Video-Based Online Face Recognition Using Identity Surfaces

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

Abstract

Recognising faces across multiple 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 change of illumination. The problem can be related to the problem of modelling the spatio-temporal dynamics of moving faces from video input for unconstrained live face recognition. Both problems remain largely under-developed. To address the problems, a novel approach is presented in this paper. A multi-view dynamic face model is designed to extract the shape-and-pose-free *texture patterns of faces. The model provides a precise correspondence to the task of recognition since the 3D shape information is used to warp the multi-view faces onto the model mean shape in frontal-view. The* identity surface *of each subject is constructed in a discriminant feature space from a sparse set of face texture patterns, or more practically, from one or more learning sequences containing the face of the subject. Instead of matching templates or estimating multi-modal density functions, face recognition can be performed by computing the pattern distances to the* identity surfaces *or trajectory distances between the object and model trajectories. Experimental results depict that this approach provides an accurate recognition rate while using trajectory distances achieves a more robust performance since the trajectories encode the spatio-temporal information and contain accumulated evidence about the moving faces in a video input.*

1 Introduction

The issue of face recognition has been extensively addressed over the past decade. Various models and approaches have been proposed aiming to solve the problem under different assumptions and constraints. Among them, the eigenface approach proposed in [16, 17] uses Principal Component Analysis (PCA) to code face images and capture face features. In [19], face recognition is performed by Elastic Graph matching based on a Gabor wavelet transform. The Active Shape Model (ASM) and Active Appearance Model (AAM) capturing both shape and shape-free grey-level appearance of face images have been applied to face modelling and recognition [4, 3]. These methods have been mostly applied in frontal-view or near frontal-view face recognition.

However, recognising faces with large pose variation is more challenging 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 intended for recognising faces under varying views[13]. Li *et al.* [10] presented a view-based piece-wise SVM (Support Vector Machine) model of the face space. Cootes *et al.* [5] proposed the view-based Active Appearance Models which employ three models for profile, half-profile and frontal views. But the division of the face space in these methods is rather arbitrary, ad hoc and often coarse. Both ASM and AAM have been extended to nonlinear cases across views based on Kernel Principal Component Analysis (KPCA)[14]. These nonlinear models aimed at corresponding dynamic appearances of both shape and texture across views, but the computation involved is rather intensive.

Also, in most of the previous work, the basic methodology adopted for recognition is largely based on matching static face image patterns in a feature space. 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 [9] argued that recognition of famous faces shown in photographic negatives could be significantly enhanced when the faces were shown moving rather than static. Bruce *et al.* [1, 2] 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. In a computer vision system, when faces are tracked consecutively, not only more information of those faces from different views but also the spatio-temporal connection between those faces can be obtained [7]. Yamaguchi *et al.* [20] presented a method for face recognition from sequences by building a subspace for the detected faces on the given sequence and then matching the subspace with prototype subspaces. Gong *et al.* [8] introduced an approach that uses Partially Recurrent Neural Networks (PRNNs) to recognise temporal signatures of faces. Other work of recognising faces from video sequences include [6, 15]. Nevertheless, the issue of recognising the dynamics of human faces under a spatio-temporal context remains largely unresolved.

In this paper, we present a novel approach to video-based face recognition. In the registration stage, an *identity surface* for each subject to be recognised is constructed in a discriminant feature space from one or more learning sequences , while in run-time, recognition is performed by computing the pattern distances to the *identity surfaces* or the trajectory distances between the *object trajectory* and a set of *model trajectories* which encode the spatio-temporal information of a moving face over time. A multi-view dynamic face model is designed to extract the *shape-and-pose-free* texture patterns from sequences for accurate across-view registration in both training and run-time.

The remaining part of this paper is arranged as follows: The multi-view dynamic face model is described in Section 2. *Identity surface* synthesis, object and model trajectory construction, and dynamic face recognition are presented in Section 3. Conclusions are drawn in Section 4.

2 Multi-View Dynamic Face Model

Our multi-view dynamic face model [12] consists of a sparse 3D Point Distribution Model (PDM) [4] 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 3D shape vector of a face is estimated from a set of 2D face images in different views, i.e. given a set of 2D face images with known pose and 2D positions of the landmarks, the 3D shape vector can be estimated using linear regression. To decouple the covariance between shape and texture, a face image fitted by the shape model is warped to the mean shape at frontal view (with 0° in both tilt and yaw), obtaining a *shape-and-pose-free* texture pattern. This is implemented by forming a triangulation from the landmarks and employing a piece-wise affine transformation between each triangle pair. 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. We applied PCA on the 3D shape patterns and *shape-and-pose-free* texture patterns respectively to obtain a low dimensional statistical model.

Figure 1 shows the sample face images used to construct the model, the landmarks labelled on each image, the 3D shape estimated from these labelled face images, and the extracted *shape-and-pose-free* texture patterns.

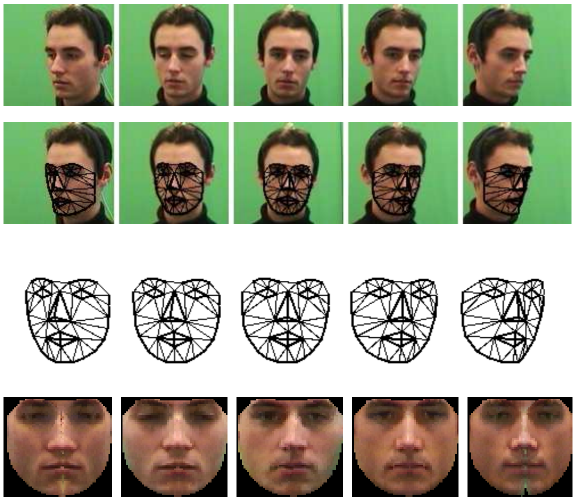

Figure 1. Multi-view dynamic face model. From top to bottom are sample training face images (first row), the landmarks labelled on the images (second row), the estimated 3D shape rotating from -40° to $+40^\circ$ in yaw and with tilt fixed on 0° , and the extracted shape-and-pose-free texture patterns.

Based on the analysis above, a face pattern can be represented in the following way. First, the 3D shape model is fitted to a given image or video sequence containing faces. A Support Vector Machine based pose estimation method [10] is employed to obtain the tilt and yaw angles of the face in the image. Then the face texture is warped onto the mean shape of the 3D PDM model in frontal view. Finally, by adding parameters controlling pose, shift and scale, the complete parameter set of the dynamic model for a given face pattern is $\mathbf{c} = (\mathbf{s}, \mathbf{t}, \alpha, \beta, dx, dy, r)^\mathtt{T}$ where $\mathbf{s}$ is the shape parameter, $\mathbf{t}$ is the texture parameter, (α, β) is pose in tilt and yaw, (dx, dy) is the translation of the centroid of the face, and r is its scale. More details of model construction and fitting are described in [12].

The *shape-and-pose-free* texture patterns obtained from model fitting are adopted for face recognition. In our experiments, we also tried to use the shape patterns for recognition, however, the performance was not as good as that of using textures.

3 Recognising Faces Using Identity Surfaces

For appearance based models, it is more challenging to recognise faces across multiple views since (1) the appearance of different people from a same view is more similar than that of one person from different views, and (2) the rotation in depth, self-occlusion and self-shading bring severe non-linearity to the task. Therefore emphasising the variation between different subjects and suppressing the variation within each subject at the same time, are the key issues of multi-view face recognition. The most widely used techniques for face recognition include computing the Euclidean or Mahalanobis distance to a mean vector or a template of a face class, and estimating the density of patterns using multi-modal models. However, these techniques cannot provide straightforward solutions to the problem of multi-view face recognition. This situation is illustrated in Figure 2 where the PCA patterns of multi-view face images from four subjects are mingled together. More precisely, the variation between different subjects is overwhelmed by that from pose change.

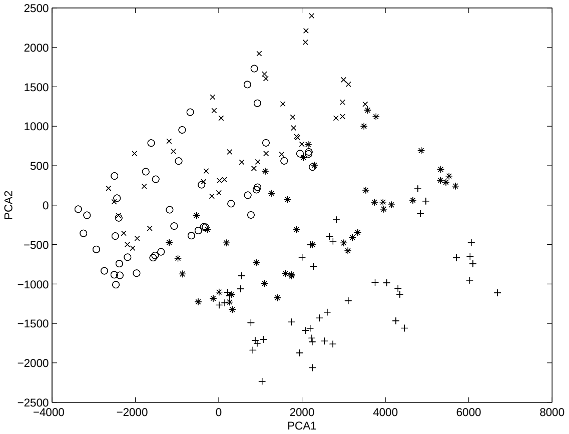

Figure 2. Distribution of multi-view face patterns from four subjects represented in two PCA dimensions. There are mainly two kinds of variation among these patterns: variation from different subjects and variation from pose change. Unfortunately the former is overwhelmed by the latter.

To address this problem, we propose an approach to construct *identity surfaces* in a discriminant feature space for multi-view face recognition where the pose information of faces is explicitly used. Each subject to be recognised is 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 discriminant feature patterns of faces. For each pair of tilt and yaw, there is one unique "point" for a face class. The distribution of all these "points" of a same face forms a hyper surface in this feature space. We call this surface an *identity surface*. Face recognition can then be performed by computing and comparing the distances between a given pattern and a set of *identity surfaces*.

3.1 Synthesising Identity Surfaces of Faces

If sufficient patterns of a face in different views are available, the *identity surface* of this face can be constructed precisely. However, we do not presume such a strict condition. In this work, we develop a method to synthesise the *identity surface* of a subject 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, $\mathbf{z}$ is the discriminant feature vector of a face pattern, e.g. the KDA [1] vector. $(x_{01}, y_{01}), (x_{02}, y_{02}), ..., (x_{0N_v}, y_{0N_v})$ are predefined views which separate the view plane into N_p pieces. On each of these N_p pieces, the *identity surface* is approximated by a plane

$$\mathbf{z} = \mathbf{a}x + \mathbf{b}y + \mathbf{c} \tag{1}$$

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

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

$$\begin{aligned} \text{subject to} \quad : \quad & \mathbf{a}_i x_{0k} + \mathbf{b}_i y_{0k} + \mathbf{c}_i = \mathbf{a}_j x_{0k} + \mathbf{b}_j y_{0k} + \mathbf{c}_j \\ & k = 0, 1, ..., N_v, \\ & \text{planes } i, j \text{ intersect at } (x_{0k}, y_{0k}). \end{aligned} \tag{3}$$

This is a quadratic optimisation problem which can be solved using the interior point method [18].

For an unknown face pattern $(x, y, \mathbf{z_0})$ where $\mathbf{z_0}$ is the KDA vector and x, y are the pose in tilt and yaw, one can compute the distance to one of the *identity surfaces* as the Euclidean distance between $\mathbf{z_0}$ and the corresponding point on the *identity surface* $\mathbf{z}$

$$d = \|\mathbf{z_0} - \mathbf{z}\| \tag{4}$$

where $\mathbf{z}$ is given by (1).

[1] Kernel Discriminant Analysis, a nonlinear approach to maximise the inter-class variation and minimise the within-class variation [11].

3.2 Video-Based Online Face Recognition

When a face is tracked continuously in a video sequence, not only additional information in different views and from multiple frames is obtained, but also the underlying dynamic characteristics of the face can be captured by the spatio-temporal continuity, variation, and distribution of face patterns.

For computer based vision systems, formulating and modelling the psychological and physiological dynamics described in Section 1 is still an unsolved problem. However, significant improvement in terms of recognition accuracy and robustness may still be achieved when the spatio-temporal information is modelled in a rather straightforward way, e.g. simply accumulating the discriminant evidence with the spatio-temporal order encoded in an input sequence. Based on this idea, we formulate the following approach to video-based face recognition.

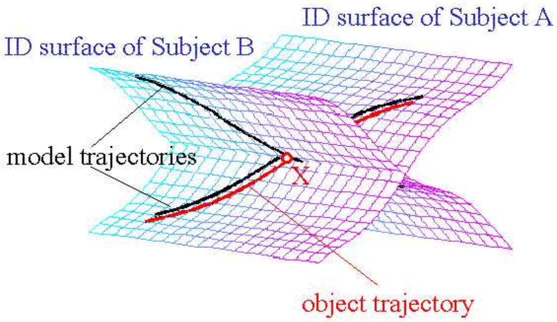

Figure 3. Identity surfaces for face recognition. When a face is tracked continuously from a video input, a robust recognition can be performed by matching the object trajectory to a set of model trajectories.

As shown in Figure 3, 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 on each of the *identity surface* with the same poses and temporal order forms a *model trajectory* of the specific subject. It can be regarded as the ideal trajectory of this subject 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 robust and accurate. For example, it is difficult to decide whether the pattern X in Figure 3 belongs to subject A or B for a single pattern, however, if we know that X is tracked along the object trajectory, it is much clear that it is more likely to be subject A than B.

3.3 Constructing Identity Surfaces from Learning Sequences

Before recognition is carried out, a face class should be registered to the system by one or more learning sequences containing the face of the subject. For example, we can record a small video clip of this 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. Then these patterns are stored to construct the *identity surface* of this subject and, if necessary, to train (or re-train) the KDA.

To simplify computation, normally we do not use all the patterns of each subject to train the KDA since the size of the kernel matrix is directly related to the number of training examples. A pragmatic way to select the KDA training patterns is to factor-sample the patterns from the training sequences so that the result patterns uniformly cover the view sphere.

After the KDA training, all face patterns can be projected onto the feature space spanned by the significant KDA base vectors. Then the method described in Section 3.1 can be employed to construct the *identity surfaces*.

3.4 Object Trajectory and Model Trajectories

In the recognition stage, we apply the same multi-view dynamic face model on a novel sequence containing faces to be recognised, then an object trajectory can be obtained by projecting the face patterns into the KDA feature space. On the other hand, according to the pose information of the face patterns, it is easy to build the model trajectory on the *identity surface* of each subject using the same pose information and temporal order of the object trajectory. Those two kinds of trajectories, i.e. object and model trajectories, encode the spatio-temporal information of the tracked face. And finally, the recognition problem can be solved by matching the object trajectory to a set of identity model trajectories.

A preliminary realisation of this approach is implemented by computing the trajectory distances, i.e. the weighted sum of the distances between patterns on the object trajectory and their corresponding patterns on the model trajectories. At frame t of a sequence, we define the distance between the object trajectory and an identity model trajectory m as:

$$d_m = \sum_{i=1}^{t} w_i d_{mi} \qquad (5)$$

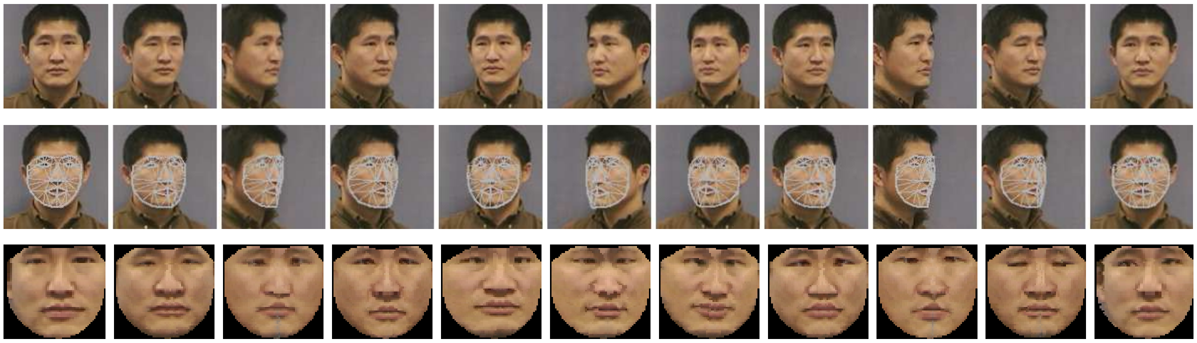

Figure 4. Video-base multi-view face recognition. From top to bottom, sample images from a test sequence with an interval of 10 frames, images fitted by the multi-view dynamic face model, and the shape-and-pose-free texture patterns.

where d_{mi}, the pattern distance between the face pattern captured in the ith frame and the *identity surface* of the mth subject, is computed from (4), and w_i is the weight on this distance. Considerations on determining w_i include: confidence of model fitting, variation from the previous frame, and view of the face pattern (e.g. profile face patterns are weighted lower since they carry less discriminant information). Finally, recognition can be performed by:

$$id = argmin_{m=1}^{M} d_m \tag{6}$$

3.5 Experiments

We demonstrate the performance of this approach on a small scale multi-view face recognition problem. Twelve sequences, each 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. Only 10 KDA dimensions were used to construct the *identity surfaces*. Then recognition was performed on new test sequences of these subjects. Figure 4 shows the sample images fitted by our multi-view dynamic model and the warped *shape-and-pose-free* texture patterns from a test sequence. The object and model trajectories (in the first two KDA dimensions) are shown in Figure 5. The pattern distances from the *identity surfaces* in each individual frame are shown in the left side of Figure 6, while the trajectory distances shown in the right side. These results depict that a more robust performance is achieved when recognition is carried out using the trajectory distances which include the accumulated evidence over time though the pattern distances to the *identity surfaces* in each individual frame already provides a sufficient recognition accuracy.

4 Conclusions

Most of the previous research in face recognition is mainly on frontal-view (or near frontal-view) and from static images. In this work, an approach to video-based online face recognition is presented. The contributions of this work include:

1. A multi-view dynamical face model is adopted to extract the *shape-and-pose-free* texture patterns of faces from a video input. The model provides a precise correspondence for recognition since 3D shape information is used to warp the texture patterns to the model mean shape in frontal view.

2. Instead of matching templates or estimating multimodal density, the *identity surfaces* of face classes are constructed in a discriminant feature space. Then recognition is performed by computing the pattern distances to the *identity surfaces*.

3. A more robust performance can be achieved by performing recognition dynamically on video sequences. The *identity surfaces* can be constructed from learning sequences. Trajectory distances between the object and model trajectories, which encode the spatio-temporal information of a moving face, is computed for recognition.

For visual interaction and human-computer interaction, the problem of face recognition involves more than matching static images. At a low-level, the face dynamics should be accommodated in a consistent spatio-temporal context where the underlying variations with respect to changes in identity, view, scale, position, illumination, and occlusion are integrated together. At a higher level, more sophisticated behaviour models, including individual-dependent

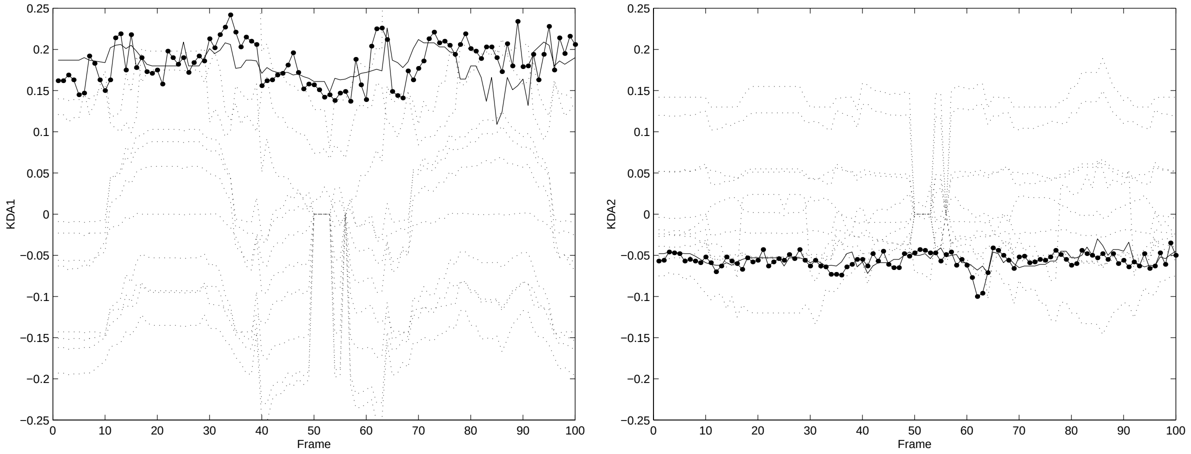

Figure 5. The object and model trajectories in the first two KDA dimensions. The object trajectories are the solid lines with dots denoting the face patterns in each frame. The others are model trajectories where the ones from the ground-truth subject highlighted with solid lines.

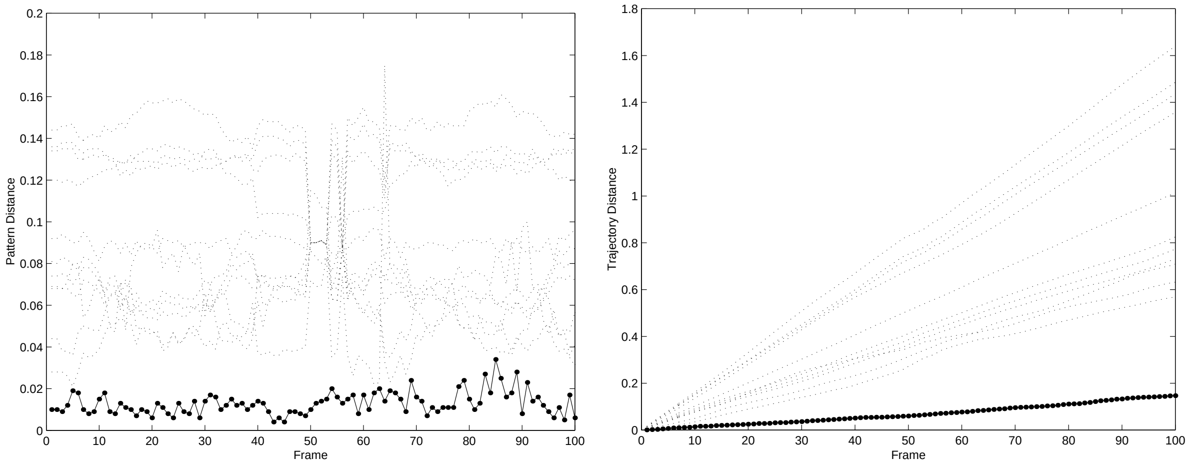

Figure 6. Pattern distances and trajectory distances. The ground-truth subject is highlighted with solid lines. By using KDA and identity surfaces, the pattern distances can already give an accurate result. However, the trajectory distances provide a more robust performance, especially its accumulated effects (i.e. discriminative ability) over time.

and individual-independent models, may supervise and cooperate with all the low level modules.

In this paper, we highlighted the nature of the problem and showed the potential of modelling face dynamics as an effective means to face recognition. Nevertheless, some of the implementation such as the trajectory distance is still simplistic in its present form and rather ad hoc. Future work using more sophisticated temporal models to capture face dynamics with respect to subject, expression, movement and illumination changes is to be extensively conducted.

References

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

[2] V. Bruce, P. Hancock, and A.Burton. Human face perception and identification. In Wechsler, Philips, Bruce, Fogelman-Soulie, and Huang, editors, *Face Recognition: From Theory to Applications*, pages 51–72. Springer-Verlag, 1998.

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

[4] 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.

[5] 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.

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

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

[8] 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.

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

[10] 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.

[11] Y. Li, S. Gong, and H. Liddell. Constructing structures of facial identities using Kernel Discriminant Analysis. Technical report, Queen Mary, University of London, 2001. www.dcs.qmw.ac.uk/∼yongmin/papers/kda.ps.gz, submitted to IJCV.

[12] Y. Li, S. Gong, and H. Liddell. Modelling faces dynamically across views and over time. In *IEEE International Conference on Computer Vision*, Vancouver, Canada, July 2001.

[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] S. Romdhani, S. Gong, and A. Psarrou. On utilising template and feature-based correspondence in multi-view appearance models. In *European Conference on Computer Vision*, volume 1, pages 799–813, Dublin, Ireland, June 2000.

[15] S. Satoh. Comparative evaluation of face sequence matching for content-based video access. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 163–168, Grenoble, France, 2000.

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

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

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

[19] L. Wiskott, J. Fellous, N. Kruger, and C. Malsburg. Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):775–779, 1997.

[20] 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.