Learning to Associate Faces across Views in Vector Space of Similarities to Prototypes*

Shaogang Gong†, Eng-Jon Ong† and Stephen McKenna‡

† Dept of Computer Science, Queen Mary and Westfield College,
London E1 4NS, England. sgg@dcs.qmw.ac.uk

‡ Dept of Applied Computing, University of Dundee,
Dundee DD1 4HN, Scotland.

Abstract

We present a method for learning appearance models that can be used to recognise and track both 3D head pose and identities of novel subjects with continuous head movement across the view-sphere. We describe an automatic face data acquisition system based on a magnetic sensor and a calibrated camera. The system enabled us to obtain systematically a database of face images with labelled 3D poses across a view-sphere of $\pm 90^\circ$ yaw and $\pm 30^\circ$ tilt at intervals of 10°. The database was used to learn appearance models of unseen faces based on similarity measures to prototype faces. The method is computationally efficient and enables real-time performance.

1 Introduction

To be able to recognise faces of moving people not only requires the ability to label novel face images with known identities, but also needs detecting and tracking of faces over time [1]. We refer to this as the task of *associating faces*. We adopt the view such a task can be better achieved using view-based appearance models rather than explicit 3D models [2]. One of the difficulties in associating faces using view-based representations is that face images of the same person from different viewpoints are significantly more dissimilar than images of different people appearing in the same view. However, the task can be significantly simplified if poses are known [3]. The ability to estimate and predict the 3D orientation of faces and the ways in which they change over time also imposes temporal continuity in recognition. Consequently, the ability to locate, track and predict head pose of a moving person is an integral part of recognition. Here we present a method for learning to associate faces across the view-sphere based on similarity measures to prototypes in multiple views. Although similar work was proposed for recognition using similarity measures [4], and for novel view generalisation and synthesis using linear combination of prototypes [5], this work extends the idea to a unified method that addresses the problems of both pose and identity recognition and tracking over time. The method uses training

* E.-J. Ong and S. McKenna were funded by EPSRC Grant No. GR/K44657.

BMVC 1998 doi:10.5244/C.12.6

data from a database of 3D pose labelled face images across the view-sphere captured by an automated data acquisition system.

The problems addressed here are (1) automatic acquisition of labelled face data across the view-sphere for learning, (2) real-time recognition and tracking of face image location, scale, and pose relative to the image-plane and (3) identities over time. For simplicity, rotation in the image-plane is ignored, i.e. the subject is assumed to be upright. The pose refers to the rotation in depth relative to the image axes. A "nodding" head undergoes x-axis rotation whilst a "shaking" head undergoes y-axis rotation. The former is also referred to as "tilt" and the latter as "yaw".

The proposed method takes a view-based approach in which face appearance models are learned from example views without recourse to any explicit 3D model. Furthermore, the views are aligned using only simple image-plane transformations such as translation and scaling, or at most affine transformation. In particular, no dense correspondences between feature points on different faces are required and as a result, real-time performance is obtained. The models are constructed from aligned image data labelled with pose angles. Efficient focus of attention based on colour and motion cues is used to bootstrap face image search in the image plane [6, 7].

2 Acquisition of Labelled Views across the View-sphere

In order to build appearance models, example views labelled with 3D pose angles (both tilt and yaw) are required. A system was designed that utilises both a magnetic sensor attached to the subject's head and a camera calibrated relative to the sensor's transmitter. The sensor was then used to provide pose labels for the face images of the subject captured by the camera. Figure 1 shows the acquisition system.

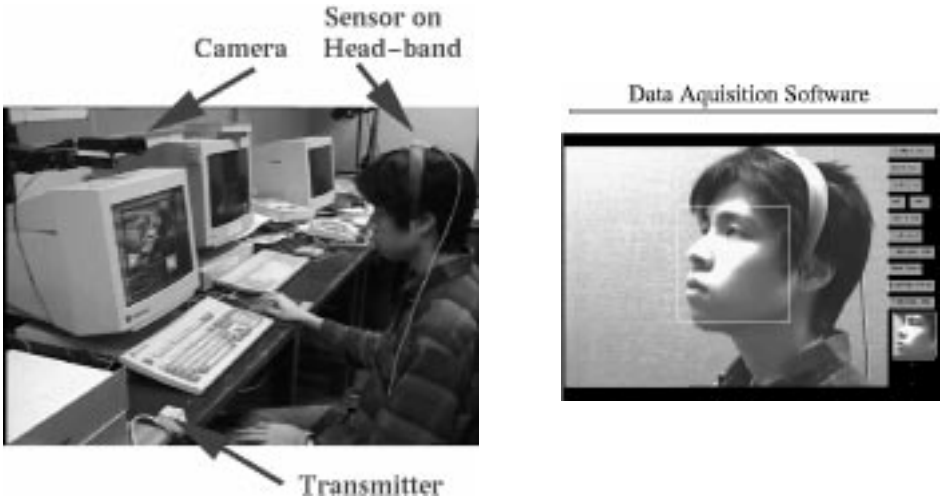

Figure 1: The system for acquiring labelled views across the view-sphere.

More precisely, an electromagnetic 6 DOF Polhemous tracker with a sensor and a transmitter was used to provide 3D coordinates (in cm) and orientation of the sensor rel-

Figure 2: An example labelled head image set. The images of labelled views are from +90° to -90° in yaw and from +30° to -30° in tilt at 10° intervals.

ative to the transmitter. The tilt, yaw and roll correspond to rotations about the x, y and z axes respectively and are Euler angles (in degrees). The sensor is rigidly attached to a head-band worn by the user so that it follows the head's movements and changes in orientation. The image acquisition system used has a single camera which has been calibrated to the sensor's coordinate system. The location and size of the head in the image are determined by back-projection onto the image-plane and an appropriately cropped image is thus acquired. The sensor orientation is used to label the image with head pose. In order to locate and align the 2D head images, camera calibration with respect to the sensor is needed. This involves determining camera parameters using the 3D positions provided by the sensor and their corresponding 2D projections on the camera's image-plane. Both intrinsic and extrinsic parameters were estimated. The intrinsic parameters are focal length and radial distortion. The extrinsic parameters are the position and orientation of the camera relative to the sensor's coordinates. We adopted the camera model used by [8].

The position of the sensor with respect to the head is somewhat arbitrary. However, the position and scale of the heads in the images acquired need to be consistent across different people. Therefore, a few facial features were manually located for each subject in order to bootstrap the acquisition process by determining a scaling factor and a 3D point inside the head. This point was rigidly "attached" to the facial features (eyes and upper lip) and was used to project onto the centre of the acquired head images. In other words, the facial features' 3D coordinates were used to determine the coordinates of a 3D point inside the head relative to the sensor's 3D coordinates. The image was then cropped as determined by the scale factor and re-sampled to a fixed number of pixels. Labelled images were captured with y-axis rotation in the range $\pm 90^\circ$ and x-axis rotation in the range $\pm 30^\circ$ at intervals of 10°. Examples for one subject can be seen in Figure 2.

3 View-based Face Appearance Models using Prototypes

Face appearance models are essentially view-based holistic templates. A simple way to obtain a generic appearance model is to estimate an average face template at each pose. These mean templates could conceivably be used to associate face images in order to recognise and track poses of faces across viewpoints. Figure 3 shows some of

the mean templates computed by averaging filtered views of 11 different subjects. However, although these view-based mean templates can be used to perform reasonably well in recognising and tracking poses, they are sensitive to illumination changes and image noise. Furthermore, they do not capture identity information. More elaborated appearance models use linear combinations of training samples. Given sufficient data, such linear combinations can also be statistical. This includes the use of PCA [9], LDA [10] and Gaussian mixtures [11].

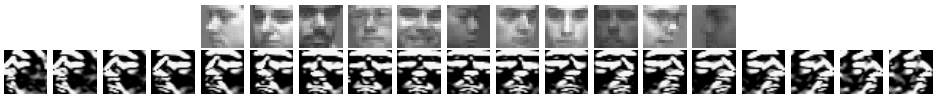

Figure 3: Average templates for views from profile to profile.

In order to generalise between views, rather than assuming that face appearances are linear combination of prototypes at given views [5], an image can be represented as a vector of similarities to prototype views [12]. Here we exploit this approach to both face pose tracking and view-based recognition. Let a face image $\mathbf{x}$ at a given pose be represented as a vector $\boldsymbol{\alpha}$ of similarities to q prototype faces $\mathbf{y}_i$ at the same pose as follows:

$$\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_q], \qquad \alpha_i = h(\mathbf{x}, \mathbf{y}_i) \tag{1}$$

where $i = 1, \ldots, q$ and $h()$ is a similarity function that defines a similarity measurement. A straightforward $h()$ can be the inverse Euclidean distance between a face image $\mathbf{x} = [x_1, \ldots, x_N]$ and a prototype $\mathbf{y} = [y_1, \ldots, y_N]$ at a given view, where N is the dimensionality of the images. To take normalisation for overall intensity and contrast into consideration, a better measurement should be the inverse of the Pearson's linear correlation coefficient:

$$h(\mathbf{x}, \mathbf{y}) = \frac{\sqrt{\sum_{i=1}^{N}(x_i - \mu_{\mathbf{x}})^2}\sqrt{\sum_{i=1}^{N}(y_i - \mu_{\mathbf{y}})^2}}{\sum_{i=1}^{N}(x_i - \mu_{\mathbf{x}})(y_i - \mu_{\mathbf{y}})} \tag{2}$$

where $\mu_{\mathbf{x}}$ and $\mu_{\mathbf{y}}$ are the mean of the elements of $\mathbf{x}$ and $\mathbf{y}$ respectively. Furthermore, a deviation weighted distance measure can also be adopted using Gaussian:

$$h(\mathbf{x}, \mathbf{y}) = \exp\left(\frac{-||\mathbf{x} - \mathbf{y}||^2}{2\sigma^2}\right) \tag{3}$$

By measuring similarity vectors of images of novel faces to prototypes across changes in yaw (y-axis rotation), it can be observed that they form separable but also approximately linear manifolds (see Figure 4). The model is therefore useful for recognition. Let us first consider its use in pose recognition and tracking across views.

4 Person-Independent Pose Recognition and Tracking

McKenna and Gong [3] described a real-time system for tracking and estimating head pose based on person-specific templates. A user-specific head model consisting of multiple view templates was used. These templates were filtered with Gabor wavelets or

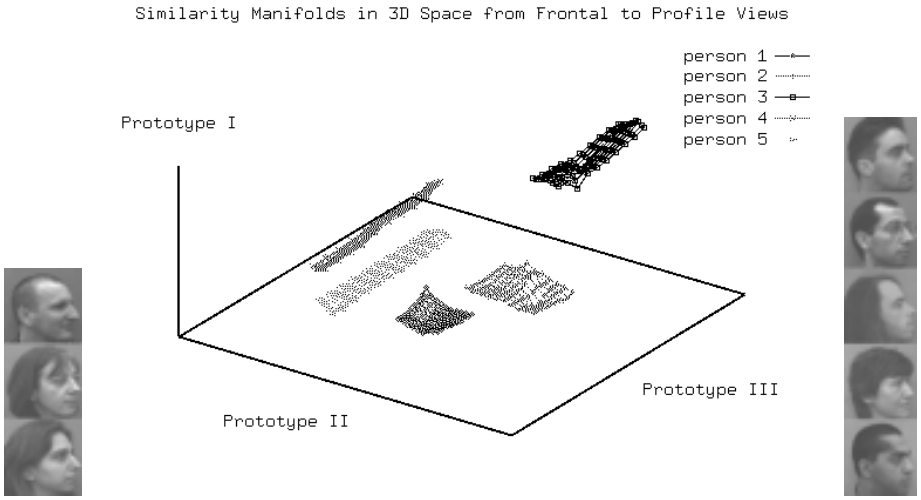

Figure 4: Left: Prototypes at profile view. Centre: Pose manifolds of novel faces in the vector space of similarity to prototypes. Right: Images of novel faces at profile view.

simpler oriented filters in order to obtain a degree of invariance to illumination and to aid the pose estimation (see Figure 5). Each template was labelled with pose angles and in each frame, the best matching template was used to give the pose estimate. Exploitation of temporal constraints allowed a level of performance suitable for driving an avatar in a manner perceptually acceptable to the user. This was achieved in real-time using modest hardware (133MHz P5 with Matrox Meteor board). The performance of such person-specific models was evaluated against the ground-truth measured by the magnetic sensor and it shows acceptable performance, as shown in Table 1. However, it is obviously undesirable to require models for every individual. Here we describe a person-independent pose tracking and prediction system based on similarity to prototypes.

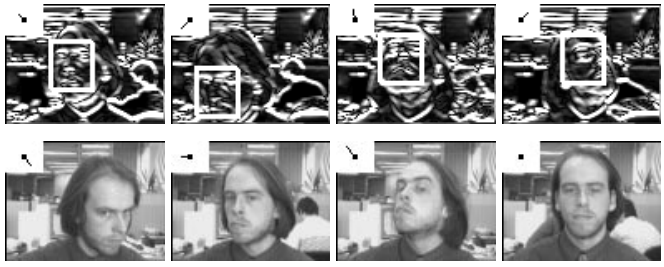

Figure 5: Tracking poses using person-specific templates. The pin diagram indicates estimated pose. Top: the best matching templates' bounding boxes are shown overlaid on a filtered sequence. Bottom: An unfiltered sequence is shown here for visualisation.

Given a database of multiple views of different people, a generic view-based appearance model can conceivably be learned for tracking head pose in a person-independent

manner given sufficient training data. In practice, the number of examples available at each view is small. Alternatively, appearance models based on similarity vectors to a limited number (in tens) of prototype faces at multiple views can be adopted. Given that face images at the frontal view can be readily detected [6], let a similarity vector $\boldsymbol{\alpha}$ to prototypes for a detected face image at the frontal view be measured using Equation (1). Pose recognition and tracking can then be performed by finding the next pose θ (both yaw and tilt) which maximises

$$\mathcal{L}(\theta) = ||\boldsymbol{\alpha}_\theta^t|| + \kappa\, h(\boldsymbol{\alpha}_\theta^t,\, \boldsymbol{\alpha}^{t-1}) \tag{4}$$

where $||\boldsymbol{\alpha}_\theta^t||$ is the L_2 norm of the similarity vector at the most likely pose at time t. Function $h(\boldsymbol{\alpha}_\theta^t,\, \boldsymbol{\alpha}^{t-1})$ is the similarity measure between the two similarity vectors at the previously known pose and the currently likely pose. Maximising $\mathcal{L}(\theta)$ imposes two constraints. The first term maximises the magnitude of similarity regardless identity in a neighbourhood centred at the likely pose at time t, therefore performing a generic face matching at the likely pose at time t. The second term assumes identity constancy in similarity vector space, provided that all other sources of variation such as lighting and translational shift in the images have been eliminated (as shown in Figure 4). The constant κ controls a trade-off between the two factors and its value will depend on the expected smoothness in the pose change and the changes existed in a face's similarity measures to prototypes in different views.

It may also seem to be obvious that the tracked poses (tilt and yaw) can be further Kalman filtered in order to provide a degree of consistency and prediction. However, our experiments show that such attempts have little effect due to the fact that 3D pose change in real-time sequences tends to be highly nonlinear and can vary significantly between frames. Such linear predictive filters are more effective in modelling temporal changes in the image plane [13].

5 Recognising and Tracking Identities across Views

Face recognition of novel identities across pose using similarity to prototypes was proposed by Duvdevani-Bar *et. al.* [4]. The method was based on the assumptions that view-invariant prototypes can be learnt using RBF networks and that the recognition of static faces at novel views can be performed using similarity measures to the view-independent prototypes. We adopt a different approach based on view-specific prototypes and the assumption that similarity measures to prototypes between views vary smoothly (observed in Figure 4). The approach provides a model for recognising and tracking both pose and identity of moving faces.

For a chosen set of q prototypes at a given view θ, let the similarity vectors $\boldsymbol{\alpha}_i^t$ of M different people to the prototypes be their identity measures at view θ. Recognition of a novel face image $\mathbf{x}$ at time t can then be performed by maximising

$$\mathcal{L}(i) = h(\boldsymbol{\alpha}_\mathbf{x},\, \boldsymbol{\alpha}_i^t) + \kappa\,(t-1)\, h(\boldsymbol{\alpha}_i^t,\, \boldsymbol{\alpha}_i^{t-1}) \tag{5}$$

where $i = 1, \ldots, M$, $t \geq 1$ and $h(\boldsymbol{\alpha}_\mathbf{x},\, \boldsymbol{\alpha}_i^t)$ is the similarity measure between the similarity vectors of image $\mathbf{x}$ and a known face i at view θ, given by either Equation (2) or (3). Function $h(\boldsymbol{\alpha}_i^t,\, \boldsymbol{\alpha}_i^{t-1})$ is the similarity measures between similarity vectors of the previously (at $t-1$) and the currently (at pose θ) recognised faces. In other words, whilst

the first term finds the best match to a known face, the second term imposes an identity constancy assumption over time. The constant κ controls a trade-off between the two factors and its value will also depend on the degree of change in a face's similarity measures across views.

Now, if similarity vectors to prototypes are assumed to be invariant across the view-sphere, the recognition could then be generalised to novel views where similarity vectors of known faces are not available, provided that the pose of the current image $\mathbf{x}(t)$ is known. However, our experiments indicate that similarity measures vary across the view-sphere (see Figure 4), although slowly or even linearly. Let us first consider view changes in yaw only. If similarity measures of known faces at two views, θ_1 and θ_2, are available, recognition at novel views θ between θ_1 and θ_2 can then be performed using linear interpolation as follows:

$$\boldsymbol{\alpha}(\theta) = \boldsymbol{\alpha}(\theta_1) + \frac{\theta - \min(\theta_1, \theta_2)}{|\theta_1 - \theta_2|} \left(\boldsymbol{\alpha}(\theta_2) - \boldsymbol{\alpha}(\theta_1)\right), \qquad \theta_1 \leq \theta \leq \theta_2 \tag{6}$$

This interpolation can be easily extended to a 2D surface for both yaw and tilt. If similarity vectors to prototypes at more than two views are known, Equations (5) and (6) can then be used to recognise and track face identities of moving people across the view-sphere.

6 Results and Discussions

In our experiments, a database were acquired having 4450 face images across a view-sphere of $\pm 90^\circ$ yaw and $\pm 30^\circ$ tilt. It was composed of 5 continuous sequences of different people moving their heads freely through the view-sphere with each sequence lasting 350 frames and 20 sets of different people randomly exposing to all the poses in the pose-sphere with each having 133 frames. An example of the view-sphere captured by the database was shown in Figure 2. Among them, a disjoint sub-set of 11 different people were selected as prototypes, as shown in Figure 3. The images were aligned, intensity and contrast normalised and expression changes were minimal. The aim of the experiments was then to recognise and track both the poses and the identities of the other subjects in the database using similarity measures to the 11 prototypes.

A person-independent model was learned based on the 11 prototypes. This model was then used to recognise and track head pose of a novel subject who was not one of the prototypes. Figure 6 shows a few examples in which the tracked pose was compared with the ground-truth provided by the magnetic sensor. Overall, the average error (in degrees) over some 1300 frames from different sequences of different people moving in pose range of $\pm 90^\circ$ yaw and $\pm 30^\circ$ tilt is shown in Table 1. It is worth pointing out that pose tracking with Kalman filter-based prediction did not improve the performance over tracking without prediction.

For recognising and tracking identities across views, we again used 11 prototypes. Figure 7 shows examples of sequences of moving faces with their identities been recognised and tracked across views (1) using interpolation of similarity measures between known views at -40° and +40° yaw for all the tilt angles (top plot), and (2) between known views at 20° yaw intervals for all tilt (bottom plot). The average error rates with different known views over 400 different images from different sequences is shown in Table 2. Simultaneous pose and identity recognition and tracking can be achieved over 20 Hz on a 200MHz P5 with Matrox Meteor board.

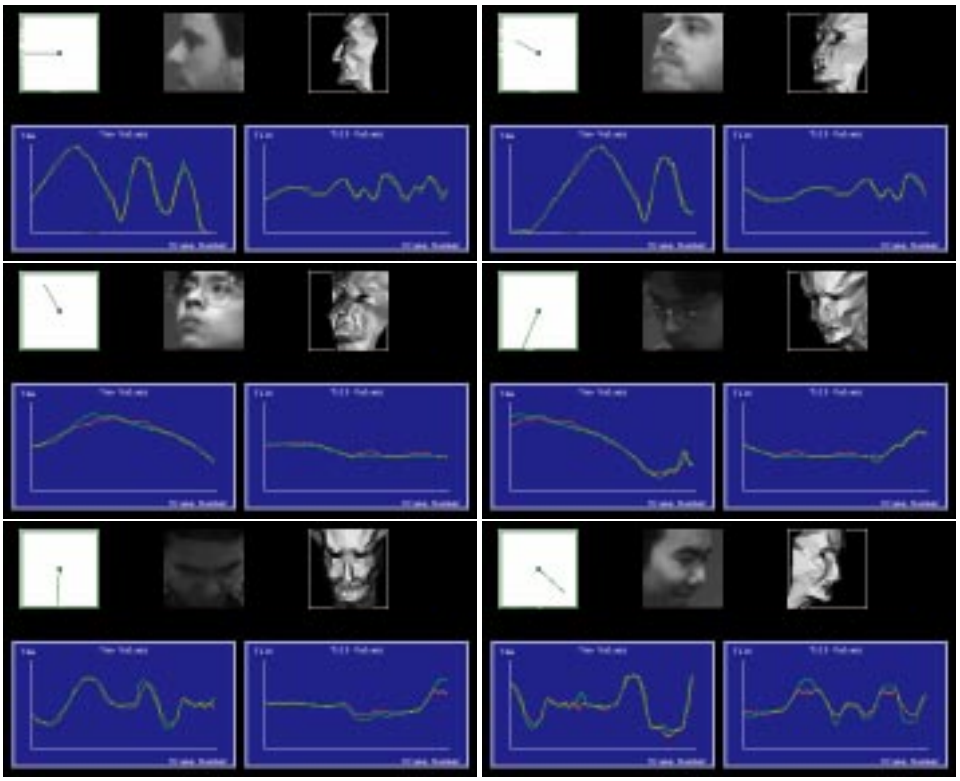

Figure 6: Head pose tracked by a person-independent model shown with the ground-truth measured by the magnetic sensor, yaw on the left and tilt on the right. All the faces been tracked were unknown to the prototypes used for building the model.

Appearance models	Error with prediction		Error without prediction	
	yaw	tilt	yaw	tilt
Person-specific	9.03°	4.50°	10.03°	5.00°
Mean templates	5.50°	8.02°	6.00°	8.02°
Similarity measures	3.50°	3.50°	3.53°	3.50°

Table 1: Average error in pose recognition and tracking of both known and unknown subjects using the person-independent model.

Known Views (yaw)	Error from Generalisation to Novel Views
-90°, 0°, +90°	30.1%
50°, 130°	22.55%
Every 20°	3%

Table 2: Average error in identity recognition and tracking.

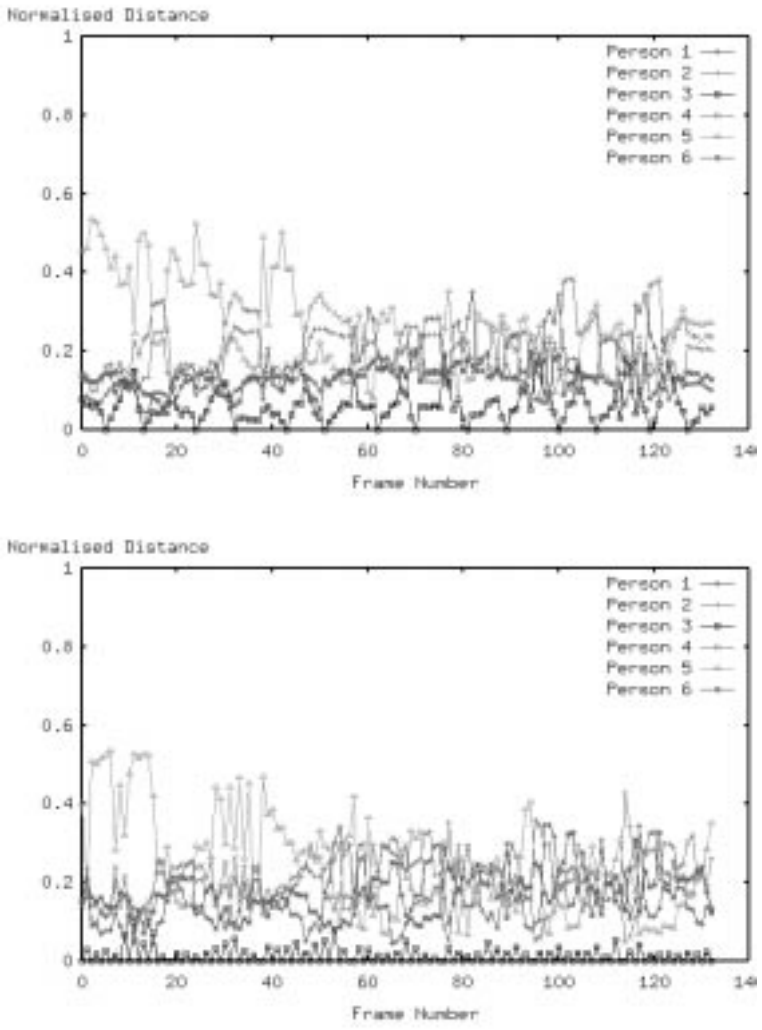

Figure 7: Example results for 1-in-6 identity recognition and tracking of novel faces across the view-sphere using view-based similarity measures to 11 prototypes. Top: recognition by interpolation based on two known views at -40° and +40° yaw. Bottom: interpolation between every 20° yaw intervals. The trajectories were normalised similarity measures of face images from a test sequence to 6 known faces. Person 3 was recognised and tracked across views over time.

To conclude, the method described in this work can be extended in a number of useful ways. Firstly, in all of the above experiments we have established only simple coarse alignment. In theory, more correspondence (affine, dense) could be established since images are only ever compared with similar poses. Secondly, a global principal components analysis could be used to reduce dimensionality prior to applying the above methods.

This makes computation more efficient and will be especially useful as the number of prototypes becomes large. It might however preclude the use of dense correspondence.

References

[1] S. Gong, A. Psarrou, I. Katsoulis, and P. Palavouzis, "Tracking and recognising face sequences," in *European Workshop on Combined Real and Synthetic Image Processing for Broadcast and Video Production*, Hamburg, Germany, November 1994, pp. 96–115, Springer-Verlag.

[2] S. McKenna and S. Gong, "Recognising moving faces," in *Face Recognition: From Theory to Applications*, Wechsler, Philips, Bruce, Fogelman-Soulie, and Huang, Eds. Springer-Verlag, July 1998.

[3] S. McKenna and S. Gong, "Real-time face pose estimation," *Real Time Imaging*, 1998, To appear in the Special Issue on Real-time Visual Monitoring and Inspection.

[4] S. Duvdevani-Bar, S. Edelman, A. J. Howell, and H. Buxton, "A similarity-based method for the generalisation of face recognition over pose and expression," in *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, Nara, Japan, April 1998.

[5] T. Vetter and T. Poggio, "Linear object classes and image synthesis from a single example image," *IEEE PAMI*, vol. 19, no. 7, pp. 733–742, July 1997.

[6] S. McKenna and S. Gong, "Tracking faces," in *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, Killington, Vermont, U.S., October 1996, pp. 271–277.

[7] Y. Raja, S. McKenna, and S. Gong, "Tracking and segmenting people in varying lighting conditions using colour," in *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, Nara, Japan, April 1998.

[8] R. Y. Tsai, "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses," *IEEE J. Robotics and Automation*, vol. RA-3, no. 4, pp. 323–344, August 1987.

[9] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces for face recognition," in *CVPR*, 1994, pp. 84–91.

[10] K. Etemad and R. Chellappa, "Discrimenant analysis for recognition of human face images," *Optical Society of America*, vol. 14, no. 8, pp. 1724–1733, August 1997.

[11] S. Gong, E. Ong, and P. Loft, "Appearance-based face recognition under large rotations in depth," in *ACCV*, Hong Kong, January 1998, IEEE.

[12] S. Edelman, "Representation is representation of similarity," *Behavioral and Brain Sciences*, 1998, To Appear.

[13] F. De la Torre, S. Gong, and S. McKenna, "View-based adaptive affine alignment," in *ECCV*, Freiburg, Germany, June 1998.