Exploiting the Dynamics of Faces in Spatial-temporal Context

Yongmin Li, Shaogang Gong and Heather Liddell
Department of Computer Science, Queen Mary and Westfield College, University of London
London, E1 1NS, UK {yongmin, sgg, heather}@dcs.qmw.ac.uk

Abstract

We present an approach to dynamically recognise faces in a spatial-temporal context from video sequences. Linear Discriminant Analysis (LDA), head pose estimation, and multi-view face detection are integrated together to accomplish the task of multi-view face recognition. By synthesizing virtual views, our approach can work well when only a few views which sparsely cover the view sphere are available. We model the dynamics of faces by constructing the object trajectory and a set of identity model trajectories in the LDA feature space. Face recognition is performed dynamically by matching those trajectories. Compared with the static face matching methods, this approach is more robust and accurate under a coarse correspondence of face images, and has potential to visual interaction and advanced human behaviour recognition in real-world scenarios.

1 Introduction

During the past decade, the issue of face recognition has been extensively addressed. The eigenface approach proposed in [18] uses Principal Component Analysis (PCA) to code face images and capture face features. This approach has then been extended to view-based and modular eigenspaces intended for recognising faces under varying views[13]. In [21], face recognition is performed by Elastic Graph matching based on a Gabor wavelet transform. An alternative approach uses similarity vectors to estimate head pose and recognise faces across views [8]. The Active Shape Model (ASM) and Active Appearance Model (AAM) capturing both shape and shape-free grey-level appearance of face images have been successfully applied to face modelling and recognition [2, 3, 5]. Both ASM and AAM have been extended to nonlinear cases across views based on Kernel Principal Component Analysis (KPCA)[14, 16, 15]. These nonlinear models are designed to correspond dynamic appearances of both shape and texture across views.

In most of the previous work, the basic methodology adopted for recognition is largely based on matching static face image patterns in a given feature space. More recently, there has been some work on face recognition using video sequences [9, 6, 22, 17]. Nevertheless, the issue of recognising human faces dynamically in a spatio-temporal context remains largely unresolved.

In this paper, we describe an integrated approach to face detection and recognition with a focus on modelling and recognising faces dynamically in a spatio-temporal context. First, an overall framework is briefly introduced in Section 2. Section 3 describes the issue of multi-view face recognition in static images. Then in Section 4 we extend the research to recognising human faces dynamically in a spatial-temporal context. Conclusions are drawn in Section 5.

2 A Framework for Multi-view Face Detection, Tracking and Recognition

We aim to model the dynamics of human faces undergoing rotations in depth, changes in scale, transformation in position, and variation to identities. The input to our system is adopted as live video sequences of captured but unsegmented faces that vary continuously in a temporal context. At the initializing phase, the system performs the follow tasks:

1. segmentation: use motion estimation, skin colour detection, and background subtraction to segment sub-images containing faces.
2. scanning: scan the sub-images with different scales.
3. pose estimation: for each scan, estimate the likely "pose" in tilt and yaw of the image patch.
4. face detection: choose a suitable face detector from a set of multi-view face detectors according to the estimated "pose" to determine whether the pattern is a face. If the output of the face detector is above a preset threshold, then a face is detected, and the position, scale and pose of the detected face are used for recognition. Otherwise, the patch is rejected as a non-face pattern.
5. face recognition: combine the detected face pattern, its pose estimate, and, if available, the history information from the previous frames to recognise the face dynamically.

After a face is successfully detected, it is not necessary to repeat this whole procedure again. For example, a Kalman filter can be used to predict the position, scale, and pose of the detected face in the successive frame. When the tracking fails, the system can be restored by re-initializing as described above. The predicted pose can be directly used, while face detection is still carried out on a slightly enlarged sub-image around the predicted position using exhaustive scanning for a robust performance. However, the size of this sub-image can be smaller than that segmented in the initializing stage. Therefore, face detection is the most computational intensive part in the system. An efficient algorithm using Support Vector Machines [19] and eigenspace methods [13] has been designed to solve this problem [12].

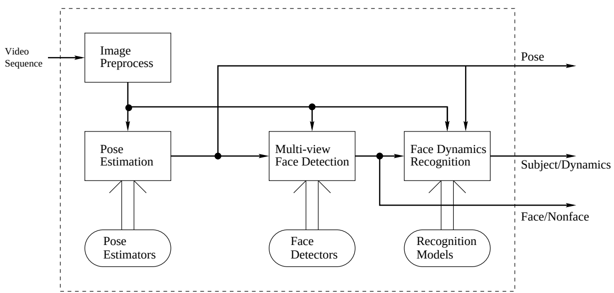


Figure 1. *System framework.*

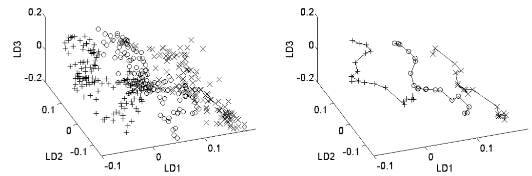
It is worth noticing that pose estimation is performed on multi-scale image patches before detection. If an image patch is detected as a face, then the tilt and yaw angles of the face are obtained, and the face detectors may give a positive output for the pattern. However, when the image patch is not exactly a face, the estimated pose is just a meaningless value. At the detection stage, one of the detectors will give a negative output to reject it as non-face. It seems that extra computation is exerted on non-face patterns at the stage of pose estimation. However, more saving in computation is achieved at the stage of multi-view face detection as only one detector is chosen for the given pose [11]. The proposed framework for dynamic face detection and recognition is illustrated in Figure 1.

3 Recognising Multi-view Faces in Static Images

Recognising faces across multiple views is one of the most challenging problems in this research area, especially for 2D appearance based methods. In this section, we discuss this problem in the domain of static image matching. We adopt Linear Discriminant Analysis (LDA) as the representation. The key points of our approach are combining pose information and LDA features for recognition, and synthesizing virtual views from a sparse set of available views.

3.1 Linear Discriminant Analysis

LDA seeks to find a linear transformation from input space to a feature space (LDA space) by maximizing the between-class variance and minimizing the within-class variance at the same time. Computationally, LDA is similar to eigen-decomposition [7]. Since it is quite effective to select discriminant features for two or more object classes, LDA has been widely adopted in many pattern recognition applications. Zhao et al. [23] used LDA as a representation for frontal-view face recognition. Edwards et al. [4] adopted LDA to select *Discriminant Parameters* based on Active Appearance Models.



(a) (b)

Figure 2. *Distribution of multi-view face patterns in the LDA space spanned by the first 3 LDA features. (a) Patterns from 3 subjects, 133 views of each. (b) Selected patterns by pose from the left figure, where yaw changes from 0° to 180° and tilt is 90° for all patterns.*

In this paper, we adopt a similar approach to [23] for the representation of face images. However we extend the problem to multi-view face recognition. Intuitively, the image appearance of different people at the same view is more similar than that of the same person at different views. This makes LDA less effective when applied to multi-views. This is illustrated in Figure 2(a), where the distribution of face patterns from 3 subjects, 133 views of each, is shown in the space spanned by the first 3 significant LDA features. One can observe that the variance from different subjects is not significantly isolated from the variance from different views, which suggests that LDA alone is insufficient for multi-view face recognition.

However, if we consider the pose information, the problem becomes different. Figure 2(b) shows the selected patterns from Figure 2(a) where tilt is 90° and yaw changes from 0° to 180° . We form the trajectories of the patterns from the same subjects. One notices that, even under a very low dimensional feature space (only the first 3 LDA features used), the identities of different subjects captured by their trajectories are separable, therefore can be discriminated.

3.2 Synthesizing Virtual Views for Multi-view Face Recognition

If all views of subjects are available, the problem of face recognition can be simply matching the face patterns on a specific view. However, this assumption may be too strict in many situations. That leads the following question: Given a few views of a subject, can we still perform recognition on unknown views? One solution is to synthesize virtual views in some feature space based on face patterns available, and then to match the unseen face pattern to the synthesized patterns. Vetter and Poggio [20] used Linear Object Classes for this purpose. In their work, the synthesis of virtual views was based on a dense correspondence between 2D face images of different people at the same view. Alternatively similarity vectors to a set of prototypes were adopted to construct the feature space, where face recognition on novel views was performed by linear interpolation of similarity vectors of patterns on available views [8].

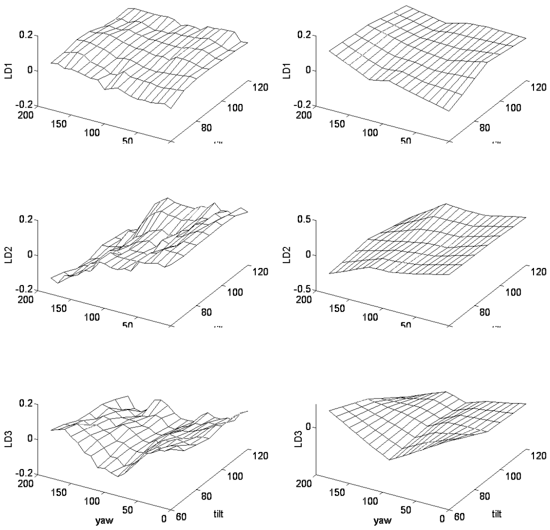


Figure 3. *Comparison between original face patterns and synthesized patterns from 15 views. The three axes are tilt, yaw, and the first 3 LDA features respectively.*

In this paper, we present an approach to synthesize virtual views in LDA space. The problem can be formulated as follows: Given a sparse set of face patterns $P_1, P_2, \cdots, P_m$ of an object on prototype views $v_1, v_2, \cdots, v_m$, one approximates the distributions of this object in feature space $P(v)$ by the patterns available. Figure 3 shows real distributions of face patterns from one subject with respect to different views and the synthesized distribution from 15 out of 133 views. The views are defined by tilt and yaw angles.

If more prototype views are available, a high order approximation of $P(v)$ is preferable for high accuracy. However, if only a few prototype views are available, one can still perform the synthesis by simple linear techniques such as a bilinear interpolation which approximates a novel view by the four nearest prototype views, or by a triangle interpolation which uses the 3 nearest prototype views. In the experiments presented in this paper, the bilinear interpolation was employed to synthesize virtual views.

In our system, after face detection, the position, scale, and pose both in tilt and yaw of the detected face are obtained. Then for each subject to be recognised, a synthesized pattern in the detected view is generated in the feature space from the prototype views of the corresponding subject. Finally recognition is performed by matching the detected pattern with all the synthesized patterns.

n	tilt(°)	yaw(°)	rate(%)	known(%)	novel(%)
4	80,100	40,140	68.54	74.15	62.92
6	80,100	40,90,140	82.92	87.08	78.77
9	70,90,110	30,90,150	90.00	94.62	85.38
15	70,90,110	30,60,90,120,150	94.31	96.62	92.00

Table 1. *Recognition results of using synthesized patterns in LDA feature space. n: number of prototype patterns; tilt/yaw: view angles of prototype patterns; rate: overall recognition accuracy; known/novel: accuracy on known/novel subjects respectively.*

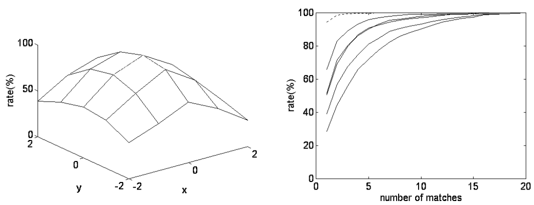
We trained the LDA on a set of face images from 10 subjects, 133 views of each (-90° -90° in yaw and -30° -30° in tilt with an interval of 10°). The training images were grouped by subjects, hence 10 LDA basis vectors were received after training. The test set included face images from 20 subjects, where 10 subjects did not appear in the training set. Table 1 lists the results of recognition using different selections of prototype views. In this experiment, we did not perform pose estimation on the test images, instead the ground-truth pose information was used for synthesizing virtual views. The results illustrate that given accurate pose estimation and face detection, a small number of prototype views are sufficient to synthesize novel views for recognition. Also, the recognition accuracy on novel subjects is only slightly lower than that on known subjects.

4 Recognising Faces Dynamically from Video Sequences

Matching static images is one of the fundamental issues in face recognition, which can still find many applications in real-life. However, in a situation where the appearance of faces changes spatially and temporally, recognition by matching static patterns on a set of isolated and independent pieces of information will be inefficient as it is not robust enough against misalignment, occlusion, and changes of pose, illumination, and expression. In this section, we analysis the recognition sensitivity to two major factors: image alignment and pose estimation, then propose an approach to address the problem of recognising faces dynamically by using the spatial-temporal information from video sequences.

4.1 Recognition Sensitivity to Image Alignment and Pose Estimation

It is crucial to point out that *no explicit alignment is performed in our approach*. In fact, face detection takes the role of coarse correspondence based on the appearance of image patterns. Compared with other methods such as optical flow or Active Shape Models where the correspondence between images is aimed to be well established either densely or sparsely, this method is only based on registering the holistic patterns approximately. Also, the synthesis of virtual views is based on pose estimation where the average error in our system is around 10° in both tilt and yaw. This error is another factor which influences the accuracy of recognition.



(a) (b)

Figure 4. *Sensitivity analysis with respect to position shift of detected faces. (a) Recognition rate with respect to the shift along horizontal/vertical direction of images. (b) Recognition rate of subjects being recognised in the top n matches. Only 6 curves plotted here. The dotted curve is from the results of the accurate position. The corresponding shifts for the other 5 curves are -2 in x, and -2 to 2 in y. In this experiment, 15 prototype views as shown in Table 1 were used, and the face image size is 20x20.*

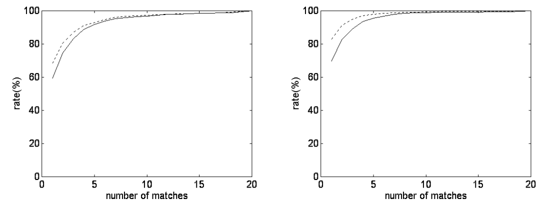
We performed sensitivity analysis based on two factors: Sensitivity to the image shift off the centre, and sensitivity to pose estimate. Figure 4 shows the recognition rates of the top match (the nearest distance) when shifting the ground-truth positions of faces around centres and the recognition rates obtained from the top n matches (the first n nearest distances). Figure 5 gives the recognition results when all the view information is estimated by our pose estimators and compared to the results on ground-truth pose information.

In Figure 4 and Figure 5, we note:

1. the error both in position alignment and view estimation have significant effects on the top recognition match *alone*;
2. most of the patterns that were not correctly recognised by the top match were correctly recognised within the top 3-5 matches.

If we improve the accuracy in both detection and pose estimation, the performance would be improved to some extent. However, this assumption may be too strict in many situations. An alternative solution is to incorporate explicit shape models into the system for an accurate correspondence, but usually this approach is expensive and can also be too restrictive in real-world dynamic scenes.

If recognition is performed on a continuously moving face by accumulated identification evidence rather than snapshots of the face independently, the requirement for alignment and pose estimation can be relaxed. Based on this, we present an approach to address the problem by using the spatial-temporal information from video sequences.



(a) (b)

Figure 5. *Sensitivity analysis to the error in pose estimation. The two figures show the results when 4 and 6 prototype views were used for recognition respectively (see Table 1). The horizontal axis stands for the number of top matches, and vertical axis for recognition rate. Solid curves are obtained from estimated pose, and dotted curves from ground-truth pose.*

4.2 Face Recognition by Matching Trajectories in Feature Space

In many real-world applications, a sequence of images containing the subjects to be recognised are necessarily acquired. With a sequence which records a face varying continuously over time and across views, not only can more information be obtained, but also the dynamics of faces can be captured [10, 1].

Yamaguchi et al. [22] 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. [9] introduced an approach that uses Partially Recurrent Neural Networks (PRNNs) to recognise temporal signatures of faces.

We present here an approach to recognising human faces dynamically by matching trajectories in LDA feature space. For a given sequence containing faces to be recognised, after head pose estimation and face detection, one can obtain a trajectory by projecting the face patterns into LDA space. On the other hand, according to the pose information of the face patterns, it is easy to build the identity model trajectory for each subject using the known prototype patterns. Therefore, the recognition problem can be solved by matching the object trajectory to a set of identity model trajectories.

At frame t of a sequence, we define the distance between the object trajectory and a identity model trajectory

m as:

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

where d_{mi} is the distance in feature space between model point and object point at frame i, and w_i is the weight on this distance. Considerations on determining w_i include:

1. confidence of face detection;
2. variation from the previous frame;
3. view of the face pattern. For example, profile face patterns are weighted lower than those at 3/4 views since they carry less discriminating information.

Finally, result of recognition can be given as:

$$id = argmax_{m=1}^{M} d_m \quad (2)$$

Figure 6 illustrates an example of recognising moving faces dynamically. Recognition through matching static face patterns in individual frame can lead to wrong results. However, if recognition is performed by matching the accumulated trajectories, a more robust and accurate recognition can be achieved. In particular, it is important to observe in Figure 6(c,d) that the true identity of a facial appearance is often recognised among the best few matches although it may not be consistently the best match in every frame over time. In other word, the accumulation of positive identity information will overwhelm any misidentification over time if recognition is performed on accumulated evidence.

5 Conclusions

From the viewpoint of visual interaction and human-computer interaction, the problem of face recognition involves more than matching static images. At a low-level, the face dynamics can 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 and individual-independent models, may supervise and co-operate with all the low level modules.

In this work, we have presented the following key issues to address the problem:

1. integrating head pose estimation, multi-view face detection and LDA to recognise faces dynamically;
2. synthesizing virtual views in LDA feature space, which are used to construct the identity model trajectories, from a sparse set of views available;
3. recognising faces dynamically by matching a object trajectory to the model trajectories.

Establishing correspondence between faces across views and from different subjects is one of the fundamental problems in face recognition. Sophisticated shape models may solve the problem accurately. However, most shape models are computationally expensive. Here we try to solve the problem alternatively based on *coarse registration*.

Another problem is how many known views of each subject are required. We do not constrain ourselves to the extent that all views of a subject should be available. In fact, by synthesizing virtual views, only a few views are sufficient for capturing the whole view sphere. Furthermore, those prototype views need not be fixed, a randomly sampled sequence clip can be conveniently used for modelling.

We believe that exploiting the dynamics of human faces is the most promising issue in face recognition. It is true that measuring the distance between the object and model trajectories is only a simple implementation of the approach. More elaborated methods, such as matching the temporal variation of the trajectories using higher order modes other than only positions in feature space, can be more effective. In this paper, we highlighted the nature of the problem and shown the potential of modelling faces in a spatial-temporal manner to the problem of face recognition. Nevertheless, we only touched the surface of the problem by modelling faces dynamically across views and between subjects in an LDA feature space. Substantial future work using sophisticated spatial-temporal models to capture face dynamics with respect to expression, movement and illumination changes is required.

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

[3] T. Cootes and C. Taylor. A mixture model for representing shape variation. *Image and Vision Computing*, 17:567–573, 1999.

[4] G. Edwards, A. Lanitis, C. Taylor, and T. Cootes. Statistical models of face images - improving specifi city. In *British Machine Vision Conference*, volume 2, pages 765–774, Edinburgh, Scotland, 1996.

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

[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] K. Fukunaga. *Introduction to statistical pattern recognitiion*. Academic Press, 1972.

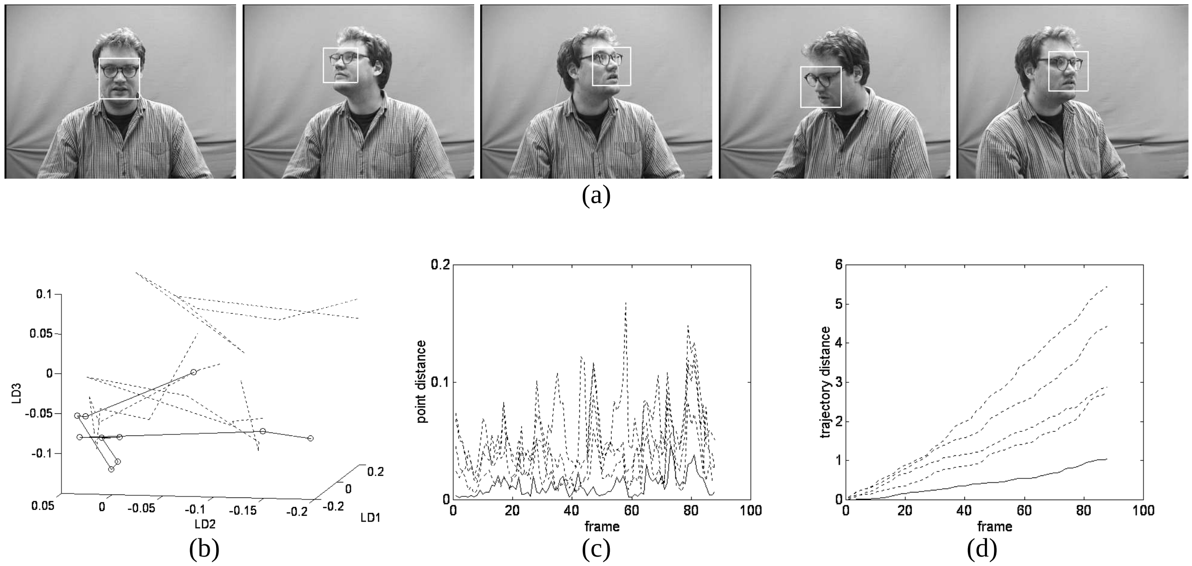


Figure 6. *Recognising faces dynamically by matching trajectories in LDA feature space. (a) Successive frames from a test sequence where detected faces are marked with white boxes. (b) Object and identity model trajectories. The object trajectory is shown by solid curve with frames labelled by small circles. The others are model trajectories. For clarify, only 3 out of 20 model trajectories are illustrated here. (c) Distance measured independently in each frame between object pattern and model patterns. (d) Distance between object and different model trajectories over time. Results on the first 5 subjects are illustrated here, where the solid line is from the true subject.*

[8] S. Gong, E.-J. Ong, and S. McKenna. Learning to associate faces across views in vector space of similarities to prototypes. In *British Machine Vision Conference*, pages 54–64, Southampton, England, 1998.

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

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

[11] 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, 2000.

[12] Y. Li, S. Gong, J. Sherrah, and H. Liddell. Multi-view face detection using support vector machines and eigenspace modelling. In *Fourth International Conference on Knowledge-Based Intelligent Engineering System & Allied Technologies*, pages 241–244, Brighton, UK, 2000.

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

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

[15] S. Romdhani, S. Gong, and A. Psarrow. On utilising template and feature-based correspondence in multi-view appearance models. In *European Conference on Computer Vision*, Dublin, Ireland, June 2000.

[16] S. Romdhani, S. Gong, and A. Psarrow. A generic face appearance model of shape and texture under very large pose variations from profile to profile views. In *International Conference on Pattern Recognition*, Barcelona, Spain, September 2000.

[17] 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.

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

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

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

[21] 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.

[22] 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.

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