Modelling Facial Colour and Identity with Gaussian Mixtures

Stephen J. McKenna[a,1], Shaogang Gong[b] and Yogesh Raja[b,2]

[a]*Department of Applied Computing, University of Dundee, Dundee DD1 4HN, Scotland*

[b]*Department of Computer Science, Queen Mary and Westfield College, London, England*

Abstract

An integrated system for the acquisition, normalisation and recognition of moving faces in dynamic scenes is introduced. Four face recognition tasks are defined and it is argued that modelling person-specific probability densities in a generic face space using mixture models provides a technique applicable to all four tasks. The use of Gaussian colour mixtures for face detection and tracking is also described. Results are presented using data from the integrated system.

Key words: Face recognition, Biometrics, Gaussian mixtures, Colour models.

1 Introduction

Face recognition in general and the recognition of moving people in natural scenes in particular, require a set of visual tasks to be performed robustly. These include (1) *Acquisition*: the detection and tracking of face-like image patches in a dynamic scene, (2) *Normalisation*: the segmentation, alignment and normalisation of the face images, and (3) *Recognition*: the representation and modelling of face images as identities, and the association of novel face images with known models. These tasks seem to be sequential and have traditionally often been treated as such. However, it is both computationally and psychophysically more appropriate to consider them as a set of co-operative visual modules with closed-loop feedbacks. In order to realise such a system,

[1] Supported by EPSRC Grant No. GR/K44657.

[2] Supported by an EPSRC/BBC CASE Studentship.

an integrated approach has been adopted which will perform acquisition, normalisation and recognition in a coherent way. Figure 1 illustrates the system design. Images of a dynamic scene are processed in real-time to acquire normalised and aligned face sequences. Typical examples can be seen in Figure 2. In essence, this process is a closed-loop module that includes the computation and fusion of three different visual cues: motion, colour and face appearance models. Face tracking based upon motion and a face appearance model has been addressed in greater detail elsewhere [1,2]. The use of colour is described here. The remainder of this paper then focuses upon person identification within such a framework. Complementary to recognition, appearance-based mechanisms for real-time face pose estimation have been developed which can be used to improve the robustness of detection and alignment [3].

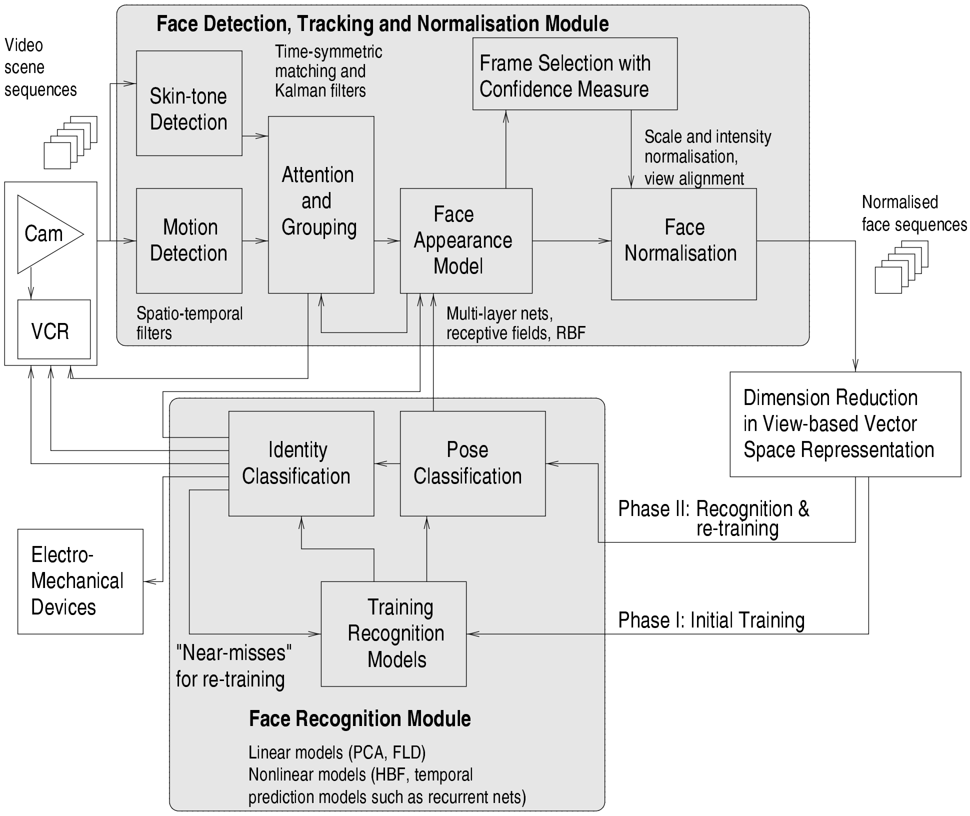
Face Detection, Tracking and Normalisation Module
Video scene sequences
Skin-tone Detection
Time-symmetric matching and Kalman filters
Frame Selection with Confidence Measure
Scale and intensity normalisation, view alignment
Attention and Grouping
Cam
Motion Detection
Face Appearance Model
Face Normalisation
Normalised face sequences
VCR
Spatio-temporal filters
Multi-layer nets, receptive fields, RBF
Dimension Reduction in View-based Vector Space Reppresentation
Identity Classification
Pose Classification
Phase II: Recognition & re-training
Electro-Mechanical Devices
Training Recognition Models
Phase I: Initial Training
"Near-misses" for re-training
Face Recognition Module
Linear models (PCA, FLD)
Nonlinear models (HBF, temporal prediction models such as recurrent nets)

Fig. 1. *A framework for face recognition in dynamic scenes.*

Much research effort has been concentrated on face recognition tasks in which only a single image or at most a few images of each person are available. A major concern has been scalability to large databases containing thousands of people (e.g. [4]). However, large intra-subject variability casts doubt upon the possibility of scaling face recognition, at least in this form, to very large populations. A form of biometric “facial” recognition using the iris is better suited to such populations [5]. In contrast, the face recognition tasks considered in this paper are characterised by the availability of many images of relatively small groups of individuals. Such data arise from the type of integrated approach to face recognition in dynamic scenes illustrated in Figure 1. Since these tasks involve recognition of fewer people with more images, they might appear initially to be simpler. However, applications of the "many people with few images" variety typically use images captured in highly constrained conditions. In contrast, the tasks considered here require recognition to be performed using sequences acquired and normalised automatically in poorly constrained dynamic scenes. These are characterised by low resolution, large scale changes, variable illumination and occasionally inaccurate cropping and alignment. Recognition based upon isolated images of this kind is highly inconsistent and unreliable. However, the poor quality of the data can be compensated by accumulating recognition scores over time. Many images of a person can be acquired in a few seconds. Given sufficient data, it becomes possible to model class-conditional structure, i.e. to estimate probability densities for each person.

In section 2, the use of Gaussian mixture colour models for face detection and tracking is described. In section 3, four face recognition tasks are defined and possible approaches to each of these are discussed. It is argued that estimating class-conditional densities in a "face space" provides appearance-based models of identity suited to all four tasks. Gaussian mixtures are then presented and evaluated for this purpose. Conclusions are drawn in section 6.

2 Locating and tracking faces using colour

A system for detecting and tracking faces was previously described [1,2]. It combined motion detection by spatio-temporal filtering with an appearance-based face model in the form of a neural net. Multiple person tracking was performed using time-symmetric matching and Kalman filtering. In this section, the use of colour as a cue for detection and tracking is described. Colour provides a computationally efficient yet effective method which is robust under rotations in depth and partial occlusions. It can be combined with motion and appearance-based face detection.

Human skin forms a relatively tight cluster in colour space even when different races are considered [6]. Figure 3 shows the colour distribution of three faces in hue-saturation (H-S) space. Face colour distributions were modelled as Gaussian mixtures of the form:

$$p(\mathbf{x}) = \sum_{j=1}^{M} p(\mathbf{x}|j)P(j) \tag{1}$$

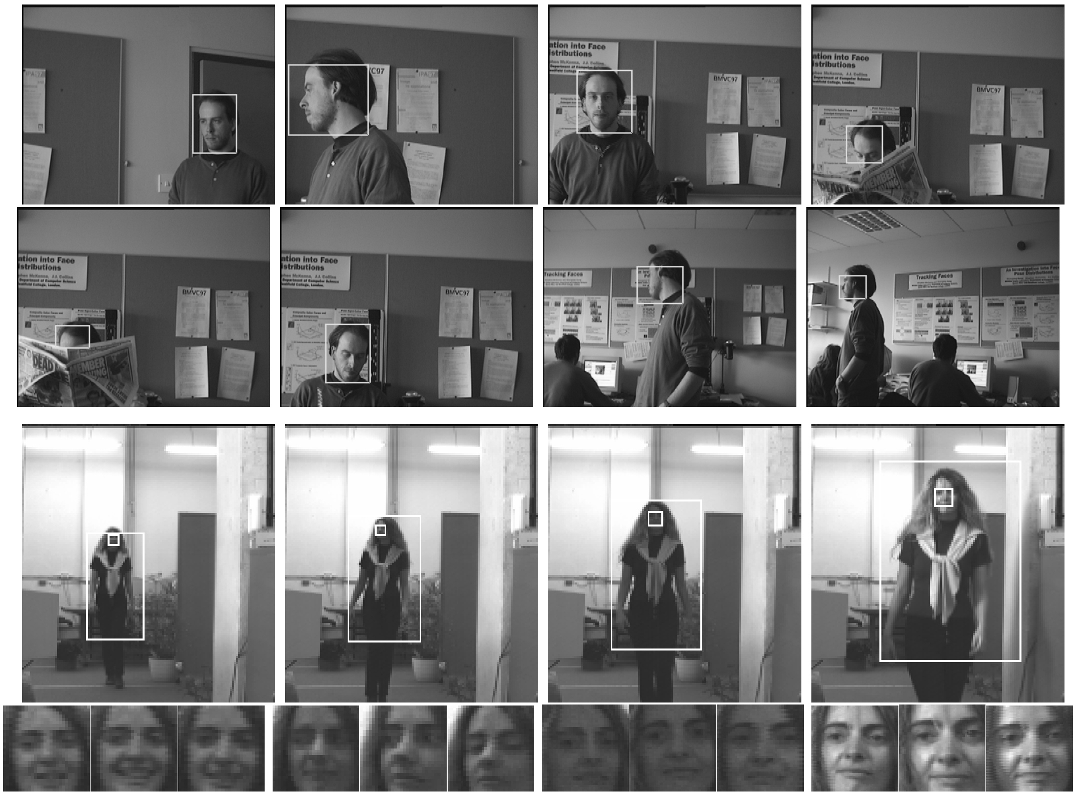

Fig. 2. *Real-time tracking and normalisation in dynamic scenes using colour, motion and neural nets based face appearance models. In the first sequence the system uses a colour model to cope with large pose variations and partial occlusion whilst the camera pans and zooms. In the second sequence, approximate body bounding boxes are also shown along with aligned and scale-normalised faces.*

The mixing parameter $P(j)$ corresponds to the prior probability that the data, $\mathbf{x}$, was generated by component j. Each of the M mixture components, $p(\mathbf{x}|j)$, is a Gaussian with mean $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$. Given n face pixels $\mathbf{x}_i$, $i = 1 \ldots n$, Expectation-Maximisation (EM) provides an effective maximum-likelihood algorithm for learning a Gaussian mixture model [7]. An expectation (E) step consists of evaluating the posterior probabilities $P(j|\mathbf{x}_i)$ for each mixture component. Let the sum of these probabilities be $S_j = \sum_{i=1}^{n} P(j|\mathbf{x}_i)$. A maximisation (M) step then updates the mixture components as follows:

$$P(j)^{new} = \frac{S_j}{n}, \qquad \boldsymbol{\mu}_j^{new} = \frac{1}{S_j} \sum_{i=1}^{n} \mathbf{x}_i P(j|\mathbf{x}_i), \tag{2}$$

$$\Sigma_j^{new} = \frac{1}{S_j} \sum_{i=1}^{n} [\mathbf{x}_i - \boldsymbol{\mu}_j^{new}] \cdot [\mathbf{x}_i - \boldsymbol{\mu}_j^{new}]^T P(j|\mathbf{x}_i) \tag{3}$$

The E and M steps are iterated until convergence. If $M = 1$, the parameters of the Gaussian are estimated directly.

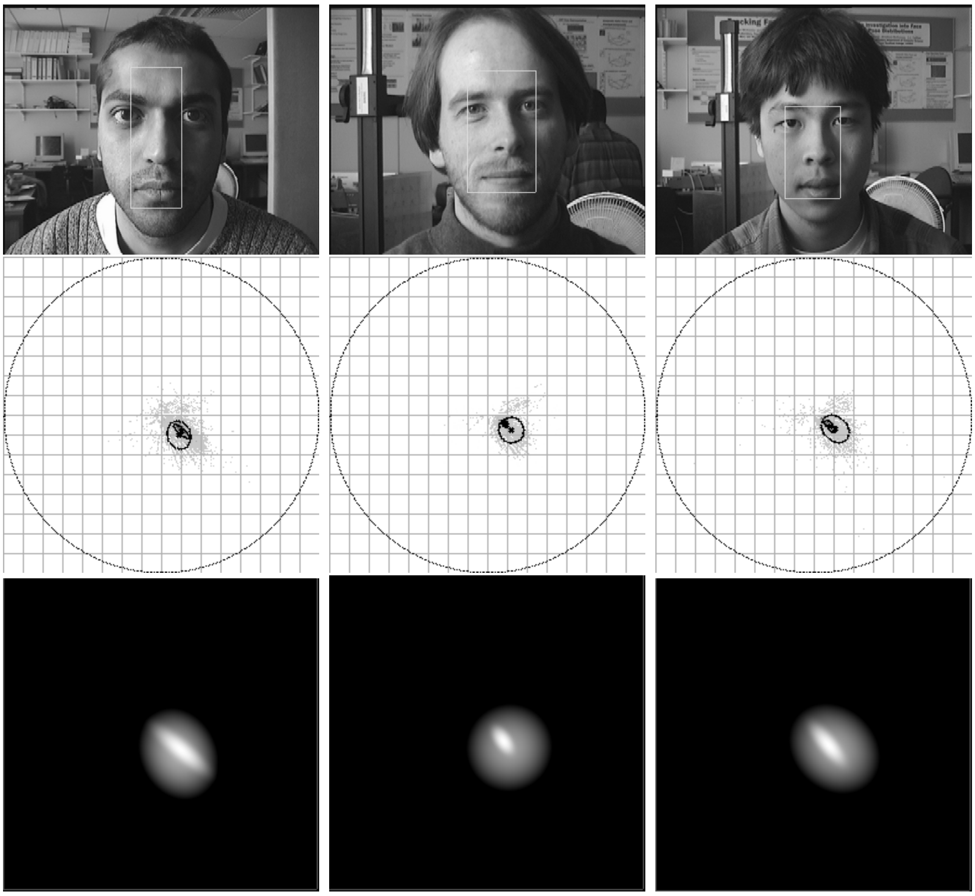

Fig. 3. *The tight clustering of skin colour for three different races is illustrated here. The top row shows the face regions used to build the mixture models. The middle row shows the colour distributions and mixture components plotted in H-S space. The bottom row shows the probability density models in H-S space.*

In practice, an H-S model of a single person functions well with other races. The mixture model is used to assign a probability to each pixel in an image and faces are detected by grouping suitably sized areas of high probability.

The tracking dynamics involve estimating the position of the object and the size of a bounding box. This box provides a focus of attention for further processing. The position and size of the box are found by computing the mean $\mathbf{m}^t = (m_x, m_y)$ and standard deviations $\boldsymbol{\sigma}^t = (\sigma_x, \sigma_y)$ of the local colour probability distribution within a rectangular search area centred on $\mathbf{m}^{t-1}$ in the image domain at time t. The dimensions of this search area are determined by scaling the dimensions of the bounding box at time $t-1$. The experiments presented in this paper were performed with search areas $\frac{3}{2}$ times the height and width of the bounding box.

For a given frame t, the box position $\mathbf{m}^t$ is estimated as an offset from the position $\mathbf{m}^{t-1}$:

$$\mathbf{m}^t = \mathbf{m}^{t-1} + \frac{\sum_{\boldsymbol{\xi}} p(\mathbf{x}_{\boldsymbol{\xi}})(\boldsymbol{\xi} - \mathbf{m}^{t-1})}{\sum_{\boldsymbol{\xi}} p(\mathbf{x}_{\boldsymbol{\xi}})} \tag{4}$$

where the sums are computed over the search region i.e. $\boldsymbol{\xi}$ ranges over all image coordinates in the search region and $\mathbf{x}_{\boldsymbol{\xi}}$ is the colour point at image position $\boldsymbol{\xi}$. To improve accuracy, probabilities $p(\mathbf{x}_{\boldsymbol{\xi}})$ are thresholded. Values lower than the threshold are taken to be background and are consequently set to zero in order to nullify their influence on the estimation of $\mathbf{m}^t$ and $\boldsymbol{\sigma}^t$.

The size of the bounding box is estimated by computing the standard deviation of the image probability density:

$$\boldsymbol{\sigma}^t = \sqrt{\frac{\sum_{\boldsymbol{\xi}}[p(\mathbf{x}_{\boldsymbol{\xi}})\{(\boldsymbol{\xi} - \mathbf{m}^{t-1}) - \mathbf{m}^t\}^2]}{\sum_{\boldsymbol{\xi}} p(\mathbf{x}_{\boldsymbol{\xi}})}} \tag{5}$$

Figure 2 shows a sequence of a face being tracked with a moving camera against a cluttered background. The tracker's ability to deal with changes in scale, large rotations in depth and partial occlusion are all clearly demonstrated.

The colour-based tracking system has been implemented on a 200MHz Pentium PC equipped with a Matrox Meteor frame grabber and a Sony EVI-D31 active camera. The camera can be driven by maintaining the mean position, $\mathbf{m}$, at the centre of the image. Tracking is performed at approximately 15 frames per second. Some problems are inevitably caused by large changes in the spectral composition of scene illumination. It has been found necessary to use at least two colour models, one for interior lighting and one for exterior natural daylight. Adaptive colour models can be used to perform tracking under varying illumination conditions [8].

3 Face recognition tasks

Given a database consisting of a set, $\mathcal{S}$, of N known people, different face recognition tasks can be envisaged. Four tasks are defined here as follows:

(1) *Face classification*: The task is to identify the subject under the assumption that the subject is a member of $\mathcal{S}$.
(2) *Known/Unknown*: The task is to decide if the subject is a member of $\mathcal{S}$.

(3) *Identity verification*: The subject's identity is supplied by some other means and must be confirmed. This is equivalent to task 2 with $N = 1$.

(4) *Full recognition*: The task is to determine whether or not the subject is a member of $\mathcal{S}$, and if so to determine the subject's identity.

When considering appearance-based approaches to these tasks it is helpful to know something of the topology of sets of face images in an image space. The set of all faces forms a small number of extended, connected regions [3]. Furthermore, a face undergoing transformations such as rotation, scaling and translation results in a connected but strongly non-convex subregion in the image space [9]. Whilst these transformations might be approximately corrected using linear image-plane transformations, large rotations in depth, illumination changes and facial expressions cannot be so easily "normalised". Therefore, the set of images of a single face will form at least one and possibly several, highly non-convex, connected regions in image space.

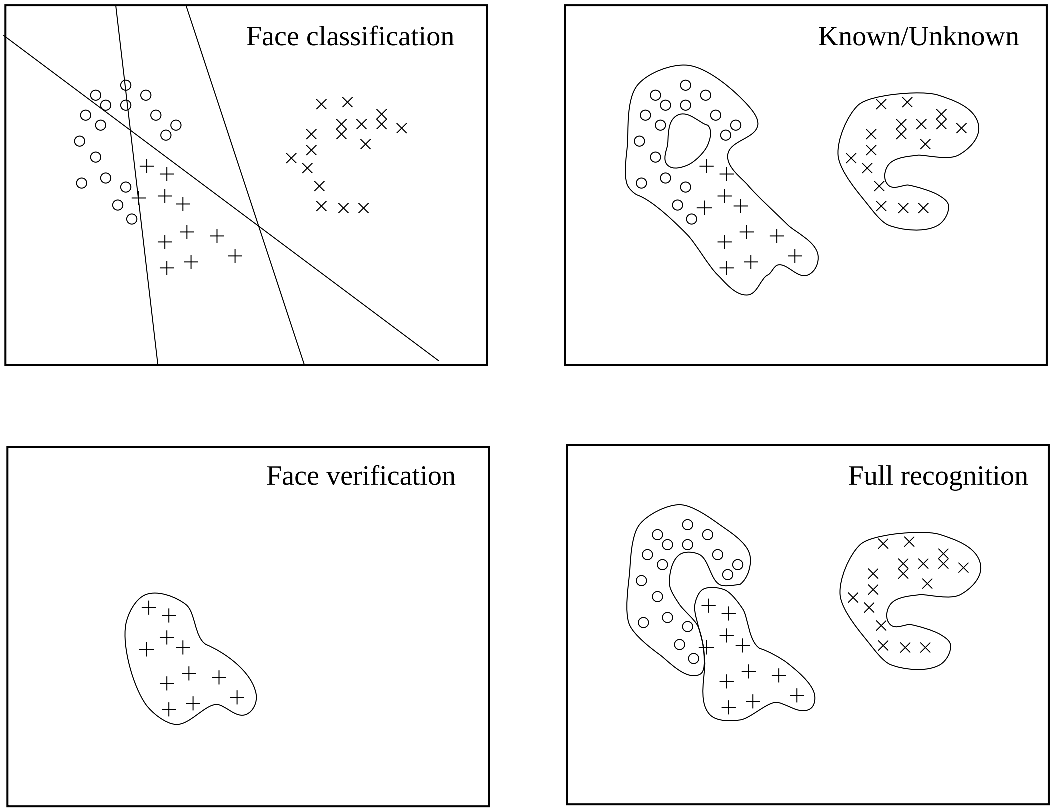

Fig. 4. *Plotted in a hypothetical face space, $\mathcal{F}$, are example faces from 3 different people. Suitable decision boundaries are shown for the four recognition tasks.*

Figure 4 illustrates the four recognition tasks defined above in a hypothetical face space $\mathcal{F}$, where $\mathcal{F}$ is assumed to contain all possible face images and to

[3] A single connected region may not be sufficient because of certain binary properties of faces such as the presence or absence of glasses.

exclude all other images. Plotted in $\mathcal{F}$ are example faces for three different people[4]. Suitable decision boundaries for performing the recognition tasks are shown. The separability of face identities in $\mathcal{F}$ will depend upon the technique used to model $\mathcal{F}$. However, it is likely that each identity will form strongly non-convex regions in this subspace. In the *face classification* task, all N classes can be modelled. In contrast, the other three tasks all suffer from the need to consider the class of unknown faces. Each task will now be discussed in greater detail.

3.1 Face classification

The face classification task is an N-class classification problem in which all N classes can be modelled. It can be tackled by collecting representative data for each of the N classes and applying one of many possible pattern classification techniques. The probability of misclassifying a face x is minimised by assigning it to the class C_k with the largest posterior probability $P(C_k|x)$, where

$$P(C_k|x) = \frac{p(x|C_k)P(C_k)}{p(x)} \tag{6}$$

$p(x)$ is the unconditional density, $p(x|C_k)$ is the class-conditional density and $P(C_k)$ is the prior probability for class C_k. Since $p(x)$ is the same for every class it need not be evaluated in order to maximise posterior probability [10]. Therefore, one approach to the classification task is to model the class-conditional probability densities, $p(x|C_k)$, for each class. This approach is explored in this work. An alternative approach is to estimate discriminant functions using, for example, Linear Discriminant Analysis [11,12].

3.2 Face verification

Face verification can be treated as a 2-class classification problem. The two classes C_0 and C_1 correspond to the cases where the claimed identity is true and false respectively. In order to maximise the posterior probability, x should be assigned to C_0 if and only if

$$p(x|C_0) > \frac{p(x|C_1)P(C_1)}{P(C_0)} \tag{7}$$

[4] Several simplifications have been made here for illustrative purposes. Firstly, the identities are likely to overlap significantly. Secondly, the space has high dimensionality and visualisation in two dimensions would in fact reveal little or no structure. Thirdly, each identity has been shown as a single connected region whereas several such regions may be required in reality.

Density $p(x|C_1)$ represents the distribution of faces other than the claimed identity. This is difficult to model but a simple assumption is that it is constant over the relevant region of space, falling to zero elsewhere. In this case, Inequality (7) is equivalent to thresholding $p(x|C_0)$. Perhaps a more accurate assumption is that the density $p(x|C_1)$ is smaller in regions of space where $p(x|C_0)$ is large. If $p(x|C_1)$ is chosen to be of the form $F[p(x|C_0)]$, where F is a monotonically decreasing function, then this assumption is also equivalent to thresholding $p(x|C_0)$. In this case, the threshold takes the form $G^{-1}\left[\frac{P(C_0)}{P(C_1)}\right]$, where $G(z) \equiv F(z)/z$. Since G is monotonic, G^{-1} is unique[5]. Utilising only data from class C_0, it is therefore reasonable to perform verification by thresholding $p(x|C_0)$.

In order to achieve more accurate verification, negative data, i.e. data from class C_1, would need to be used in order to better estimate the decision boundaries. Only data which are "close" to C_0 are relevant here. An iterative learning approach can be used in which incorrectly classified unknown faces are selected as negative data. Furthermore, the face images used to train the face detection network also provide a suitable source of negative examples for identity verification [1].

3.3 Known/Unknown

This task can also be treated as a 2-class classification problem. The two classes C_0 and C_1 correspond to the cases where the subject is and is not a member of the known group $\mathcal{S}$, respectively. The methods discussed above for face verification can be similarly applied to this 2-class problem.

A slightly different approach involves building an identity verifier for each person in $\mathcal{S}$. The known/unknown task is performed by carrying out N identity verifications. If the numerator in the threshold of Inequality (7) is the same for all verifiers then they can be combined in a straightforward manner.

3.4 Full recognition

The full recognition task can be performed by combining N identity verifiers similarly to the second approach described above for known/unknown.

[5] This fact was pointed out by Bishop [13] in the context of neural network validation.

4 Methods for face recognition tasks

The approach proposed in this work provides a recognition framework that can be applied to any of the four tasks defined in Section 3. The main idea is to model a class-conditional density for each person in a representation space of relatively low dimensionality. Given such class-conditional densities, all four recognition tasks can be performed in a well-founded, statistical way. However, the method chosen to estimate these densities needs to be sufficiently general in order to model the highly non-convex distributions generated by different images of a face. It should also allow for a range of model complexity in order to model people for whom a relatively small amount of data are available. As more data are collected through recognition the model should be able to adapt to capture the underlying distribution more accurately.

The method selected here for density estimation was Gaussian mixture models. Modelling face classes with mixture models has several attractive characteristics. Density estimation is performed in a semi-parametric way so that the size of the model (number of mixture components) scales with the complexity of the data rather than with the size of the data set. The method is sufficiently general to model highly complex, non-linear distributions given enough data. However, it can also be constrained in a straightforward manner to obtain well-conditioned estimation given limited data. When classification is performed, other models emerge as special cases of using Gaussian mixtures, e.g. nearest neighbour and nearest mean classification.

4.1 Modelling identity using Gaussian mixtures

Let each person k constitute a class C_k. A person's identity is modelled by estimating the class-conditional density, $p(x|C_k)$, from examples of that person's face. This density takes the form of a mixture of M components estimated using the EM algorithm described in section 2:

$$p(x|C_k) = \sum_{j=1}^{M} p(x|j)P(j) \tag{8}$$

However, appearance-based face representations usually have high dimensionality and in practice fitting a mixture of Gaussians is often highly under-constrained due to limited data and the "curse of dimensionality". There are a number of complementary approaches to making the modelling tractable.

Firstly, the number of parameters in the model can be reduced by constraining the form and the number of Gaussian mixture components. In the most general case, each Gaussian, j, has a covariance matrix, Σ_j, which is completely determined by the data. If Σ_j is constrained to be a diagonal matrix then there are only $2d$ parameters to be determined, where d is the dimensionality of the data. If $\Sigma_j = \sigma_j I$ for some σ_j then the Gaussian is radially symmetric and there are only $d+1$ parameters to be determined. If $\Sigma = I$ then only the mean must be estimated.

Secondly, the data set can be artificially enlarged by synthesising new *virtual* images for each person using models of possible variations of a face image. In its simplest form, this approach consists of applying a set of simple transformations to the images e.g. small translations, scalings, rotations and mirroring about the vertical axis. Noise can also be artificially added to the images. Models of deformation have been employed for a more complex synthesis of virtual views, e.g. [14].

Thirdly, the dimensionality of the face representation vectors can be reduced. A simple way to reduce dimensionality in the image domain is to consider only a restricted part of the face or to subsample the image. A significant reduction in dimensionality is achieved by representing faces as vectors in the subspace of faces $\mathcal{F}$ rather than as image vectors in the space of all possible images although $\mathcal{F}$ can be difficult to model.

4.2 Modelling face space

Since the intrinsic dimensionality of face space, $\mathcal{F}$, is much less than that of the space of all images, $\mathcal{I}$, a significant reduction in dimensionality can be obtained without loss of significant information provided that two criteria can be met:

(1) The recognition algorithm only ever has to deal with correctly normalised images of faces, i.e. face tracking provides perfect data.
(2) The subspace $\mathcal{F}$ is accurately modelled in such a way that separability of identity is preserved.

A face tracking system has been developed that can largely fulfill the first criterion by using a measure of confidence to discard nearly all the poorly aligned face images [1]. However, there will always be some error in this process, particularly under demanding illumination conditions and with low resolution images.

A representative data set containing a large number of different identities is needed in order to build a *generic* model of the face space. In practice, a *specific* approximation, $\mathcal{F}_\mathcal{S}$, is usually obtained from images in the set $\mathcal{S}$ of N known people. When N is small, $\mathcal{F}_\mathcal{S}$ is a poor approximation to $\mathcal{F}$. If a specific model is used, it must be updated each time the set $\mathcal{S}$ changes. Furthermore, any identity-specific models which make use of $\mathcal{F}_{\mathcal{S}}$ must also be updated. In contrast, a generic model need never be updated. An important point here is that *face classification* is easier to perform in $\mathcal{F}_{\mathcal{S}}$ than in $\mathcal{F}$ while *identity verification*, *known/unknown* and *full recognition* are best performed in a generic face space, $\mathcal{F}$.

In theory, if exact pointwise correspondences can be established between all face images, face space can be accurately modelled using linear vector spaces [14]. In practice, establishing even a small set of feature correspondences between faces is highly problematic, especially at low resolution. In experiments described in section 5, only approximately aligned frontal or near-frontal views of faces are considered and linear models can provide reasonably accurate representation [15]. Principal Components Analysis (PCA) has been used to obtain face space models for face classification [16]. The models are computed without the use of any identity class information. PCA is therefore suitable for data sets with only a few example images per person and (or) large numbers of people. Linear discriminant analysis has also been used (e.g. [17]) and can preserve class linear separability when applied to data sets with many images per person and relatively few people. It is therefore suitable for computing specific face space models for face classification using many training images of a few people.

In experiments described in the next section, a large data set containing many different people with only a few images per person was used to compute a generic face space using PCA. First, a brief description is given of the PCA "eigenface" methods used.

4.3 Normalised eigenfaces

Given n face images of size $m = p \times q$ pixels, a face eigenspace is calculated as follows. Each image defines an m-dimensional column vector $\mathbf{x}$. The mean, $\boldsymbol{\mu}$, and the $m \times m$ covariance matrix, $\boldsymbol{\Sigma}$, of the set of n face images are computed. Let $\mathbf{u}_j$, $j = 1 \ldots n'$, be the n' eigenvectors of $\boldsymbol{\Sigma}$ which have the largest corresponding eigenvalues λ_j. The n' eigenvectors are the principal components. For an image, $\mathbf{x}$, an n'-dimensional "pattern vector", $\boldsymbol{\Omega}(\mathbf{x}) = [\omega_1\, \omega_2 \ldots \omega_{n'}]$, can be computed by projection onto each of the eigenvectors $\mathbf{u}_j$:

$$\omega_j = \mathbf{u}_j^T(\mathbf{x} - \boldsymbol{\mu}) \qquad j = 1, \ldots, n' \tag{9}$$

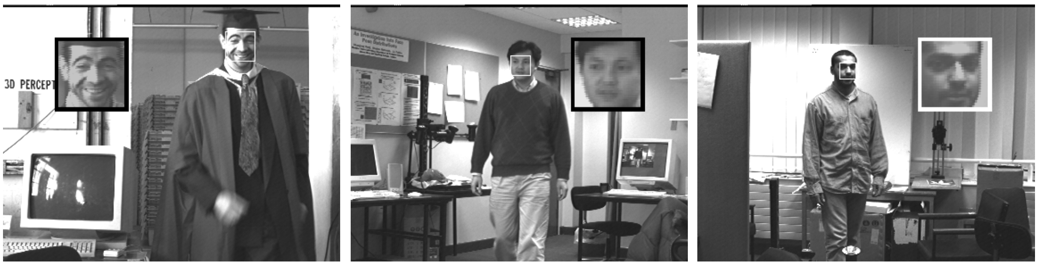

Fig. 5. *Example frames from tracked sequences. The face bounding box is overlaid on each image. The extracted face images are shown inlaid.*

This pattern vector can be normalised by the eigenvalues in order to give the data equal variance along each principal component axis:

$$\mathbf{\Omega_{norm}}(\mathbf{x}) = [\frac{\omega_1}{\lambda_1} \frac{\omega_2}{\lambda_2} \dots \frac{\omega_{n'}}{\lambda_{n'}}] \tag{10}$$

Class-conditional densities can be modelled in a principal subspace by estimating either $P[\mathbf{\Omega}(\mathbf{x})|C_k]$ or $P[\mathbf{\Omega_{norm}}(\mathbf{x})|C_k]$.

5 Experiments

This section describes experiments using Gaussian mixture models of identity. Face image data were acquired and normalised fully automatically by the face tracking system. The neural network model used to perform tracking was trained using 9000 example face images rotated by ±10° and scaled to 90% and 110% [1]. The normalised faces from the tracker therefore varied by at least these amounts in scale and rotation. Since the aim of these experiments was to compare methods for modelling identity rather than to optimise recognition accuracy, no attempt was made to reduce these variations.

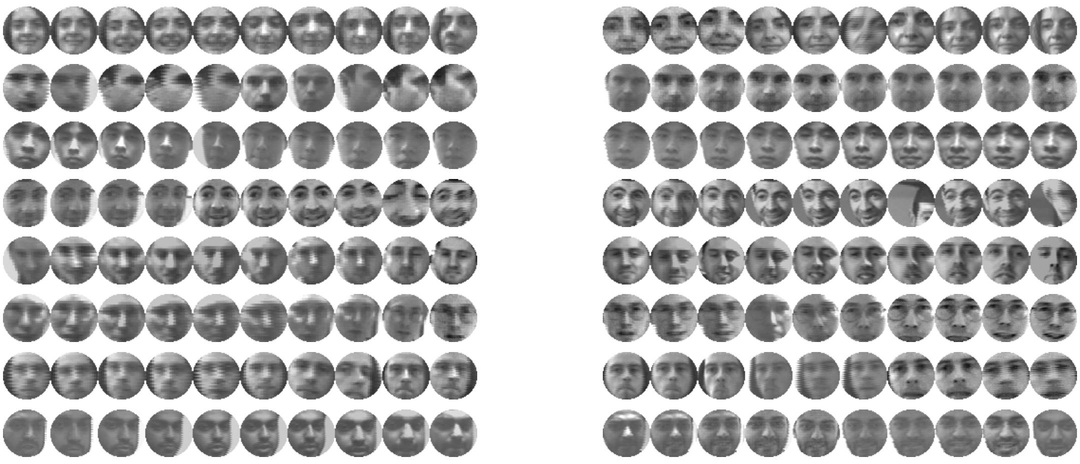

Fig. 6. *Subset of data (Left: training, Right: test).*

Face space	Person (% images correct)								Total %	Seq. (Maj.)
	0	1	2	3	4	5	6	7		
Specific	75	64	74	85	56	78	29	11	55.1	7
Generic	57	67	66	20	13	72	25	29	43.6	4

Table 1
Test set results with generic and specific face space models using 40 principal components. Identities were modelled by fitting a single radial Gaussian to each of the 8 people.

Eight subjects were tracked through relatively unconstrained indoor scenes as they walked towards a fixed camera. Overhead lighting resulted in variations in facial illumination. Figure 5 shows three examples from the tracking process. The resolution of the area of the face tracked ranged from approximately 10 × 10 pixels when the subject was far from the camera to 80 × 80 pixels when the subject approached the camera. Two normalised face sequences were obtained for each subject. The first sequence of each subject was used for training and the second sequence for testing. In total, there were 326 training images and 296 test images. The number of training images per person varied from 21 to 60 and the number of test images from 21 to 53. Figure 6 shows 10 of the images used to form the training and test sets for each of the subjects.

Face space was modelled by performing PCA on the training images. A specific model was computed from the training set. A generic model was computed using 644 of the images used to train a face detection neural network in the tracking system. These images were highly suitable, having similar variations in scale and rotation to the tracked data to be recognised. The training images were projected onto the first n' eigenvectors and each person's identity was modelled by estimating either $P[\mathbf{\Omega}(\mathbf{x})|C_k]$ or $P[\mathbf{\Omega_{norm}}(\mathbf{x})|C_k]$ with Gaussian mixtures. The 8 mixture models' parameters were stored along with the n' eigenvectors and eigenvalues and subsequently used to perform classification of the test sequences.

Initially, both a specific and a generic eigenspace were computed using the first 40 eigenvectors. Table 1 shows a comparison of face classification using the specific and generic models. Identities were modelled by fitting a single radial Gaussian to each person's data. The percentage of images correctly classified for each person along with the percentage of total images classified correctly are given. Sequence classification results are also given based upon a majority vote i.e. the sequence is classified as the person with the most images. The result illustrates the fact that the use of a generic face space which could be used to facilitate identity verification, known/unknown or full recognition, in turn makes face classification more difficult.

A reduction in the dimensionality of the generic face space from 40 to 20 did

Name	M	Σ type	$\frac{y_i}{\lambda_i}$	Tot. %	Seq.	
					Maj.	Pr.
T-P	1	σ_f	N	25.0	2	2
1-NN	n	$\sigma_{\rightarrow 0}$	N	32.1	1	1
T-P$_{norm}$	1	σ_f	Y	46.3	4	4
Radial	1	σ_j	Y	44.3	4	4
Diag	1	Σ_d	Y	42.9	4	3
2-Rad	2	σ_j	Y	52.0	5	7
3-Rad	3	σ_j	Y	42.2	5	5
2-Diag	2	Σ_d	Y	41.9	4	5

Table 2
Test results with a 20-dimensional generic face eigenspace and identity mixture models. Column 2 indicates the number of Gaussians, M. Column 3 indicates the type of Gaussian where Σ_d denotes a diagonal covariance, σ_j an independent variance and σ_f a variance equal to that of all other components. A 'Y' in column 4 indicates that normalised pattern vectors were used.

not result in any significant loss of accuracy. Face classification results using the 20-dimensional generic space are given in Table 2. Sequences were classified (1) by a majority vote (Maj.) and (2) by accumulating probabilities (Pr.). Gaussian mixture models of various complexity were compared for modelling identity.

The first two methods in Table 2 used unnormalised pattern vectors. The first method (T-P) used single radial Gaussians of equal variance resulting in a nearest-mean classifier which was equivalent to the eigenfaces method of Turk and Pentland [16]. The second method was a nearest neighbour classifier (1-NN). Both these methods performed poorly. However, the use of normalised pattern vectors resulted in a significant improvement with T-P$_{norm}$ classifying 4 sequences correctly. The mixture models had either radial or diagonal covariance Gaussians with between 1 and 3 components. A mixture of 2 radial Gaussians provided the best performance. The use of sequences as opposed to single images yielded improved recognition performance.

6 Conclusions

An integrated approach to face recognition in dynamic scenes was presented. The recognition tasks to be performed by such a system are typically characterised by poor resolution and variable lighting. In contrast to most previously developed face recognition methods, the data sets consist of many images of relatively small groups of known people. Four recognition tasks were defined: face classification, face verification, known/unknown and full recognition. All but face classification require consideration of the class of unknown people. As a consequence, identities should be modelled in a generic face space rather than a face space which is specific to the set of known people.

Mixture models provide an effective way to model identities as class-conditional probability densities in face space. Model complexity adapts to the structure of the data and simplified models are easily obtained when data is lacking. Face data used to compute a face space model for face detection were also used to compute a linear face space model for recognition. The eigenface method of [16] can be viewed as a special case and was outperformed by simple mixture models. It was shown that modelling identities using such models is beneficial given an appropriate level of mixture complexity. This approach to recognition results in a system which can learn and update identity models independently of one another. Recognition was performed using sequences of faces tracked in near real-time under poorly constrained conditions. The use of sequences yielded better recognition performance than the use of single images.

Gaussian mixture colour models were also used to provide an efficient and effective focus of attention for use in face detection and tracking. Adaptive colour models for face tracking under varying illumination conditions are currently being developed. Future work will explore other methods for learning view-based models using density estimation in relatively high-dimensional spaces. A promising approach which has been successfully applied in other application domains such as character recognition is the use of mixtures of principal component analysers [18,19].

References

[1] S. McKenna, S. Gong, and J. J. Collins, "Face tracking and pose representation," in *British Machine Vision Conference*, Edinburgh, 1996.

[2] S. J. McKenna and S. Gong, "Tracking faces," in *Proc. 2nd Int. Conf. on Automatic Face and Gesture Recognition*, Killington, Vermont, US, October 1996.

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

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

[5] J. G. Daugman, "High confidence visual recognition of persons by a test

of statistical independence," *IEEE PAMI*, vol. 15, no. 11, pp. 1148–1161, November 1993.

[6] M. Hunke and A. Waibel, "Face locating and tracking for human-computer interaction," in *28th Asilomar Conf. Signals, Systems and Computers*, California, 1994.

[7] R. A. Redner and H. F. Walker, "Mixture Densities, Maximum Likelihood and the EM Algorithm," *SIAM Review*, vol. 26, no. 2, pp. 195–239, 1984.

[8] S. J. McKenna, Y. Raja, and S. Gong, "Tracking colour objects using adaptive mixture models," *Image and Vision Computing*, 1998, In Press.

[9] M. Bichsel and A. P. Pentland, "Human face recognition and the face image set's topology," *CVGIP: Image Understanding*, vol. 59, no. 2, pp. 254–261, March 1994.

[10] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley, New York, 1973.

[11] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *PAMI*, vol. 19, no. 7, pp. 711–720, July 1997.

[12] K. Etemad and R. Chellappa, "Discriminant analysis for recognition of human face images," *J. Opt. Soc. Am. A*, vol. 14, no. 8, pp. 1724–1733, August 1997.

[13] C. M. Bishop, "Novelty detection and neural network validation," *I.E.E. Proc.-Vis. Image Signal Process.*, vol. 141, no. 4, pp. 217–222, 1994.

[14] D. Beymer and T. Poggio, "Image representations for visual learning," *Science*, vol. 272, pp. 1905–1909, 28 June 1996.

[15] M. Kirby and L. Sirovich, "Application of the karhunen-loeve procedure for the characterization of human faces," *IEEE PAMI*, vol. 12, no. 1, 1990.

[16] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. of Cognitive Neuroscience*, vol. 3, no. 1, 1991.

[17] K. Etemad and R. Chellappa, ," in *Int. Conf. on Audio- and Video-Based Biometric Person Authentication, LNCS 1206*, 1997, pp. 127–142.

[18] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analysers," Tech. Rep. NCRG/97/003, Neural Computing Research Group, Aston University, June 1997.

[19] G. E. Hinton, P. Dayan, and M. Revow, "Modelling the manifolds of images of handwritten digits," *IEEE Trans. Neural Networks*, vol. 8, no. 1, pp. 65–74, January 1997.

Stephen McKenna is a lecturer in the Department of Applied Computing, University of Dundee, Scotland. Dr. McKenna was born in 1969 and graduated B.Sc. (Hons) in Computer Science (Edinburgh 1990), M.Sc. and Ph.D. (Dundee 1993, 1994). He was a post-doctoral researcher at Queen Mary and Westfield College, London (1995-98) and was a visiting research fellow at the Robotics and Automation Laboratory, Tecnopolis CSATA, Italy (1994-95) and at BT Labs, England (1996). His research interests include face recognition, visual surveillance, visually mediated interaction, statistical learning and neural networks.

Shaogang Gong is a senior lecturer in computer science and artificial intelligence in the Department of Computer Science, Queen Mary and Westfield College, London, England. Dr Gong was born in ChungKing, Sichuan Province, China in 1964. He received his BSc with a distinguished graduate award from The University of Electronic Sciences and Technology of China in 1985 and his DPhil in computer vision from Oxford University in 1989. He was a recipient of a Sino-Anglo Queen Elizabeth II Research Scientist Award in 1987, Royal Society Research Fellow (1987-1988), GEC-Oxford Industrial Research Fellow (1989), computer vision research fellow at QMW (1989-1993). His research interests include dynamic vision, visual learning, Bayesian and statistical learning theories, neural networks, visual surveillance, face recognition, visually augmented immersive virtual reality and visually mediated interaction.

Yogesh Raja was born in 1973. He received his B.Sc. Hons in Computer Science (London 1994) and M.Sc. (London 1995). He is currently pursuing a Ph.D. at QMW. His interests include colour for tracking, human body segmentation for virtual studios and neural networks.