hoc and often coarse.

Another limitation of the previous work is that the methods proposed for recognition are largely based on matching static face images. Psychology and physiology research depicts that the human vision system's ability to recognise animated faces is better than that on randomly ordered still face images (i.e. the same set of images, but displayed in random order without the temporal context of moving faces). Knight and Johnston (1997) showed that recognition of famous faces in photographic negatives can be significantly enhanced when the faces were shown moving rather than static. Bruce et al. (1998a, 1998b) extended this result to other conditions where recognition is made difficult, e.g. by thresholding the images or showing them in blurred or pixellated formats. Although some preliminary results obtained from techniques such as the temporal signature method (Gong et al., 1994), the subspace method (Yamaguchi et al., 1998) and the identity trajectory method (Li et al., 2000a), have been reported, the issue of recognising the dynamics of faces under a spatio-temporal context remains largely unresolved.

1.3. Our Approach

In this paper, we present a novel and comprehensive approach to modelling facial identities across views and over time. To model faces with large pose variation, a multi-view face model with 3D shape, normalised facial texture and affine geometrical information is developed. With a stochastic fitting algorithm, this model can be automatically fitted to face images or sequences to obtain the shape, texture and geometry descriptions of faces. To address the severe non-linearity of multi-view faces, a non-linear discriminatory method, Kernel Discriminant Analysis (KDA), is adopted, which employs the kernel technique to maximise the between-class variance and minimise the within-class variance. Aimed at dynamic face recognition, a spatio-temporal identity surface of each face class is constructed in a kernel discriminatory feature space. A video-based approach using pattern distances and trajectory distances to the identity surfaces is presented to perform online face recognition dynamically. The rest of this paper is arranged as follows: The multi-view face model is presented in Section 2, including the model components, model training and fitting algorithm. The issue of extracting non-linear discriminatory features from multi-view faces using KDA is discussed in Section 3. Section 4 describes an approach to video-based face recognition using identity surfaces. 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) (Cootes et al., 1995) 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 Labelled 2D Images

Modelling the appearance of faces with large pose variation is non-trivial for 2D models owing to the severe non-linearity. But if 3D geometrical information is available, this situation can be alleviated to a certain extent. A straight-forward way to collect 3D information about faces is to use sensors such as a 3D laser scanner. However, the huge amount of 3D range data involved may bring a heavy burden to the computation. Another difficulty comes from establishing the correspondence between 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 at different views.

Our database includes 2D face images from 12 subjects, 133 poses of each subject. All face images were chosen without spectacles on (see Gong et al. (2000)) 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 the image plane is not taken into account on the basis that human heads are assumed to be mostly upright. A sparse set of 44 landmarks locating the mouth, nose, eyes, and face contour were semi-automatically labelled on each face image. Figure 1 illustrates the landmarks used in this work and the triangulation formed from these landmarks which can be used to warp multi-view faces onto the frontal view (details will be discussed in Section 2.3). Figure 2 shows the sample face images used to construct the model (a) and the landmarks labelled on each image (b).

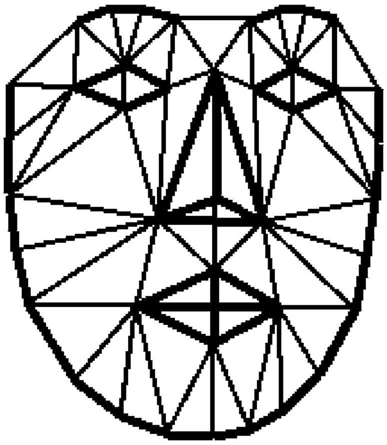


Figure 1. Landmarks and triangulation of the face model.

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 at the centre of the eye centres and the mouth centre. We define 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:

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

where $\mathbf{R}(\alpha, \beta)$ 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.

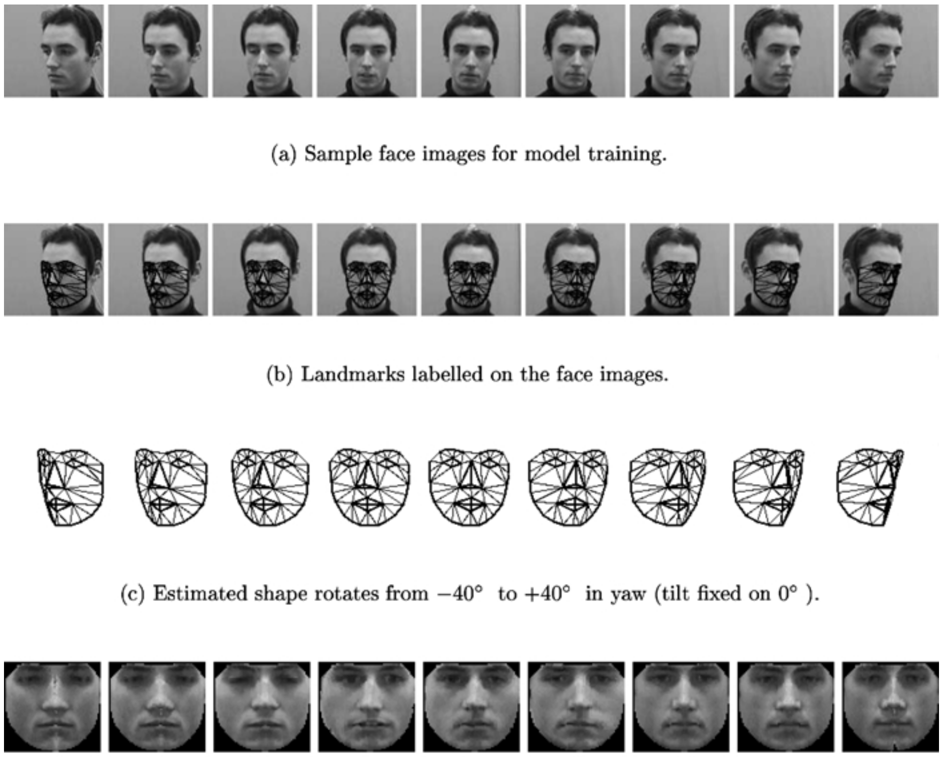
$$\mathbf{R}(\alpha, \beta) = \begin{bmatrix} \cos(\beta) & 0 & \sin(\beta) \\ \sin(\alpha)\sin(\beta) & \cos(\alpha) & -\sin(\alpha)\cos(\beta) \end{bmatrix} \tag{2}$$

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

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

$$\text{Subject to} \quad (x_i, y_i)^{\mathrm{T}} = \mathbf{R}(\alpha_i, \beta_i)\,(X, Y, Z)^{\mathrm{T}}, \quad i = 1, 2, \ldots, M \tag{4}$$



(a) Sample face images for model training.

(b) Landmarks labelled on the face images.

(c) Estimated shape rotates from $-40°$ to $+40°$ in yaw (tilt fixed on $0°$).

(d) Normalised facial texture patterns by warping to the mean shape in frontal view.

Figure 2. Multi-view face model.

where (x_i, y_i) is the known 2D position of the landmark and (α_i, β_i) is the pose of the landmark in the ith face image. Then the 3D shape vector **p** is obtained as:

$$\mathbf{p} = \left(X_1, Y_1, Z_1, X_2, Y_2, Z_2, \ldots, X_{N_l}, Y_{N_l}, Z_{N_l}\right)^{\mathrm{T}} \quad (5)$$

where N_l is the number of landmarks.

Figure 2(c) shows a 3D shape pattern with tilt fixed on $0°$ and yaw changing from $-40°$ to $+40°$. This shape pattern is estimated from the labelled face images in Fig. 2(b).

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 adequately large. 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 in order to learn the statistical 3D PDM of faces. This will be further discussed in Section 2.2.

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. The shapes of human faces are able to be represented in an even lower dimensional shape space since they share a very similar structure. The PDM is adopted to construct this low dimensional shape space.

Performing PCA on N given 3D face shape vectors $\{\mathbf{p}_i, i = 1, 2, \ldots, N\}$, which are estimated using the method described in Section 2.1, one obtains the mean shape $\bar{\mathbf{p}}$ and the matrix $\mathbf{U}$ which is comprised of the first N_s significant eigenvectors

$$\mathbf{U} = \left[\mathbf{u}_1\mathbf{u}_2 \ldots \mathbf{u}_{N_s}\right] \tag{6}$$

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

$$\mathbf{s} = \mathbf{U}^{\mathrm{T}}(\mathbf{p} - \bar{\mathbf{p}}) \tag{7}$$

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

$$\mathbf{p}_r = \mathbf{U}\mathbf{s} + \bar{\mathbf{p}} \tag{8}$$

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. The first 10 eigenshapes account for 95.5% of all variance.

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 locate the positions of these landmarks. However, this constraint is not imposed when fitting the model onto a novel image or sequence. It will be shown in Section 5 that the model can be fitted successfully even in the profile view where nearly half of a face is invisible in a 2D image.

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

There is no doubt that texture carries as important information as shape. However, accurately modelling facial texture is non-trivial owing to its sensitivity to changes in illumination, pose, and expression. In this work, we focus mainly on the problem of modelling facial texture variation arising from pose change. Explicitly modelling surface reflection and shading properties provides one possible solution to this problem (Atick et al., 1996; Zhao and Chellappa, 2000). 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 facial texture and shape, the facial texture 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 of the triangle pairs (see Fig. 1). By warping to the mean shape, one obtains the shape-free texture of a given face image. Furthermore, by warping to the frontal view, a pose-free texture representation is achieved.

The aim of warping the facial texture to the shape-and-pose-free patterns is to establish the correspondence for faces in different shapes and across multiple views, so that an accurate statistical model can be constructed from these normalised texture patterns. The so-called "pose-free" in this context is in the sense of geometry only, i.e. facial texture is normalised by geometrical parameters (landmark positions and pose angles) only. The variation from other sources which may also be related to pose, for example, the self-shading and illumination change while a face is rotating out of the image plane, has not been studied at the current stage.

Figure 2(d) illustrates the shape-and-pose-free texture patterns of the face images shown in Fig. 2(a). It is noted that when one side of a face becomes partially invisible, the texture pattern is constructed from the visible side using the bilateral symmetry of faces.

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 eigenmodes account for 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

$$\mathbf{t} = \mathbf{V}^{\mathrm{T}}(\mathbf{q} - \bar{\mathbf{q}}) \tag{9}$$

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

$$\mathbf{V} = \left[\mathbf{v}_1\mathbf{v}_2 \ldots \mathbf{v}_{N_s}\right] \tag{10}$$

The reconstruction of the texture pattern is given by

$$\mathbf{q}_r = \mathbf{V}\mathbf{t} + \bar{\mathbf{q}} \tag{11}$$

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 a given image or video sequence containing faces. The shape parameters of the fitted face are given by Eq. (7). The face texture is warped onto the mean shape of the 3D PDM model at the frontal view. Then the texture parameters of the face are computed using Eq. (9). Finally, by adding parameters controlling pose, shift and scale, the complete parameter set of the dynamic model for a given face pattern is

$$\mathbf{c} = (\mathbf{s}, \mathbf{t}, \alpha, \beta, dx, dy, r) \tag{12}$$

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.

2.5. Model Fitting

Model fitting in this context is the problem of searching for the optimal parameters of the model for an unknown face image to be interpreted, and it is given by:

$$\mathbf{c}^* = \operatorname{argmin}(L(\mathbf{c})) \tag{13}$$

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

2.5.1. Loss Function for Fitting. We formulate the loss function as

$$L(\mathbf{c}) = \|\mathbf{q}_r(\mathbf{c}) - \mathbf{q}\| + \xi \sum_{i=1}^{N_l} w_i \mathcal{M}(\hat{\mathbf{F}}_i(\mathbf{c}), \mathbf{F}_{i0}) + \eta \sum_{i=1}^{N_l} w_i \mathcal{M}(\hat{\mathbf{F}}_i(\mathbf{c}), \hat{\mathbf{F}}_i(\mathbf{c}_{-1})) \tag{14}$$

The first term on the right-hand side evaluates the difference between the image appearance and the synthesised appearance, where $\mathbf{q}_r(\mathbf{c})$ is the reconstructed texture given by (11), and $\mathbf{q}$ is the original texture warped onto the mean shape at frontal view. This is based on the principle of *analysis-by-synthesis* (Ezzat and Poggio, 1996; Cootes et al., 1998; Vetter and Blanz, 1998). 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 (Lades et al., 1993) or derivative of Gaussian, on the current position of the ith landmark. The same filters have been applied to the training face images described in Section 2.3. 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 the notion of distance-in-feature-space (DIFS) (Moghaddam and Pentland, 1997). 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 the simpler derivative 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 (14) 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 pattern in the previous frame.

2.5.2. A Fitting Algorithm. Based on stochastic search, the fitting algorithm of the multi-view face model is described in Table 1. The evaluation of the loss function used in step 4 is carried out as described in Table 2.

The Support Vector Machine based method described in Li et al. (2000b) was used for real-time pose estimation in Step 1. Figure 3 illustrates the process of applying the above algorithm to a face image.

2.6. Fitting the Model to Sequences Over Time

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 more information across views and over time be used for model fitting, but also, the temporal continuity provides the possibility

Table 1. Fitting algorithm.

1. Assume initial parameter $\mathbf{c}_0 = (\mathbf{s}, \alpha, \beta, r, dx, dy)$;
2. Randomly sample n parameter points around the 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 2. Evaluation of $L(\mathbf{c})$.

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 (8);
   - project $\mathbf{p}_r$ to (α, β);
   - scale to r and translate to (dx, dy);
3. Evaluate the global fitting criterion given as the first term in (14)
   - warp the texture enclosed by the 2D shape to the mean shape at frontal view to obtain the shape-and-pose-free texture $\mathbf{q}$;
   - compute the texture parameter $\mathbf{t}$ by projecting $\mathbf{q}$ using (9);
   - reconstruct $\mathbf{q}_r$ using (11);
   - calculate the similarity;
4. Sample and filter the local texture around each landmark;
5. Evaluate the local fitting criterion of landmarks given by the second term in (14);
6. Evaluate the temporal fitting criterion of landmarks, if necessary, given by the third term in (14);
7. Compute the overall loss in (14).

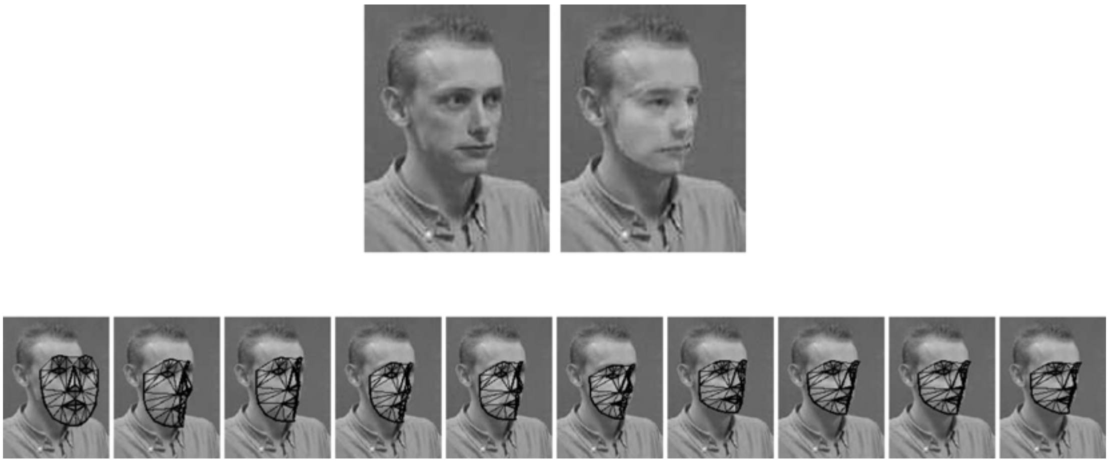


Figure 3. Fit the multi-view face model to a face image. The first row shows the original face image and the reconstructed texture of the fitted pattern warped on the original image. The second row lists the fitting results in 10 iterations.

to exploit the facial dynamics encoded in the input stream.

Suppose an input sequence contains one subject whose identity is unchanged throughout the sequence. Fitting the model onto a sequence frame by frame independently is likely to yield a fluctuating estimate of the model parameters for the following reasons:

1. 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 (14).
2. The fitting algorithm may be attracted to local optima and image noise.
3. Expression and illumination changes may also affect the estimation of model parameters.

Under these circumstances, the model fitting problem should be regarded as dynamic parameter estimation of an underlying stochastic process where the identity parameters (**s**, **t**) are kept constant and the geometrical parameters change freely. In the following discussion, we assume that the purpose of model fitting is face recognition, i.e. temporally estimating the identity parameters of faces.

A straightforward approach to estimate the identity parameters temporally is performing Gaussian estimation (Brammer and Siffling, 1989) based on the least squares principle. However, this method computes all the information accumulated in a batch way which is not appropriate for dynamic model fitting. Alternatively, a temporal model such as Kalman filters (Brammer and Siffling, 1989) provides a recursive solution to this problem.

3. Extracting Non-Linear Discriminatory Features of Multi-View Faces

Owing to the severe non-linearity caused by rotation in depth, self-occlusion, self-shading and illumination change, modelling the appearance of faces across multiple views is much more challenging than that from a fixed view, e.g. frontal view. Moreover, the appearance similarity between different persons from a same view is not less than that of one person from different views, i.e. the variation from different face identities may be overshadowed by that from pose change. This makes the task of recognising identities of multi-view faces even more challenging. However, it is not a difficult task for the human vision system to recognise faces across views, which suggests that our biological vision system may perform this task with a discriminatory mechanism rather than based on the low-level appearance features.

3.1. Statistical Methods of Feature Extraction

The high dimensionality of raw images is problematic in computation. To address this, PCA (Bishop, 1995; Sirovich and Kirby, 1987) has been adopted to reduce dimensionality and extract abstract features of faces. We have reviewed the previous work of using PCA for face recognition in Section 1.1. However, it is worth noting that the features extracted by PCA are actually "global" features for all face classes, thus they are not necessarily representative for discriminating one face class from others.

On the other hand, LDA, which seeks to find a linear transformation by maximising the between-class variance and minimising the within-class variance (Fisher, 1938; Fukunaga, 1972), proved to be a more suitable technique for class separation. Computationally, LDA can be solved as an eigen-decomposition problem similar to PCA. Although LDA can provide a significant discriminatory improvement to the task of face recognition, it is still a linear technique in nature. When severe non-linearity is involved, this method is intrinsically poor. Another shortcoming of LDA lies in the fact that the number of basis vectors are limited by the number of face classes, therefore it would be less representative when small set of subjects are concerned.

To extract the non-linear principal components, Kernel PCA (KPCA) was developed for pattern recognition (Scholkopf et al., 1997; Scholkopf, 1997). However, as with PCA, KPCA captures the *overall* variance of all patterns which are inadequate for discriminatory purposes.

In this work, we adopt Kernel Discriminant Analysis (KDA) (Roth and Steinhage, 1999; Mika et al., 1999; Baudat and Anouar, 2000) to extract the nonlinear discriminatory features for face recognition across multiple views.

3.2. Kernel Discriminant Analysis

The principle of KDA is illustrated in Fig. 4. It is difficult to directly compute the discriminatory features between the two classes of patterns because of the severe non-linearity. By defining a non-linear mapping from the input space to a high-dimensional feature space, patterns are linearly separable in the feature space. Then LDA, the linear technique, can be performed in the feature space to extract the most significant discriminatory features. However, the computation may be problematic or even impossible in the feature space owing to the high dimension. By introducing a kernel function which corresponds to the nonlinear mapping, all the computation can conveniently be carried out in the input space.

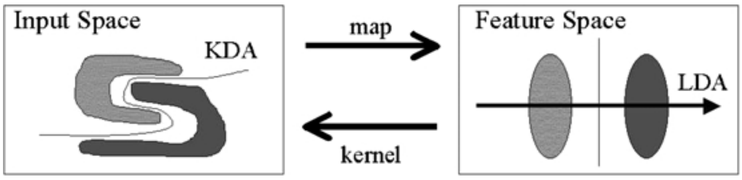


Figure 4. Kernel Discriminant Analysis. The non-linear discriminating problem can be solved as a linear problem by projecting the patterns onto a high-dimensional feature space. Furthermore, all computation can be performed conveniently in the original space through a kernel function.

The algorithm of training a KDA is briefly described in the following. Please refer to Li et al. (2001) for more details of the algorithm derivation. Alternative algorithms can be found in Roth and Steinhage (1999), Mika et al. (1999), and Baudat and Anouar (2000).[1]

Suppose we have a set of training patterns $\{\boldsymbol{x}_i, i = 1, 2, \ldots, N\}$ which are categorised into C classes. Let N_c be the number of patterns in class c, and N the total number of patterns, i.e. $N = \sum_{c=1}^{C} N_c$. We assume that patterns in the training set have been ordered by class number, i.e. the first N_1 patters belong to class 1, the following N_2 class 2, and so on. ϕ is defined as a non-linear mapping function from the input space to a high-dimensional feature space, and its corresponding kernel function is

$$k(\boldsymbol{x}, \boldsymbol{y}) = (\phi(\mathbf{x}) \cdot \phi(\mathbf{y})) \tag{15}$$

which fulfils the Mercer's condition (Vapnik, 1995).

Construct an $N \times N$ matrix

$$(\boldsymbol{K})_{ij} := k(\phi_i \cdot \phi_j) \tag{16}$$

We then obtain the centred kernel matrix

$$\tilde{\boldsymbol{K}} = \boldsymbol{K} - \frac{1}{N}\mathbf{1}_N \boldsymbol{K} - \boldsymbol{K}\frac{1}{N}\mathbf{1}_N + \frac{1}{N^2}\mathbf{1}_N \boldsymbol{K} \mathbf{1}_N \tag{17}$$

where $\mathbf{1}_N$ is an $N \times N$ matrix with $(\mathbf{1}_{N_c})_{ij} := 1$. Note that the elements in $\tilde{\boldsymbol{K}}$ follow the class order, i.e.

$$\tilde{\boldsymbol{K}} = [\boldsymbol{K}_1, \boldsymbol{K}_2, \ldots, \boldsymbol{K}_C] \tag{18}$$

where the $N \times N_c$ matrix $\boldsymbol{K}_c, c = 1, 2, \ldots, C$ is a sub-matrix of $\tilde{\boldsymbol{K}}$.

The problem can be finally formulated as an eigen-decomposition problem

$$\boldsymbol{A}\boldsymbol{\alpha} = \lambda\boldsymbol{\alpha} \tag{19}$$

yielding an eigenmatrix

$$\boldsymbol{U} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_M] \tag{20}$$

constructed from the first M significant eigenvectors of $\boldsymbol{A}$. The $N \times N$ matrix $\boldsymbol{A}$ is defined as

$$\boldsymbol{A} = \left(\sum_{c=1}^{C} \frac{1}{N_c} \boldsymbol{K}_c \boldsymbol{K}_c^{\mathrm{T}}\right)^{-1} \left(\sum_{c=1}^{C} \frac{1}{N_c^2} \boldsymbol{K}_c \mathbf{1}_{N_c} \boldsymbol{K}_c^{\mathrm{T}}\right) \tag{21}$$

For a new pattern $\boldsymbol{x}$, its projection onto the M-dimensional KDA space is computed by

$$\boldsymbol{y} = \boldsymbol{U}^{\mathrm{T}} \boldsymbol{k}_x \tag{22}$$

where

$$\boldsymbol{k}_x = (k(\boldsymbol{x}, \boldsymbol{x}_1), k(\boldsymbol{x}, \boldsymbol{x}_2), \ldots, k(\boldsymbol{x}, \boldsymbol{x}_N))^{\mathrm{T}}. \tag{23}$$

The algorithm of training KDA is summarised as follows:

1. Construct the non-centred kernel matrix $\boldsymbol{K}$ (16);
2. Compute the centred kernel matrix $\tilde{\boldsymbol{K}}$ (17);
3. Obtain sub-matrices $\boldsymbol{K}_c$ (18);
4. Compute matrix $\boldsymbol{A}$ (21);
5. Eigen-decompose $\boldsymbol{A}$ to obtain eigenmatrix $\boldsymbol{U}$ (20).

Use the following procedure to compute the KDA vector of a new pattern $\boldsymbol{x}$:

1. Compute the kernel vector $\boldsymbol{k}_x$ (23);
2. Compute the M-dimensional projection vector $\boldsymbol{y}$ (22).

We use a toy problem to illustrate the characteristics of KDA, as shown in Fig. 5. Two classes of patterns denoted by circles and crosses respectively have a significant non-linear distribution. To make the results comparable, we try to separate them with a *one*

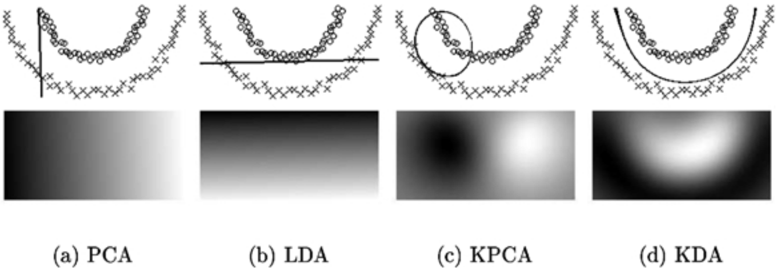


Figure 5. Solving a nonlinear classification problem with, from left to right, PCA, LDA, KPCA and KDA. The top row shows the patterns and the discriminatory boundaries computed by the four methods. The bottom row illustrates the intensity of the one-dimensional features computed using the four methods.

dimensional feature, i.e. the most significant mode of PCA, LDA, KPCA or KDA. The first column shows the patterns and the discriminatory boundaries computed by the four different methods. The second column illustrates the intensity of the one-dimensional features given by PCA, LDA, KPCA and KDA on the region covered by the training patterns. In this experiment, the discriminatory boundary is determined by the value of the discriminatory feature which minimises the misclassification from the given patterns (Bishop, 1995).

It can be seen clearly that PCA and LDA are incapable of providing correct classification because of their linear nature. Neither does KPCA do so since it is designed to extract the overall rather than the discriminatory variation though it is nonlinear in principle. KDA gives the correct classification boundary, and the feature intensity correctly reflects the actual pattern distribution.

3.3. *Extracting the KDA Features of Faces*

For the task of face recognition, one needs to consider two kinds of variation: variation from different face identities and variation from other sources such as pose, illumination and expression. An ideal representation for this problem should be able to maximise the former and minimise the latter. However, from the low-level image characteristics of multi-view face images, the former is not necessarily more significant than the latter.

Although the variation from pose change has been suppressed by normalising multi-view face patterns to shape-and-pose-free, the underlying discriminatory features for different face identities have not been represented explicitly. Also, other sources of variation like illumination and expression still exist which can not be simply ignored.

This problem can be illustrated as in Fig. 6(a) where the facial texture patterns are plotted in the first two dimensions of PCA, the arguably most widely adopted technique in face recognition. In this experiment, 540 facial texture patterns from 12 subjects (45 of each subject) were evaluated. These patterns were used to construct the multi-view face model in Section 2. For the sake of clarity, only the patterns of first four face classes are shown here. It is noted that the variation from different face classes is not efficiently separated from that of other sources of variation, or more precisely, the former is even overshadowed by the latter.

We apply KDA, LDA and KPCA, as well as PCA, to the same set of facial texture patterns. For KDA and KPCA, the Gaussian kernel is adopted,

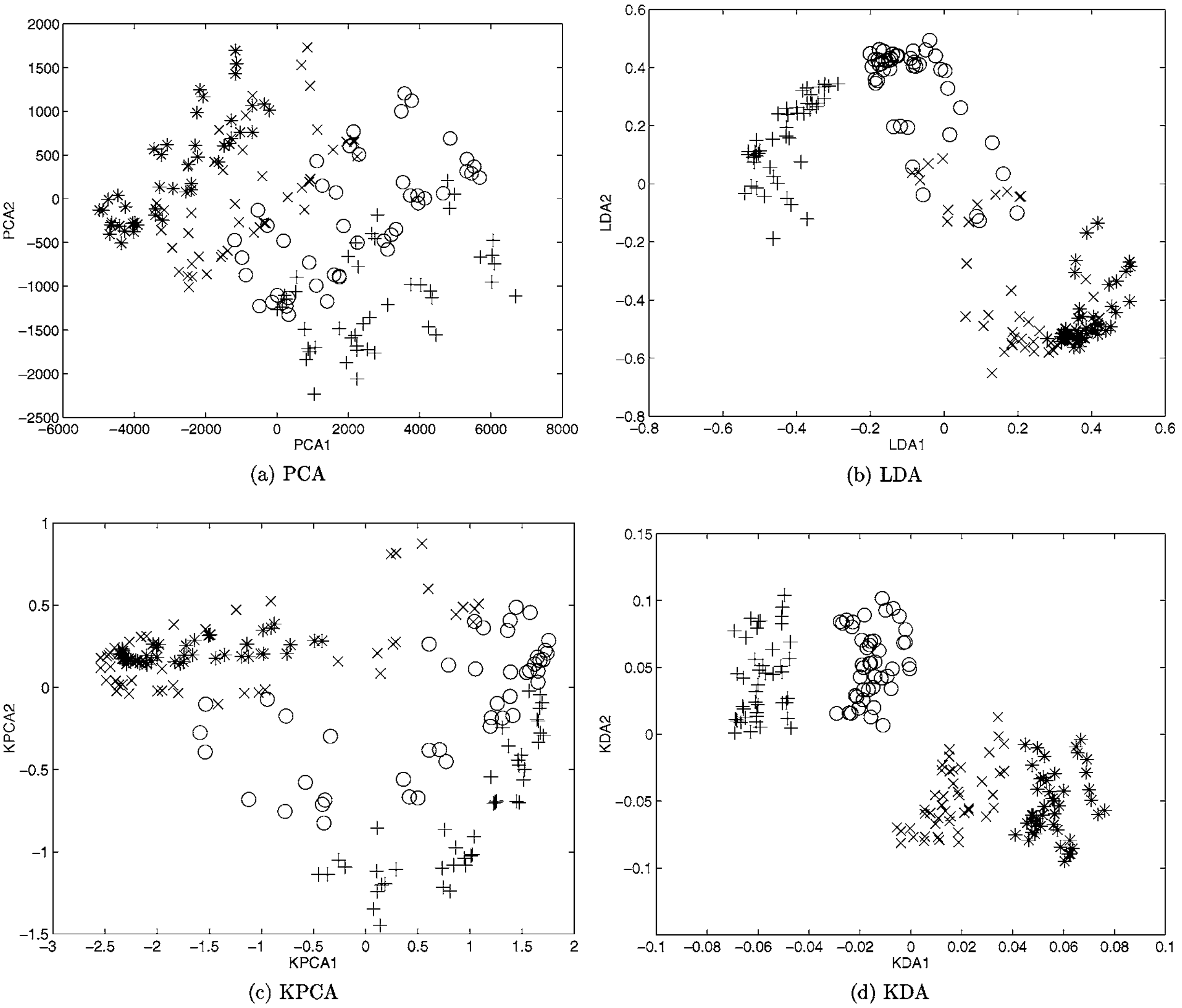
$$k(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{y}\|^2}{2\sigma^2}\right) \tag{24}$$

where $2\sigma^2 = 1$.

The distribution of the patterns are shown in Fig. 6. It is noted that

1. the pattern distributions using PCA and KPCA are not satisfactorily separable since these two techniques are not designed for discriminatory purpose;
2. LDA performs better than PCA and KPCA, but not as well as KDA;
3. KDA provides the best separation performance among the four methods.

For the KDA method, we have experimented with different types of kernel functions such as polynomial, sigmoid and Gaussian kernel functions. Similar results have been obtained for different choice of kernel. Meanwhile, when the parameter of the Gaussian kernel is chosen as $2\sigma^2 = 1$, a satisfactory result is produced in terms of recognition accuracy and reliability.



(a) PCA

(b) LDA

(c) KPCA

(d) KDA

Figure 6. Distribution of multi-view face patterns in PCA, LDA, KPCA and KDA spaces. 540 facial texture patterns from 12 subjects, 45 of each, are used in this experiment. For clarity, only patterns from the first four subjects are plotted.

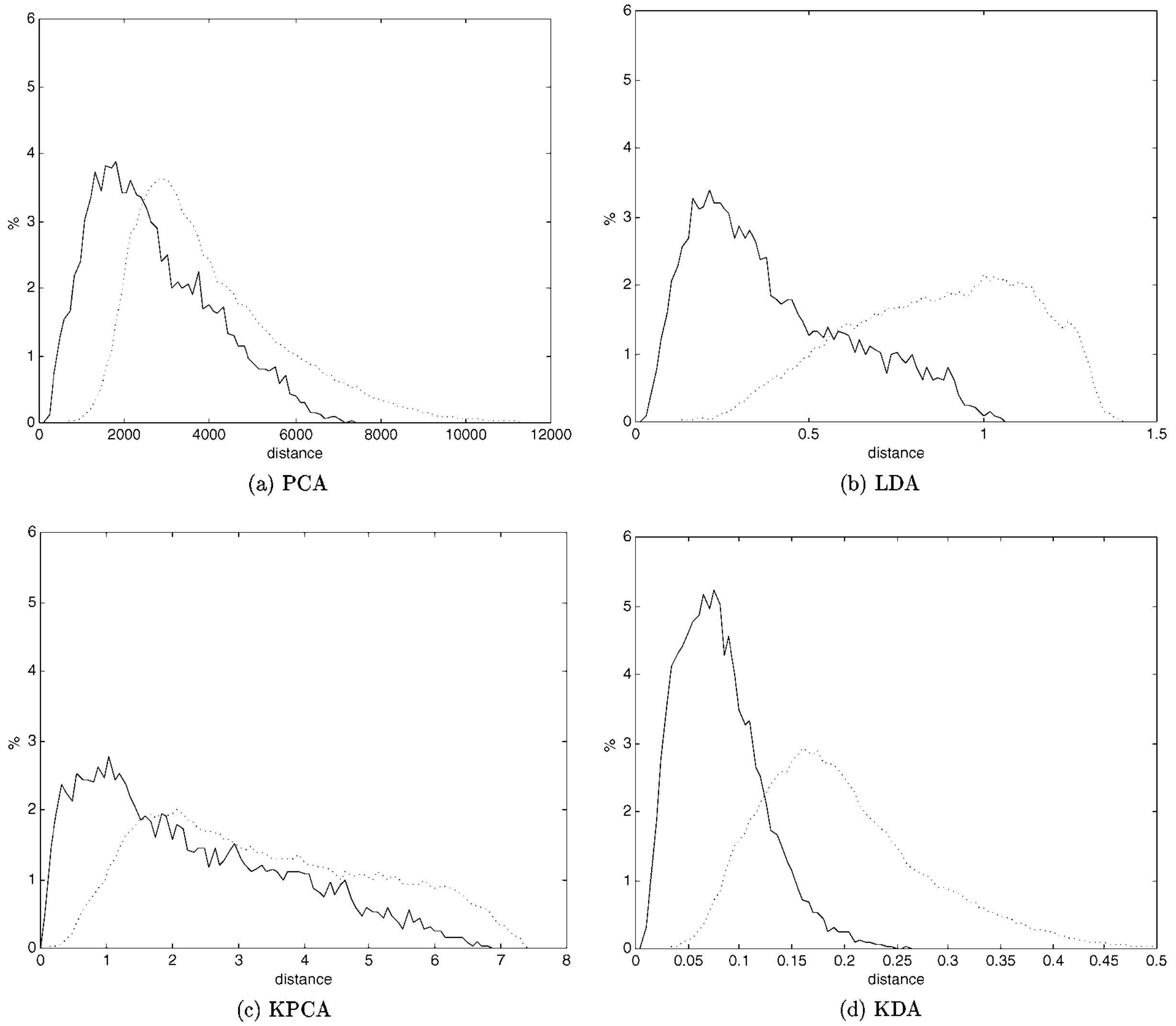
For a quantitative analysis, we plotted the histogram distribution of within-class pattern distance and between-class pattern distance in Fig. 7. The former is distance between face patterns of a same subject, while the latter is the distance between face patterns belonging to different subjects. Obviously, for a good representation, the within-class distance distribution should be dense, close to origin, having a high peak value, and well-separated from the between-class distance distribution. The average within-class distance $\bar{d}_w$ and between-class distance $\bar{d}_b$ are shown in Table 3. To make the results from different methods comparable, we compute the normalised difference of the two distances for each method as

$$dif = \frac{\bar{d}_b - \bar{d}_w}{\bar{d}_w} \tag{25}$$

which can be regarded as a measurement of how largely the within-class patterns are separated from the between-class patterns.

4. Recognising Faces Using Identity Surfaces

As reviewed in Section 1, recognising faces with large pose variation and recognising faces dynamically



(a) PCA

(b) LDA

(c) KPCA

(d) KDA

Figure 7. Histogram distribution of with-class pattern distances (solid lines) and between-class pattern distances (dotted lines). These distances are computed using 540 facial texture patterns from 12 subjects, 45 of each subject.

from video input are two of the most challenging problem in face recognition. Aiming to address these problems, we present in this section an approach to multi-view dynamic face recognition using identity surfaces.

An identity surface is constructed from the discriminatory features of a face class based on pose information. Therefore it is appropriate to deal with the variation from pose change. Moreover, it enables face recognition to be performed dynamically over time. By tracking a moving face from a video input and extracting the discriminatory features for this face, one obtains an object trajectory in a discriminatory feature space. Meanwhile, a set of model trajectories can be constructed on the identity surfaces, one of each face class, using the same pose information and temporal order. Face recognition can then be performed dynamically by matching these two kinds of trajectories.

4.1. Identity Surfaces

One of the most commonly used techniques for recognition is to compute the probabilities of a set of known patterns or the similarities among templates of different classes before selecting the optimal value using a simple metric. For example, the Euclidean distance or

Table 3. The average within-class and between-class distances and their normalised difference values. It is noted that KDA achieves the best separating performance, i.e. the highest *dif* value.

	$\bar{d}_w$	$\bar{d}_b$	*dif*
PCA	2582.497391	4010.782401	0.553063
KPCA	2.216066	3.392148	0.530707
LDA	0.386242	0.874981	1.265372
KDA	0.078392	0.198229	1.528670

the Mahalanobis distance can be adopted if the pattern distribution of each class is compact enough and separable from others. However, usually this simplistic method cannot provide satisfactory solutions to the problem of multi-view face recognition. The reasons are twofold: First, the representation adopted, e.g. the KDA, may not generate a *perfectly* compact distribution of each face class while separating one from another. Second, the distributions of each class cannot be guaranteed to be homogeneous.

When the distribution is irregular, the traditional statistical method for dealing with this problem is to estimate a multi-modal density function for each class. But a very large number of training examples are needed either for parametric or non-parametric modelling. In this work, we do not constrain ourselves to such a strict condition. Instead, we present a novel approach to construct an identity surface for each face class from a sparse sample of multi-view face patterns.

As stated previously, one of the key problems of multi-view face recognition is how to separate two kinds of variations: variation from different subjects and variation from pose. Fortunately, we can estimate the pose of a face when fitting the multi-view face model on a face image (Section 2.5.2).

If we use the pose information explicitly rather than indifferently computing all face patterns of different views, a significant improvement to face identity modelling can be expected. Based on this idea, we developed a method of multi-view face recognition using identity surfaces. The basic idea of the identity surfaces is similar to the parametric eigenspace method presented by Murase and Nayar (1994, 1995).

If we assume that only the appearance variation caused by *rotation in depth* is concerned, i.e. the variation from expression, illumination and facial make-up is excluded, each face class can be represented by a unique hyper surface based on pose information. In other words, the two basis coordinates stand for the head pose: tilt and yaw, and the other coordinates are used to represent the discriminatory features of faces, e.g. the KDA vectors. For each pair of tilt and yaw values, there is one unique "point" for a face class. The distribution of all the "points" of the same face class with regard to pose change form a hyper surface in the space spanned by the discriminatory features and pose. We call this surface an identity surface.

Even for the human vision system, the performance of face recognition is not very reliable on *static images*. However, the situation can be considerably improved when video input is available where faces move continuously. Recall the discussions in Section 1, psychological and physiological research suggests that modelling and recognising moving faces dynamically have the potential of achieving a superior performance over that on static images.

Once the identity surfaces are constructed, face recognition can be performed dynamically from a video sequence. As shown in Fig. 8, when a face is detected and tracked in an input video sequence, one obtains the *object trajectory* of the face in the feature space. Also, its projection onto each of the identity surfaces with the same poses and temporal order forms a *model trajectory* of the specific face class. It can be regarded as the ideal trajectory of this face class encoded by the same spatio-temporal information (pose information and temporal order from the video sequence) as the tracked face. Then face recognition can be carried out by matching the object trajectory with a set of model trajectories. Compared to face recognition on static images, this approach can be more reliable and accurate. For example, it is difficult to decide whether the pattern *X* in Fig. 8 belongs to subject A or B for a single pattern. However, if we know that *X* is tracked along the object trajectory, it is more likely to be subject A than B.

Based on the discussion above, we propose the following process of video-based face recognition:

Registration. Construct the identity surface for each face class from one or more training sequences;

Tracking. Fit the multi-view dynamic model (Section 2) on an input video sequence containing faces to be recognised, and extract the discriminatory features;

Recognition. Compute the object and model trajectories and match these trajectories.

The issues of model fitting and feature extraction have been presented in Sections 2 and 3. We will discuss the issues of registration and recognition in the rest of this section.

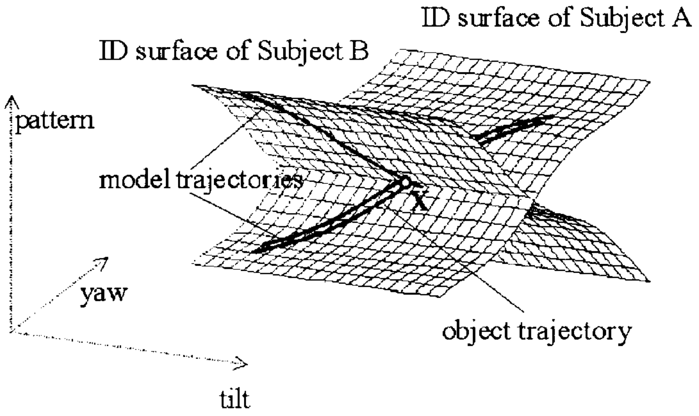


Figure 8. An identity surface is a unique hyper surface for a face class in a pose-parameterised discriminatory feature space. By matching the object trajectory and model trajectories on identity surfaces, face recognition can be performed dynamically from a video input.

4.2. Construction Algorithm

If sufficient patterns of a face class in different views are available, the identity surface of this face class can be constructed precisely. However, we do not require such a strict condition. In this work, we develop a method to synthesise the identity surface of a face class from a small sample of face patterns which sparsely cover the view sphere.

The basic idea is to approximate the identity surface using a set of N_p planes separated by a number of N_v predefined views. The problem can be formally defined as follows:

Suppose x, y are tilt and yaw respectively, z is the discriminatory feature vector of a face pattern. A list $(x_{01}, y_{01}), (x_{02}, y_{02}), \ldots, (x_{0N_v}, y_{0N_v})$ gives predefined views which divide the view sphere into N_p grids. On each grid, the identity surface of a face class is approximated by a plane

$$z = \boldsymbol{a}x + \boldsymbol{b}y + \boldsymbol{c} \tag{26}$$

Suppose the M_i sample patterns covered by the ith plane are $(x_{i1}, y_{i1}, z_{i1}), (x_{i2}, y_{i2}, z_{i2}), \ldots, (x_{iM_i}, y_{iM_i}, z_{iM_i})$, then one minimises

$$\mathcal{Q} = \sum_{i}^{N_p} \sum_{m}^{M_i} \|\boldsymbol{a}_i x_{im} + \boldsymbol{b}_i y_{im} + \boldsymbol{c}_i - z_{im}\|^2 \tag{27}$$

$$\begin{aligned} \text{subject to:}\quad & \boldsymbol{a}_i x_{0k} + \boldsymbol{b}_i y_{0k} + \boldsymbol{c}_i = \boldsymbol{a}_j x_{0k} + \boldsymbol{b}_j y_{0k} + \boldsymbol{c}_j \\ & k = 0, 1, \ldots, N_v, \\ & \text{planes } i, j \text{ intersect at } (x_{0k}, y_{0k}). \end{aligned} \tag{28}$$

This is a Quadratic Programming problem which can be solved using the interior point method (Vanderbei, 1994).

Figure 9 shows the identity surface of a face class constructed from all 45 example views (−20° ∼ +20° in tilt and −40° ∼ +40° in yaw with an interval of 10°) and the approximated identity surface using only 15 example views, i.e. the same pose ranges but with an interval of 20°. A 10-dimensional KDA feature vector was adopted to represent the multi-view faces. The identity surfaces are shown in the first three KDA dimensions only. Comparing the two identity surfaces in this figure, it is indicated that identity surfaces can be constructed from a small set of face patterns which sparsely cover the view sphere.

In practice, the identity surface of a subject can be constructed from one or more example sequences containing the face of the subject. For example, we can record a small video clip of the subject while he/she rotates the head in front of a camera. After applying the multi-view dynamic face model described in Section 2 on the video sequence, we obtain a set of face patterns of this subject. The discriminatory features, e.g. KDA features, can then be extracted from these patterns to construct the identity surface.

4.3. Video-Based Face Recognition

For an unknown face image, one first fits the multi-view dynamical face model onto the image and extracts the discriminatory features of the face to yield a pose augmented feature vector (x, y, z_0) where z_0 is

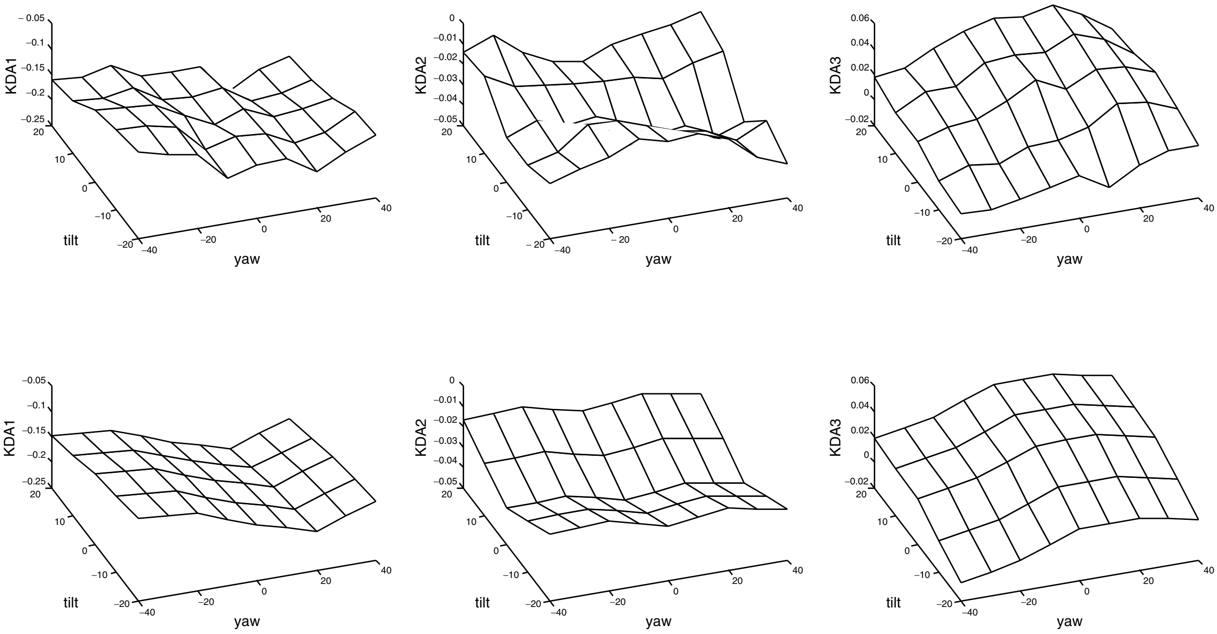


Figure 9. The identity surface constructed from all 45 views (first row) and that approximated from 15 prototype patterns (second row). A 10-dimensional KDA feature vector was adopted to represent the multi-view faces. Only the first three KDA components are shown here.

the discriminatory vector and x, y are the pose in tilt and yaw. Then the pattern distance to one of the identity surfaces can be computed as the Euclidean distance between z_0 and the corresponding point z on the identity surface

$$d = \|z_0 - z\| \tag{29}$$

where z is given by (26).

It is important to note that Euclidean distance is more appropriate for KDA and LDA while Mahalanobis distance is more efficient for PCA and KPCA since the discriminatory features are extracted in the former case and the general variation is concerned in the latter.

Recognising faces from a video sequence using (29) gives the frame-by-frame recognition results. However, more reliable and accurate recognition can be achieved by trajectory matching.

When a face is tracked in an input video sequence, an object trajectory can be obtained by projecting the face patterns into the pose-parameterised feature space. Furthermore, a model trajectory can be built on the identity surface of each subject using the same pose information and temporal order of the object trajectory. Those two kinds of trajectories, given any sequence of specific poses in a temporal order, encode the spatio-temporal information of the tracked face. Finally, recognition is performed dynamically by matching the object trajectory to a set of identity model trajectories.

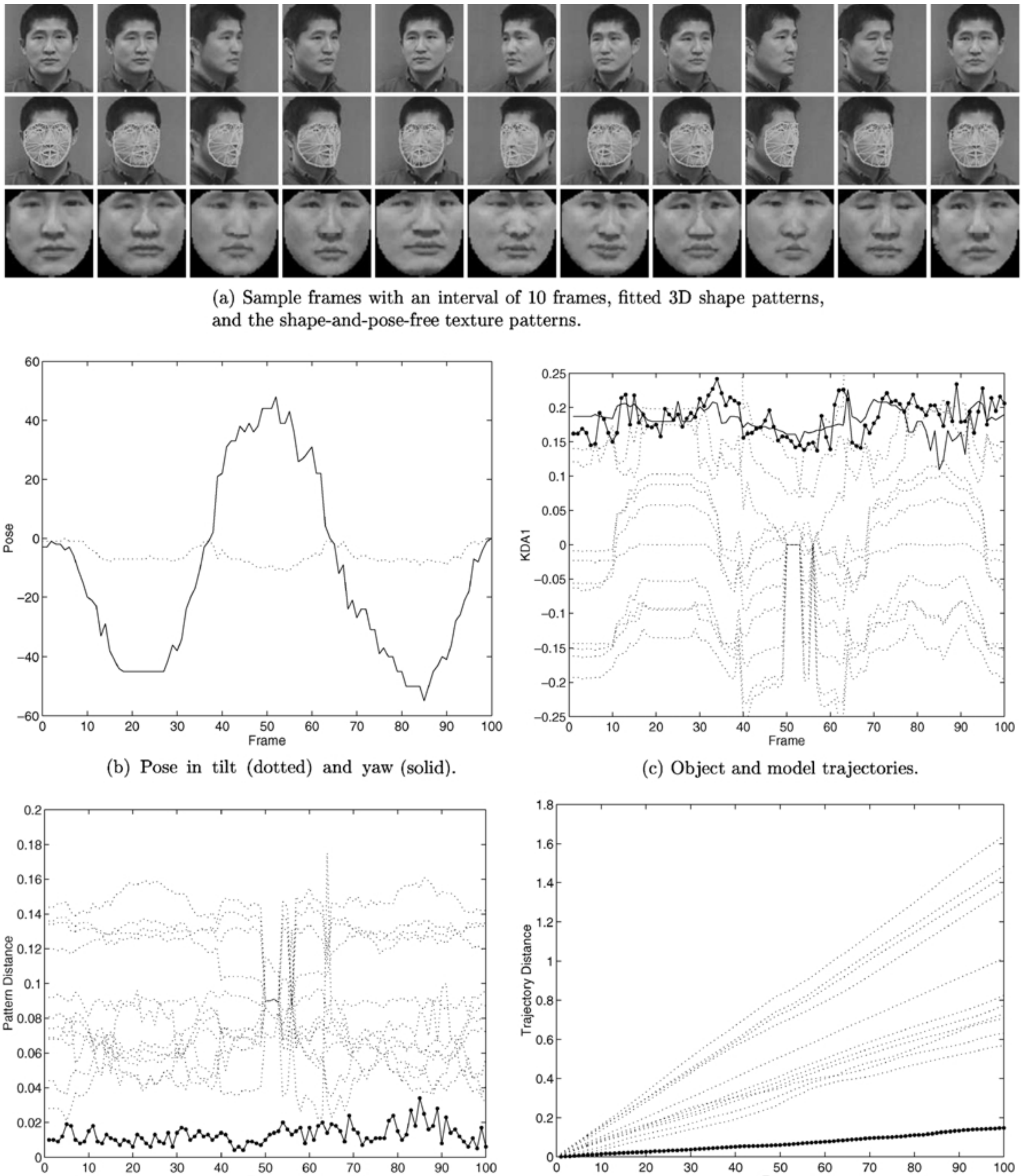
A preliminary realisation of this approach is implemented by computing a trajectory distance

$$d_m = \sum_{i=1}^{t} w_i d_{mi} \tag{30}$$

where d_{mi} is the pattern distance to the identity surface of the mth face class in the ith frame computed using (29), and w_i is the weight of this distance. Recognition is performed by selecting the subject with minimum trajectory distance.

4.4. Experiments

We applied this approach to a small scale multi-view face recognition problem. Twelve sequences, each taken from a set of 12 subjects, were used as training sequences to construct the identity surfaces. The number of frames contained in each sequence varies from 40 to 140. KDA was adopted to extract the discriminatory features.



(a) Sample frames with an interval of 10 frames, fitted 3D shape patterns, and the shape-and-pose-free texture patterns.

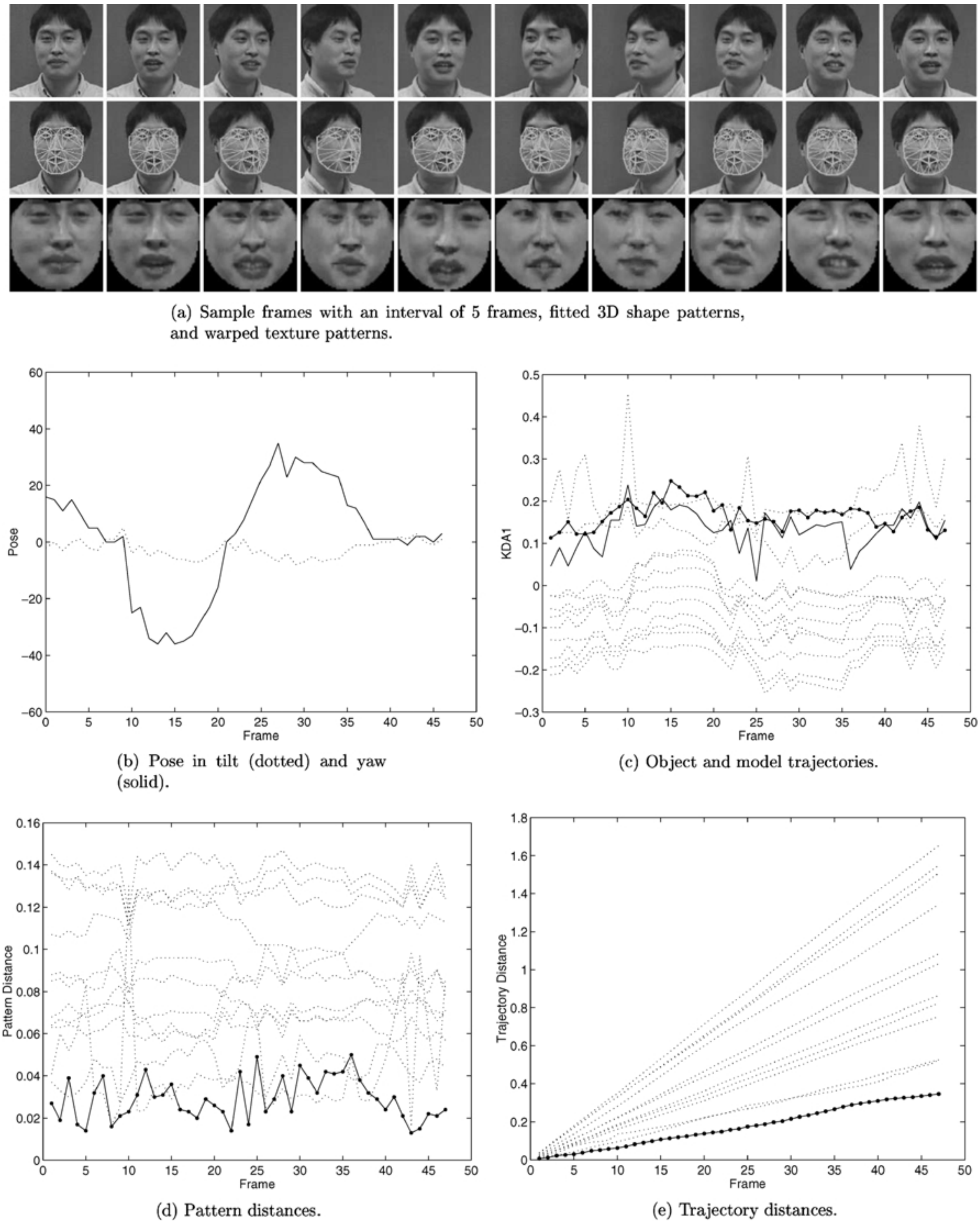
(b) Pose in tilt (dotted) and yaw (solid).

(c) Object and model trajectories.

(d) Pattern distances.

(e) Trajectory distances.

Figure 10. Video-based multi-view face recognition. (c) shows the object trajectory (solid line with dots) and model trajectories in the first KDA dimension where the model trajectory from the ground-truth subject is highlighted with solid line. It is noted from (d) and (e) that the pattern distances can give an accurate recognition result; however, the trajectory distances provide a more reliable performance, especially its accumulated effects (i.e. discriminatory ability) over time.



(a) Sample frames with an interval of 5 frames, fitted 3D shape patterns, and warped texture patterns.

(b) Pose in tilt (dotted) and yaw (solid).

(c) Object and model trajectories.

(d) Pattern distances.

(e) Trajectory distances.

Figure 11. Face recognition on a face sequence with significant expression change. The pattern distance is less reliable for a few frames, however, the trajectory distance still provides a reliable and accurate recognition.

We randomly selected 180 images (15 images of each subject which approximately cover the view sphere) to train the KDA.[2] The first ten KDA basis vectors were used to construct the identity surfaces. Then recognition was performed on new test sequences of these subjects.

Figure 10 shows the results on one of the test seqquences. It is noted that a more reliable performance is achieved when recognition is carried out using the trajectory distances which include the accumulated evidence over time, although the pattern distances in each individual frame already provides good recognition accuracy on a frame by frame basis.

Figure 11 shows the results on another sequence where the face is undergoing significant expression change. Since all the training face images are taken in neutral expression, the results of model fitting is not as good as those in Fig. 10. Also, the pattern distance from individual frames only achieved a recognition accuracy rate of 61.7% (29 out of 47 frames). However, it is important to point out that the trajectory distance still provided a reliable and accurate recognition (100% in this sequence).

In these experiments, we adopted the shape-and-pose-free texture patterns from the multi-view face model (Section 2) and the KDA vectors of these texture patterns (Section 3) to construct the identity surfaces and object/model trajectories. However, it is important to note that other kinds of representations can also be incorporated in the framework of identity surface based dynamic face recognition.

It is also noted in these experiments that, although the face patterns are represented in the shape-and-pose-free format, there are still some kinds of residual variation which are related to pose. Among them, the most noticeable ones are from illumination change and self-shading. We have discussed in Section 2.3 why only geometrical information is considered when computing the shape-and-pose-free texture patterns. However, these experimental results suggest that illumination estimation and correction may significantly improve the performance of face modelling and multi-view face recognition. We will investigate these issues in our future work.

5. Conclusions

In this paper, we have presented a comprehensive approach to modelling faces across multiple views, extracting the non-linear discriminatory features and dynamically recognising faces across views and over time. The key issues of this work can be summarised as follows:

1. Recognising faces across views is more challenging than that from a fixed view because of the severe non-linearity caused by rotation in depth, self-occlusion, self-shading, and illumination change. To model faces with large pose variation, we developed a dynamic face model, which includes a 3D PDM, a shape-and-pose-free texture model, and an affine geometrical model. By representing faces with the shape-and-pose-free texture patterns, the variance from pose change is suppressed.
2. PCA, LDA and KPCA have been widely used in face recognition. But PCA and LDA are limited to the linear applications while KPCA intends to capture the *overall* rather than the *discriminatory* variance of all patterns though it is non-linear. To efficiently extract the discriminatory features of multi-class patterns with severe non-linearity, the KDA is developed in this work. We applied this method to multi-view face recognition, and significant improvement has been achieved both in reliability and accuracy.
3. Psychological and physiological research suggests that modelling and recognising moving faces dynamically has the potential for achieving a superior performance over that on static images. Inspired by this idea, we present an approach to dynamic face recognition using identity surfaces. The identity surfaces can be constructed from a sparse sample of multi-view face images. Dynamic face recognition can then be performed by computing and matching the object and model trajectories. A more reliable recognition is achieved since these trajectories encode the spatio-temporal information of a moving face and provide the accumulated evidence of identity.

One of the main drawbacks of this approach is that only geometrical information is considered when normalising facial texture to shape-and-pose-free patterns. However, as noted in the previous sections, there are still some kinds of residual variation in the normalised patterns which are related to pose. Among them, the most noticeable ones are from illumination change and self-shading. Extended work on illumination estimation and correction is needed to improve the representative ability of the model.

Another limitation of the work is the intensive computation involved in KDA. To obtain the KDA projection of an unknown pattern, one has to compute the kernel functions of this pattern with all training examples. Actually this is a common limitation of all kernel techniques such as KPCA and SVMs. Though some methods such as the reduced set technique (Burges, 1996; Burges and Scholkopf, 1997) can be adopted for computation reduction, an additional non-linear optimisation problem is usually introduced which is not guaranteed to provide a global optimal solution.

In addition, some of the implementation such as trajectory matching is still simplistic in its present form. The trajectory distance is computed as a weighted summation, therefore it does not make any difference to the results of recognition if the information of each frame comes either in a random order or in the temporal order, as is the case here, though the temporal order is still very useful in the tracking process. We believe it is an interesting issue for both psychological and artificial vision research to exploit the underlying mechanism of this spatio-temporal dynamics, and extensive further work needs to be conducted.

Acknowledgments

The authors thank Jamie Sherrah, Eng-Jon Ong, Ting-Hsun Chang, Jeffrey Ng, Paul Verity and Dennis Parkinson for discussions on this research.

Notes

1. In particular, the algorithm presented in this paper is equivalent to that in Baudat and Anouar (2000). In our algorithm, one inverse matrix computation and one eigen-decomposition are conducted, while two steps of eigen-decomposition are performed in the latter.
2. To simplify the computation, normally we do not use all the patterns of each subject to train the KDA since the sizes of the kernel matrix $\boldsymbol{K}$ and $\boldsymbol{K}_c$ are directly related to the number of training examples. A pragmatic approach to selecting the KDA training patterns is to factor-sample the patterns from the training sequences so that the resulting patterns uniformly cover the view sphere.

References

Atick, J., Griffin, P., and Redlich, A. 1996. Statistical approach to shape from shading: Reconstruction of 3d face surfaces from single 2d images. *Neural Computation*, 8(6):1321–1341.

Baudat, G. and Anouar, F. 2000. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12:2385–2404.

Bennett, A. and Craw, I. 1991. Finding image features using deformable templates and detailed prior statistical knowledge. In *British Machine Vision Conference*, Glasgow, UK, pp. 233–239.

Bishop, C.M. 1995. *Neural Networks for Pattern Recognition*. Clarendon Press: Oxford.

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

Bruce, V., Burton, A., and Hancock, P. 1998a. Comparisons between human and computer recognition of faces. In *IEEE International Conference on Automatic Face & Gesture Recognition*, Nara, Japan, pp. 408–413.

Bruce, V., Hancock, P., and Burton, A. 1998b. Human face perception and identification. In *Face Recognition: From Theory to Applications*, H. Weschler, J. Philips, V. Bruce, F. Fogelman-Soulie, and T. Huang (Eds.), Springer-Verlag, pp. 51–72.

Brunelli, R. and Poggio, T. 1993. Face recognition: Features vs. templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10):1042–1062.

Burges, C.J.C. 1996. Simplified support vector decision rules. In *Proceedings, 13th Intl. Conf. on Machine Learning*, L. Saitta (Ed.), Morgan Kaufmann: San Mateo, CA, pp. 71–77.

Burges, C.J.C. and Scholkopf, B. 1997. Improving the accuracy and speed of support vector learning machines. In *Advances in Neural Information Processing Systems*, M. Mozer, M. Jordan, and T. Petsche (Eds.), MIT Press: Cambridge, MA, vol. 9, pp. 375–381.

Choudhury, T., Clarkson, B., Jebara, T., and Pentland, A. 1999. Multimodal person recognition using unconstrained audio and video. In *International Conference on Audio- and Video-Based Person Authentication*, pp. 176–181.

Cootes, T., Taylor, C., and Lanitis, A. 1994. Active shape models: Evaluation of a multi-resolution method for improving image search. In *British Machine Vision Conference*, York, England, vol. 1, pp. 327–336.

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

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

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

Craw, I., Tock, D., and Bennett, A. 1992. Finding face features. In *European Conference on Computer Vision*, Santa Margherita Ligure, Italy, pp. 92–96.

DeCarlo, D. and Metaxas, D. 2000. Optical flow constraints on deformable models with applications to face tracking. *International Journal of Computer Vision*, 38(2):99–127.

Edwards, G., Lanitis, A., Taylor, C., and Cootes, T. 1996. Statistical models of face images—Improving specificity. In *British Machine Vision Conference*, Edinburgh, Scotland, vol. 2, pp. 765–774.

Edwards, G., Lanitis, A., Taylor, C., and Cootes, T. 1998a. Statistical models of face images—Improving specificity. *Image and Vision Computing*, 16(3):203–211.

Edwards, G., Taylor, C., and Cootes, T. 1998b. Learning to identify and track faces in sequences. In *IEEE International Conference on Automatic Face & Gesture Recognition*, Nara, Japan, pp. 260–267.

Edwards, G., Taylor, C., and Cootes, T. 1999. Improving identification performance by integrating evidence from sequences. In *IEEE Conference on Computer Vision and Pattern Recognition*, Fort Collins, CO, USA, vol. 1, pp. 486–491.

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

Fisher, R.A. 1938. The statistical utilization of multiple measurements. *Annals of Eugenics*, 8:376–386.

Fukunaga, K. 1972. *Introduction to Statistical Pattern Recognitiion*. Academic Press.

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

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

Howell, A. and Buxton, H. 1996. Towards unconstrained face recognition from image sequences. In *IEEE International Conference on Automatic Face & Gesture Recognition*, Vermont, USA, pp. 224–229.

Jebara, T. and Pentland, A. 1997. Parametrized structure from motion for 3D adaptive feedback tracking of faces. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Kass, M., Witkin, A., and Terzopoulos, D. 1987. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331.

Knight, B. and Johnston, A. 1997. The role of movement in face recognition. *Visual Cognition*, 4:265–274.

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

Lanitis, A., Taylor, C., and Cootes, T. 1997. Automatic interpretation and coding of face images using flexible models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):743–756.

Li, H., Roivainen, P., and Forchheimer, R. 1993. 3-d motion estimation in model-based facial image coding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):545–555.

Li, Y., Gong, S., and Liddell, H. 2000a. Recognising the dynamics of faces across multiple views. In *British Machine Vision Conference*, Bristol, UK, pp. 242–251.

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

Li, Y., Gong, S., and Liddell, H. 2001. Constructing structures of facial identities using Kernel Discriminant Analysis. In *The 2nd International Workshop on Statistical and Computational Theories of Vision*, Vancouver, Canada.

McKenna, S. and Gong, S. 1998. Recognising moving faces. In *Face Recognition: From Theory to Applications*, Wechsler, Philips, Bruce, Fogelman-Soulie, and Huang (Eds.), Springer-Verlag, pp. 578–588.

Mika, S., Ratsch, G., Weston, J., Scholkopf, B., and Muller, K. 1999. Fisher discriminant analysis with kernels. In *IEEE Neural Networks for Signal Processing Workshop*, pp. 41–48.

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

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

Moghaddam, B., Wahid, W., and Pentland, A. 1998. Beyond eigenfaces: Probabilistic matching for face recognition. In *IEEE International Conference on Automatic Face & Gesture Recognition*, Nara, Japan, pp. 30–35.

Murase, H. and Nayar, S.K. 1994. Illumination planning for object recognition using parametric eigenspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(12):1219–1227.

Murase, H. and Nayar, S.K. 1995. Visual learning and recognition of 3-D objects from appearance. *International Journal of Computer Vision*, 14:5–24.

Okubo, M. and Watanabe, T. 1998. Lip motion capture and its application to 3-D molding. In *IEEE International Conference on Automatic Face & Gesture Recognition*, Nara, Japan, pp. 187–192.

Osuna, E., Freund, R., and Girosi, F. 1997a. Support vector machines: Training and applications. Technical report, Massachusetts Institute of Technology. AI Memo 1602.

Osuna, E., Freund, R., and Girosi, F. 1997b. Training support vector machines: An application to face detection. In *Proc. Computer Vision and Pattern Recognition'97*, pp. 130–136.

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

Roth, V. and Steinhage, V. 1999. Nonlinear discriminant analysis using kernel functions. In *Advances in Neural Information Processing Systems*, S. Solla, T. Leen, and K.-R. Müller (Eds.), MIT Press, vol. 12, pp. 568–574.

Scholkopf, B. 1997. *Support Vector Learning*. R. Oldenbourg Verlag: Munich.

Scholkopf, B., Smola, A., and Muller, K.-R. 1997. Kernel principal component analysis. In *Artificial Neural Networks—ICANN'97*, W. Gerstner, A. Germond, M. Hasler, and J.-D. Nicoud, (Eds.), Lecture Notes in Computer Science, Springer Berlin, pp. 583–588.

Shakunaga, T., Ogawa, K., and Oki, S. 1998. Integration of eigentemplate and structure matching for automatic facial feature detection. In *IEEE International Conference on Automatic Face & Gesture Recognition*, Nara, Japan, pp. 94–99.

Sirovich, L. and Kirby, M. 1987. Low-dimensional procedure for the characterization of human faces. *Journal of Optical Society of America*, 4:519–524.

Steffens, J., Elagin, E., and Neven, H. 1998. Personspotter—Fast and robust system for human detection, tracking and recognition. In *IEEE International Conference on Automatic Face & Gesture Recognition*, Nara, Japan, pp. 516–521.

Sung, K. and Poggio, T. 1994. Example-based learning for view-based human face detection. Technical report, Massachusetts Institute of Technology. A.I. MEMO 1521.

Swets, D. and Weng, J. 1996. Using discriminant eigenfeatures for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):831–836.

Turk, M. and Pentland, A. 1989. Face processing: Models for recognition. In *Intelligent Robots and Computer Vision VIII: Algorithms and Techniques*, Philadelphia, PA, USA, pp. 22–32.

Vanderbei, R. 1994. Loqo: An interior point code for quadratic programming. Technical report, Princeton University. Technical Report SOR 94–15.

Vapnik, V. 1995. *The Nature of Statistical Learning Theory*. Springer Verlag: New York.

Vetter, T. 1998. Synthesis of novel views from a single face image. *International Journal of Computer Vision*, 28(2):103–116.

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

Waite, J.B. and Welsh, W.J. 1990. An application of active contour models to head boundary location. In *British Machine Vision Conference*, pp. 407–412.

Wu, H., Yokoyama, T., Pramadihanto, D., and Yachida, M. 1996. Face and facial feature extraction from color image. In *IEEE International Conference on Automatic Face & Gesture Recognition*, Vermont, USA, pp. 345–350.

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

Yokoyama, T., Yagi, Y., and Yachida, M. 1998. Facial contour extraction model. In *IEEE International Conference on Automatic Face & Gesture Recognition*, Nara, Japan, pp. 254–259.

Yuille, A., Hallinan, P., and Cohen, D. 1992. Feature extraction from faces using deformable templates. *International Journal of Computer Vision*, 8(2):99–111.

Zhao, W. and Chellappa, R. 2000. SFS based view synthesis for robust face recognition. In *IEEE International Conference on Automatic Face & Gesture Recognition*, Grenoble, France, pp. 285–292.

Zhao, W., Krishnaswamy, A., Chellappa, R., Swets, D., and Weng, J. 1998. Discriminant analysis of principal components for face recognition. In *Face Recognition: From Theory to Applications*, Wechsler, Philips, Bruce, Fogelman-Soulie, and Huang (Eds.), Springer-Verlag, pp. 73–85.