Automatic Interpretation and Coding of Face Images Using Flexible Models

Andreas Lanitis, Chris J. Taylor, and Timothy F. Cootes

Abstract—Face images are difficult to interpret because they are highly variable. Sources of variability include individual appearance, 3D pose, facial expression , and lighting. We describe a compact parametrized model of facial appearance which takes into account all these sources of variability. The model represents both shape and gray-level appearance , and is created by performing a statistical analysis over a training set of face images. A robust multiresolution search algorithm is used to fit the model to faces in new images. This allows the main facial features to be located , and a set of shape , and gray-level appearance parameters to be recovered. A good approximation to a given face can be reconstructed using less than 100 of these parameters. This representation can be used for tasks such as image coding, person identification, 3D pose recovery, gender recognition , and expression recognition. Experimental results are presented for a database of 690 face images obtained under widely varying conditions of 3D pose, lighting , and facial expression. The system performs well on all the tasks listed above.

Index Terms—Face recognition, expression recognition, pose recovery, coding-reconstruction, facial feature location, deformable templates.

1 Introduction

Face images have received considerable attention from both the computer vision and signal processing communities. This interest is motivated by the broad range of potential applications for systems able to code and interpret face images. Examples include:

- Personal Identification and access control [5], [36];
- Low-bandwidth communication for videophone and teleconferencing [22], [27];
- Forensic applications including videofit and mugshot recognition [14];
- Human-computer interaction [2], [18];
- Alertness monitoring [34];
- Automated surveillance;

The functionality required to tackle these applications successfully can be expressed in terms of a number of generic capabilities:

- feature location and tracking,
- person identification,
- expression recognition,
- 3D pose recovery, and
- coding.

These are inherently difficult problems because the images involved are complex and are also highly variable, even for a particular individual. Sources of variability include 3D pose, facial expression, individual appearance, lighting, and occluding structure (facial hair, spectacles, etc.). Because of the degree of difficulty, some researchers have concentrated on particular constrained applications; this approach can lead to the development of practical systems, but makes little overall contribution to progress. Others have attempted to tackle the various generic problems independently; the drawback with this approach is that the effects of all the sources of variability are compounded, so it is extremely difficult to extract a description for one characteristic of interest (e.g., individual appearance) which is not sensitive to others (e.g., facial expression and pose) [29].

Our aim has been to develop a unified approach to the problems of face image coding and interpretation. The basis for this is a compact parametrized model of facial appearance that takes into account all the main sources of variability. A robust image search method is used to recover a parametric description for each new face image, by fitting the model to the data. The locations of all the main facial features are recovered implicitly in this process. Less than 100 parameters are required to describe each image sufficiently well to generate a good quality reconstruction of the face, irrespective of individual appearance, facial expression, 3D pose (±15 degrees of horizontal and/or vertical rotation), or lighting. Given this compact, and nearly lossless coding, low-bandwidth transmission is straightforward, while standard statistical pattern recognition techniques can be used to perform such tasks as person identification, gender recognition, expression recognition, and 3D pose recovery.

Our models of facial appearance are statistical, derived from a training set of face images. The shapes of the main features and the spatial relationships between them are represented by a Point Distribution Model (PDM) [10]. This provides a compact, parametrized description of shape for any instance of a face, and can be used in a multiresolution Active Shape Model (ASM) search [8], [9] to locate the fea-

• *The authors are with the Wolfson Image Analysis Unit, Dept. of Medical Biophysics, Stopford Building, University of Manchester, Oxford Road, Manchester M13 9PT, UK.*
E-mail: {lan, ctaylor, bim}@sv1.smb.man.ac.uk.

Manuscript received 14 Apr. 1995; revised 26 Feb. 1996. Recommended for acceptance by J. Daghman and R. Chin.
For information on obtaining reprints of this article, please send e-mail to: transpami@computer.org, and reference IEEECS Log Number P97003.

TABLE 1
DETAILS OF THE FACE DATABASE USED IN OUR EXPERIMENTS

Number of Subjects	30	CONDITIONS	TRAINING IMAGES	TEST IMAGES
Training images per subject	10	Lighting conditions	fixed*	variable
Normal test images per subject	10	3D movements	yes	yes
Difficult test images per subject	3	Expression	variable	variable
Male subjects	23	Distance from camera	variable	variable
Female Subjects	7	Spectacles	no	yes
Ethnic origin	mixed	Beards/Mustaches	yes	yes
Minimum age of subjects	17	Hairstyle changes	no	yes
Maximum age of subject	45	Background	fixed	variable
Time between capturing training/test images	3-54 weeks			

** For images of a particular individual.*

tures in new images. Gray-level appearance is modeled using flexible gray-level models [7], [8] analogous to the shape model. The primary description is provided by a shape-free gray-level model of the whole face. Local gray-level models, attached to points on the shape model, are also used to make ASM search more robust and improve person identification in the presence of partial occlusion. This extended model involves approximately 700 parameters. For our experiments we have used a database containing 690 face images from 30 individuals.[1] Details of the database are shown in Table 1. The training and test sets contain examples which are much more varied in appearance, 3D pose, and facial expression than those previously used in most successful face image interpretation experiments. In addition to the normal test set, a second test set contains images in which subjects were asked to disguise themselves by hiding part of their face; these images were intended to provide a rigorous test of the robustness of the system. Typical images from the training set and both test sets are shown in Fig. 1.

Fig. 1. Examples of images used in our experiments. Training images, test images, and difficult test images are show in top, middle, and bottom rows, respectively.

In the remainder of the paper, we review some of the most relevant literature on face coding and interpretation, describe our approach in more detail, and present results for feature location, coding and reconstruction, 3D pose recovery, person identification, gender recognition, and expression recognition.

1 The database and its contents are available at:
http://peipa.essex.ac.uk/ftp/ipa/pix/faces/manchester

2 BACKGROUND

In this section, we review, briefly, previous work on face image interpretation and coding. We concentrate mainly on approaches which aim to achieve generic functionality, particularly those which are model-based.

2.1 Locating Facial Features

Many researchers describe the use of specialized techniques designed to locate single facial features, within defined search areas. Matched filtering techniques are remarkably successful [5], [30] but do not deal very satisfactorily with variation in feature shapes.

Methods based on deformable templates have proved more effective. Kass et al. [19] describe the use of active contour models—snakes—for tracking lips in image sequences. They initialized a snake on the lips in a face image and show that it is able to deform and accurately track lip movements. A similar technique is described by Waite and Welsh [37] for locating head outlines. They initialize the snake in the image border and the snake contracts until it latches on to the face outline. Because snakes do not incorporate prior knowledge about expected shapes, this approach is easily confused by other structures present in the image and occlusion. Yuille et al. [39] describe the use of deformable templates, based on simple geometrical shapes, for locating eyes and mouths. These templates are similar to snakes, in that they can deform and move under the influence of image evidence in an attempt to minimize an energy function. Yuille's models incorporate shape constraints, but it is difficult to ensure that the form of a given model is sufficiently general or that an appropriate degree of variability has been allowed.

Craw and Cameron [13] describe a model-based approach for locating face outlines. They use a deformable model representing the outline shape, derived from a large number of training images. They place the model on a new image containing a face, and simulated annealing is used to deform, scale, and translate the model until an objective function is maximized. This function has two terms:

- The first is maximized when the model lies on strong edges;
- The second is maximized when the shape of the model looks most like a face outline based on a measure of aspect ratio.

Human faces are characterized by constrained geometrical relationships between the positions of facial features. Some systems exploit these constraints to locate groups of features [5], [13]. For example, once one feature has been located, the positions of other features can be predicted and their search areas reduced significantly. Although some success has been reported with this approach, the behavior of such systems is complex, and it has proved difficult to achieve robust performance.

2.2 Coding and Reconstruction

Kirby and Sirovich [20] propose the decomposition of face images into a weighted sum of basis images (or eigenfaces) using a Karhumen-Loeve expansion. They code a face image using 50 expansion coefficients, and subsequently reconstruct an approximation using these parameters. Many researchers [14], [22], [23], [24], [30], [36], have built on this methodology in an attempt to produce improved eigenfaces for coding and person identification applications.

Model-based coding and reconstruction has received considerable attention in the literature. Terzopoulos and Waters [33] use a model based on the physical and anatomical structure of faces, which incorporates information about tissue and muscles. They track facial features in image sequences and estimate the muscle contractions required to adjust their model in an attempt to produce a faithful reconstruction. This approach can be used for coding and expression recognition.

A number of researchers [6], [27] use variations of the CANDICE model [31]. This is a 3D wire frame model derived using a triangulation algorithm. The 3D pose and expression of the model are controlled by a number of parameters. They initialize the model to a face in a starting frame of a sequence by filling the triangles using texture mapping techniques and arranging the model shape to have the same shape as the person at the transmitting end. The 3D motion of the face in the sequence is calculated and used to drive the model at the receiving end.

2.3 Person Identification

Face identification techniques can be divided into two main categories:

- those employing geometrical features, and
- those using gray-level information.

Techniques based on geometrical features use a number of dimensional measurements, or the locations of a number of control points for classification. Since geometrical features are expression and 3D orientation dependent, explicit methods of normalization must be employed. Brunelli and Poggio [5] use 22 relative geometrical distances between features to represent faces. Correct classification rates of up to 90 percent were obtained when the method was tested on a database containing images from 47 individuals. Craw and Cameron [14] represent faces in terms of the coordinates of 59 key points. They attempt to minimize the effects of position and scale by using least squares minimization of the Euclidean distances between five control points located on each test image and the corresponding control points located on the average face shape. This approach was only tested on images which did not display significant variation in 3D pose.

Turk and Pentland [36] describe how principal component analysis of the gray-levels of face images can be used to create a set of eigenfaces. Any face can be approximated for identification purposes by a weighted sum of eigenfaces. During the eigenvalue decomposition, no shape normalization takes place, and, for the identification system no shape information is employed. Up to 96 percent correct classification was obtained when this approach was tested on a database containing images from 16 different individuals. Craw and Cameron [14] describe a similar method, except that they normalize the shapes of faces in order to ensure that only gray-level variations are modeled. Training face images are deformed to the mean shape and principal component analysis applied to obtain shape-free eigenfaces. During identification, test images are deformed to the mean shape, and the weights of the eigenfaces required to approximate the new face are calculated and used as the classification parameters. Using this approach, test faces were retrieved correctly from a database containing 100 images. These results relied on a user interactively locating 59 key points on the test images. When a similar experiment was performed using shape information alone, the results were not as good as those obtained with eigenfaces.

Cottrell et al. [11], [12] describe the use of an MLP for processing face images. They present a number of face images to the network and train it to perform various tasks such as coding, identification, gender recognition, and expression recognition. During this procedure, face images are projected onto a subspace in the hidden layers of the network; it is interesting to note that this subspace is very similar to the eigenfaces space. However, an important difference is that, in this case, the face subspace is defined according to the application for which the system is to be used. Correct identification rates of up to 97 percent were reported when the system was tested using a database of images from 11 individuals.

Lades et al. [21] describe a different approach to the problem. During training, they overlay a rectangular grid on a training image, from each individual in their face database. They measure the responses at each of the grid points for a set of 2D Gabor filters tuned to different orientations and scales. When a new image is presented to the system, the grid is overlaid and allowed to deform. A similarity measure between the new image and each training image is computed based on the responses of the same set of Gabor filters and the grid distortion. The new face image is identified as the individual for whom the similarity measure is maximized. This method uses both gray-level information (in the form of Gabor filter responses) and shape information (in the form of grid distortion). Classification experiments [17], [21] have shown that the approach can cope with changes in expression, orientation, and small changes in the lighting conditions.

2.4 Expression Recognition

Psychologists working in the area of facial expression understanding define seven distinct facial expressions [15]: happiness, sadness, surprise, fear, anger, disgust, and neutral. Many of the expression recognition systems reported in the literature are trained to classify expressions into these seven categories.

Yacoob and Davis [38] describe a method for interpreting facial expressions in image sequences based on motion detection. They analyze interframe motion of edges extracted in the area of the mouth, nose, eyes, and eyebrows. During a training phase, they establish a set of rules concerning the motion of these edges during transitions between expressions. Based on these rules they successfully interpret expressions in face image sequences. This approach to the problem is not applicable to static images.

Cottrell and Fleming [12] use a Multi-Layer Perceptron for recognizing expressions. When tested on the face images used for training, their system classifies positive emotions (like happiness, delighted, relaxed), reliably, but the results obtained for negative expressions (like sleepy, angry, sad) are not very good.

2.5 3D Pose Recovery

Tsukamoto et al. [35] describe how 3D pose can be recovered. They divide each face image into a large number of blocks which they parametrize in terms of intensity and edge strength. Once they detect a face in an image, they use a deformation algorithm to simulate rotation in three different directions, generating a temporary model for each direction of rotation. In subsequent frames, they correlate these models with the detected face; the 3D pose is estimated as a linear function of the model correlations.

Gee and Cipolla [18] describe the estimation of the direction of gaze using a simple facial model. They use the location of the eyes, nose, and mouth in face images in order to define four length measures based on which they calculate the direction of gaze.

Bichsel and Pentland [2] use motion analysis and template matching to track head movements in image sequences. At each frame, the head detector returns the location and orientation parameters of the face. By studying the variation of these parameters they interpret head movements, like nodding and shaking.

3 Overview of our Approach

Rather than treating feature location, person identification, expression recognition, 3D pose recovery, and coding as separate goals, we have attempted to develop a unified approach. The basis for this is a compact, parametrized model of facial appearance, which accounts for all the important, systematic sources of variability. Our approach can be divided into two main phases:

- Modeling, in which flexible models of facial appearance are generated, and
- Interpretation, in which the models are used for coding and interpreting face images.

3.1 Modeling

We model the shapes of facial features and their spatial relationships using a single flexible shape model [8], [10]. The model is derived from a set of training images. In each image, the main features are delineated by a large number of landmark points. The model is generated by a statistical analysis of the positions of the landmark points over the training set; it describes the mean shape, and is capable of representing variation due to differences between individuals, change in 3D pose, and change in expression.

Previous investigators have shown that gray-level information is extremely important for interpreting face images [5], [14]. We have, therefore, augmented our shape model with gray-level information using two complementary approaches. In the first, we generate a flexible gray-level model of "shape-free" appearance by deforming each face in the training set to have the same shape as the mean face, and training a flexible "shape-free" gray-level model. This is similar to the method used by Craw and Cameron [14]. In the second approach, we use a large number of local gray-level profile models, one at each landmark point of the shape model. The first approach is more complete, but the second is more robust to partial occlusion.

Shape and gray-level models are used together to describe the overall appearance of each face image; collectively, we refer to the model parameters as appearance parameters. When the face image is coded in terms of the shape model and the gray-level model of the shape-free appearance, less than 100 appearance parameters are required. It is important to note, that in this case, the coding we achieve is reversible—a given face image can be reconstructed from its shape and shape-free gray-level parameters.

3.2 Interpretation

When a new image is presented to our system, facial features are located automatically using an ASM search [9] based on the flexible shape model obtained during training. The resulting automatically located model points are transformed into shape model parameters. Gray-level information at each model point is collected and transformed to local gray-level model parameters. Then, the new face is deformed to the mean face shape and the gray-level appearance is transformed into the parameters of the shape-free gray-level model. The resulting set of appearance parameters can be used for image reconstruction, person identification (including gender recognition), expression recognition, and 3D pose recovery.

4 Flexible Models

All the models used in our system (both shape and gray-level models) are of the same mathematical form. A flexible model [7], [8], [10] is generated from a set of training examples. Each training example ($\mathbf{X_i}$) is represented by N variables.

$$\boldsymbol{X_i} = \left(x_{1i}, x_{2i}, \ldots, x_{Ni}\right) \tag{1}$$

Where x_{ki} is the kth variable in the ith training example. For instance, in constructing a shape model, the x_{ki} represent the coordinates of landmark points, expressed in a

standard frame of reference. The average example, $\overline{\mathbf{X}}$, is calculated, and the deviation of each example from the mean is established. A principal component analysis of the covariance matrix of deviations reveals the main modes of variation. Any training example $\mathbf{X_i}$ can be approximated using:

$$\boldsymbol{X_i} = \overline{\boldsymbol{X}} + \boldsymbol{Pb} \tag{2}$$

Where $\mathbf{P}$ is a matrix of unit eigenvectors of the covariance of deviations, and $\mathbf{b}$ is a vector of eigenvector weights (these are referred to as Model Parameters). By modifying $\mathbf{b}$, new instances of the model can be generated; if the elements of $\mathbf{b}$ are kept within some limits (typically $\pm 3\sigma_k$, where σ_k is the standard deviation of $\mathbf{b}_k$ over the training set) the corresponding model instances are plausible examples of the modeled objects. Since the columns of $\mathbf{P}$ are orthogonal $\mathbf{P^T P} = \mathbf{I}$, and (2) can be solved with respect to $\mathbf{b}$.

$$\boldsymbol{b} = \boldsymbol{P}^T\left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right) \tag{3}$$

Equation (3) can be used to transform examples to model parameters. Usually, the number of eigenvectors needed to describe most of the variability within a training set is much smaller than the original number of variables, allowing the model to approximate training examples using a small number of model parameters $b_1 \ldots b_t : t < N$.

The same method can be used to train both shape and gray-level models. For shape models, the variables $(x_{k\,i})$ are the coordinates of landmark points, and for gray-level models, the variables are based on gray-level intensities. For example, for flexible gray-level profile models variables $(x_{k\,i})$ may represent the absolute gray-level intensity at a specific point on each training profile.

We refer to both shape and gray-level models, as *flexible models* because they model the ways in which shapes or gray-level surfaces, respectively, are allowed to vary with respect to a mean value. The details have been described elsewhere [8], [10]. Important points to note are that

- in the shape models, the shapes of facial features and the spatial relationships between them are captured in a single model,
- although the models are flexible, they are still specific, they can vary only in ways encountered in the training set.

5 Modeling Shape

We have built a flexible shape model (or Point Distribution Model) [8], [10] representing the face using 152 points manually located on each of 160 training examples (eight examples from each of 20 individuals). Typical training examples, the mean shape, and the locations of the model points are shown in Figs. 2, 3, and 4, respectively. The model can accurately approximate the shape of any face in the training set using just 16 shape parameters; the effect of the first six is shown in Fig. 5. The first three parameters reflect variations in the 3D pose, the fourth and the sixth account for shape variation between different individuals, and the fifth changes the expression. In addition to the 16 model parameters, four 2D pose parameters are needed to define a model instance in the image plane: The x and y coordinates of the origin, a rotation angle, and a scaling factor. Given a set of model and 2D pose parameters, a face shape can be computed and projected into an image.

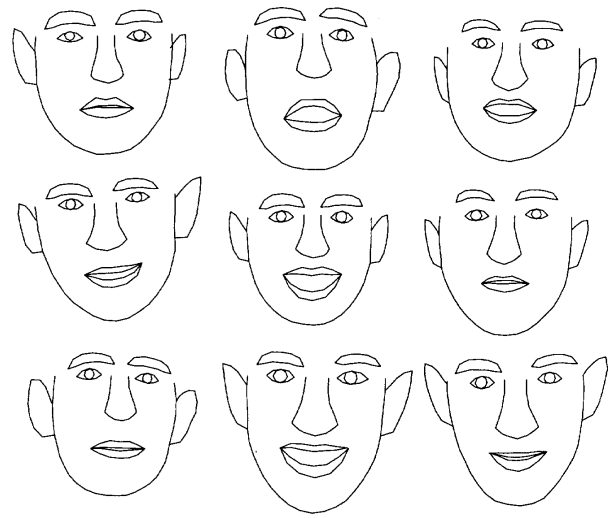


Fig. 2. Typical training shapes.

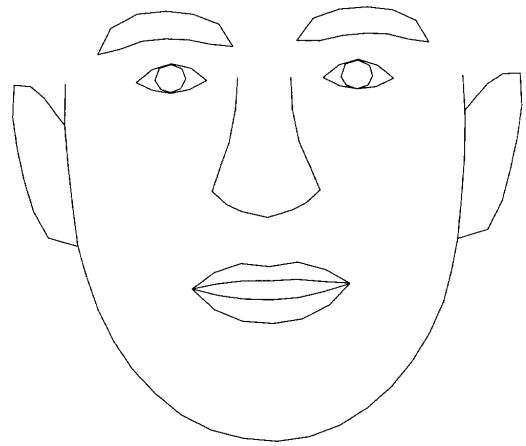


Fig. 3. The mean shape.

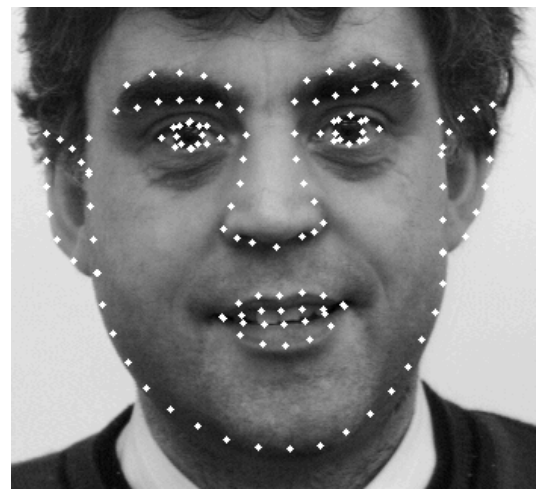


Fig. 4. Locations of model points on a training image.

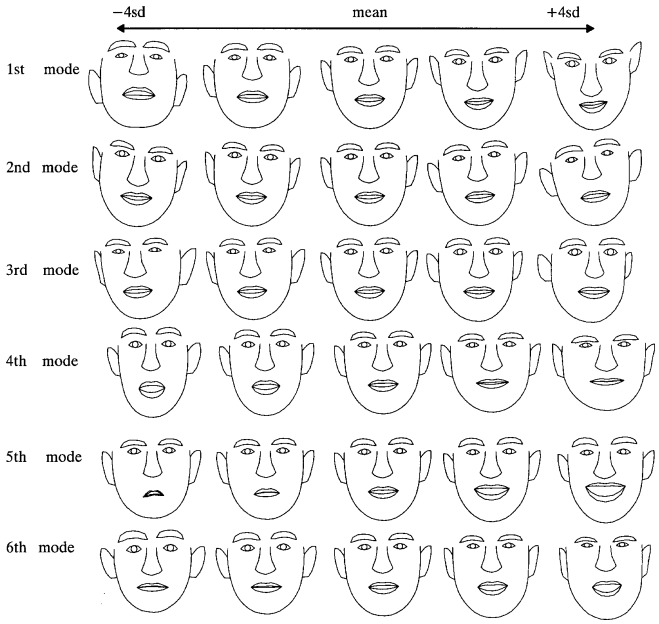


Fig. 5. The effect of the main modes of shape variation.

6 Modeling Shape-Free Gray-Level Appearance

6.1 Shape Normalization

We wish to model gray-level appearance independently of shape. To do this, we deform each face image to the mean shape in such a way that changes in gray-level intensities are kept to a minimum. For this purpose, we have used a technique developed by Bookstein [3], based on thin plate splines. This allows an image to be deformed so that a set of landmarks are moved to coincide with a set of target landmarks on the mean face in such a way that changes in the gray-level environment around each landmark are kept to minimum. We have used 14 landmarks to deform the face images. The position of these landmarks on a particular face shape and on the average face shape are shown in Fig. 6. All the landmarks are part of the shape model, thus, once the shape model has been fitted to an image, the landmarks are located trivially. Examples of face images before and after deformation are shown in Fig. 7.

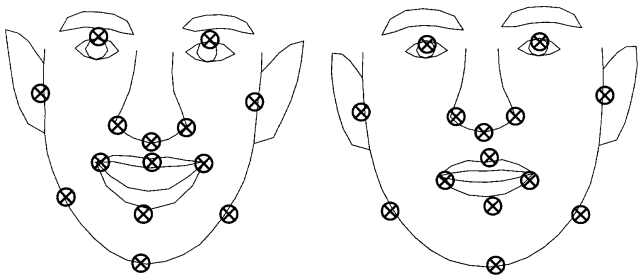


Fig. 6. The landmarks used for deforming face images.

6.2 Training the Flexible Model

Training images were deformed to the mean shape, and gray-level intensities within the face area were extracted, as

Fig. 7. Examples of original face images (left), and the respective shape-free face images (right).

Fig. 8. Examples of training shape-free patches.

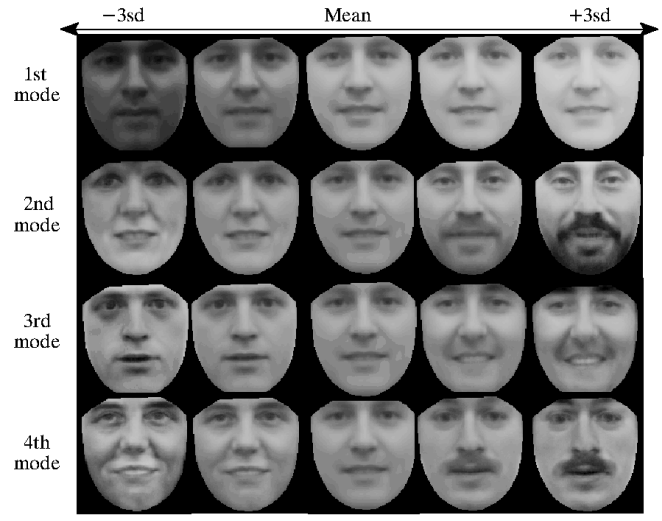


Fig. 9. The main modes of gray-level variation.

shown in Fig. 8. Each training example was represented by a vector containing the gray-level at each pixel in the patch (a total of 10,656 pixels). A flexible gray-level model for our database was generated; only 12 variables were needed to explain 95 percent of the variation in the training set. Each variable is responsible for a specific mode of variation, as shown in Fig. 9. For example, there are modes that control the lighting conditions (first mode), the addition of beards and moustaches (second and fourth modes), and the

change in expression (third mode). A problem with this model is that the first mode of variation represents 80 percent of the gray-level variation on its own so that other possible modes of variation are swamped. A second model, using gray-levels normalized with respect to the average intensity in the gray-level patch was also trained. This model needed 79 variables to explain 95 percent of the variability. For classification and reconstruction purposes the normalized gray-level model was used.

7 Modeling Local Gray-Level Appearance

We have described how we model the shape and overall gray-level appearance of face images. We also model the gray-level appearance in the vicinity of each shape model point, using a large number of local gray-level profile models. Modeling gray-level appearance locally can be important for face interpretation, because subtle but important localized effects can be swamped in the global shape-free model. Local models can also be used to achieve more robust interpretation in the presence of partial occlusion [23].

During training, shape model points were overlaid on the corresponding training images and gray-level profiles perpendicular to the boundary were extracted, as shown in Fig. 10. Because each shape model point should always correspond to the same facial feature, there is no need to apply a deformation algorithm. However, the appearance of the extracted gray-level profiles is dependent on the scale of the face in the image. To account for this, the length of training profiles was normalized with respect to the scale of the face. Each profile was represented by 10 gray-level samples, with the sampling interval varied to achieve length normalization (see Fig. 10). A flexible gray-level model was built for the profile at each model point; most of these models needed four model parameters to explain 95 percent of the variation in the training set.

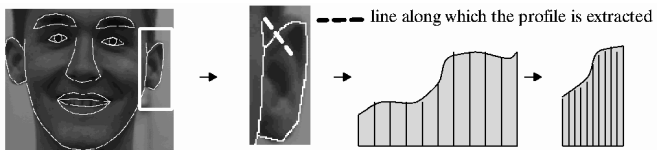


Fig. 10. Extraction of gray-level profile at a model point.

8 Calculating the Appearance Parameters

When a new face image is presented to the system, the whole set of appearance parameters can be calculated. This procedure is summarized in Fig. 11. The shape model is fitted automatically to the new face (the fitting procedure is described in Section 9) and the shape model parameters corresponding to the shape of the new face are computed. At each located shape model point, gray-level information from a profile perpendicular to the boundary is extracted, and the parameters of the local gray-level model are calculated. Based on the shape model fit, the landmarks used to deform to the mean face shape (see Fig. 6) are identified. The deformation is performed and shape-free gray-level model parameters are calculated. Together the three sets of model parameters constitute the full set of appearance parameters; in Sections 10 to 14, we describe how we use these parameters to code and interpret face images.

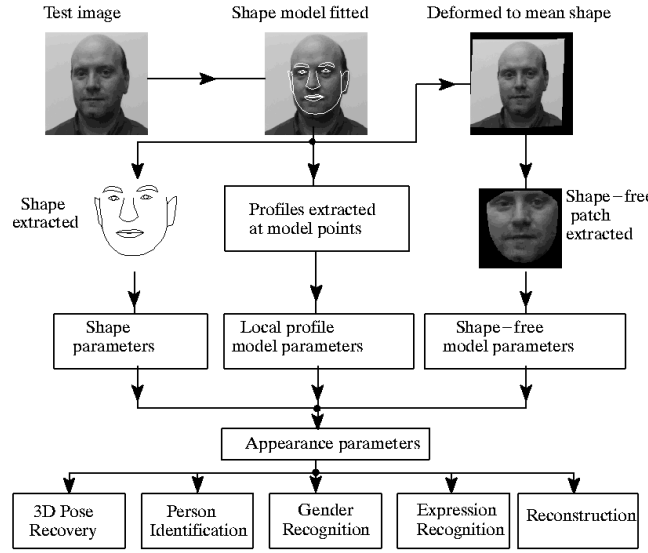


Fig. 11. Calculating the appearance parameters for a new face image.

9 Locating Facial Features

9.1 Overview

The shape model and local gray-level models described above can be used to automatically locate all the modeled features simultaneously. This is achieved using an ASM search [7], [8], [9], [10]. The mean shape model is placed in a given image and is allowed to interact dynamically until it fits to the data. Each iteration involves two main steps: Calculating a new suggested position for each model point based on matching the local gray-level models, followed by movement and deformation of the model in order to move each point as close as possible to the new preferred position. During this process, the shape model is only allowed to deform in ways which are consistent with the training set.

9.2 Calculating New Positions for the Model Points

At each model point, a gray-level profile perpendicular to the boundary is extracted, and a new preferred position for the point is selected along the profile. This is illustrated in Fig. 12. The gray-level model associated with the shape model point is scanned along the search profile (which is longer) to determine the position of best fit; details are given in [8], [9], [10]. This results in a new preferred position for each shape model point.

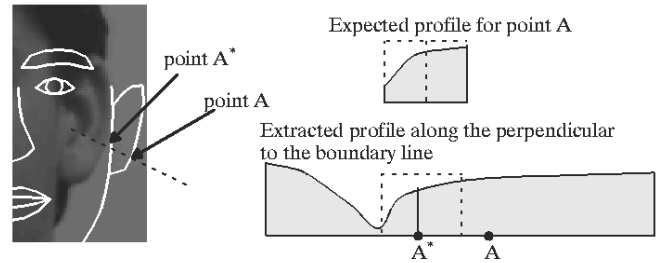


Fig. 12. Defining the new preferred position A* for a model point currently at A.

9.3 Deforming the Shape Model

The key feature of ASM search is that model points do not move individually to the new suggested positions. First, the pose parameters (i.e., translation, scaling and rotation) are adjusted to minimize the mean squared distance between model points and the suggested new positions. Next, the shape model is deformed to fit as closely as possible to the suggested points by modifying the shape parameters (**b** in (2)). The set of **b** values needed to give a least squares approximation to the new suggested shape can be obtained directly from (3). Shapes which are inconsistent with those in the training set can be avoided by constraining the values of **b** to lie within limits obtained from the training set. Further details are given in [7], [8], [10]. Examples of model fitting are shown in Fig. 13 which illustrates that the method is robust to 3D pose variation and occlusion. In practice, the fitting procedure is implemented as a multiresolution search [9]; this results in improved speed of execution and robustness.

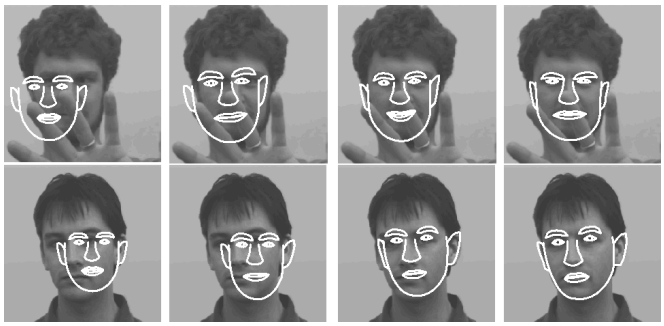


Fig. 13. Examples of the ASM fitting procedure.

9.4 Experimental Results

We have assessed the accuracy with which the 152 landmarks, shown in Fig. 4, can be located automatically in new images. For the experiment, we used a subset of 40 training images from our database to train a new shape model, and subsequently fitted this model to a different subset of 40 training images from the database. (For this experiment, we used images from the training set for testing, because, for these images, we already had manually located landmarks which we could use to assess the accuracy of the automatically located points). The model was initialized on each test image, with the mean shape scale 70 percent of the mean scale, displaced by ±20 pixels from the true position and rotated by ±12 degrees from the true orientation. The accuracy of point location at each iteration of a multiscale ASM search was assessed by calculating the mean Euclidean distance between model points and the curves defined by correct landmark points. The graph in Fig. 14 shows the results of this experiment averaged over 40 runs. Landmark points were located to an average accuracy of about three pixels within about 25 iterations. In its current form the model fitting procedure takes approximately two seconds on a Sun Sparc 20 workstation.

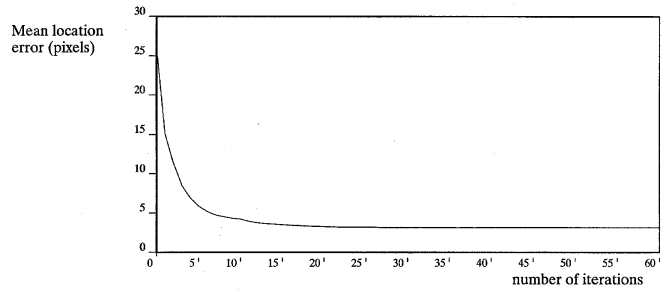


Fig. 14. Results in locating facial characteristics.

10 Tracking Coding and Reconstructing Faces

10.1 Method

Once the shape model has been fitted to a new face image, the shape-free and local gray-level model parameters can also be extracted, providing a complete description of the face area. We have shown how the shape model can be fitted to an individual face image. If a video sequence is used, the ASM search result for each frame can be used as the starting approximation for the next frame. Since only small changes in shape and position occur between frames, the shape model can be updated very rapidly. Fig. 15 shows an example of face tracking using the flexible shape model. At each frame, the shape and gray-level model parameters can be calculated and used as an extremely compact coding. To reconstruct a frame the shape-free gray-level parameters are used to generate the shape-free gray-level appearance, and the shape model parameters to define the shape of the face. The deformation algorithm described in Section 6.1 is applied again with the difference that now we deform from the average face shape to the shape corresponding to the shape parameters given.

Fig. 15. Face movement tracking using a flexible shape model.

10.2 Experimental Results

Fig. 16 shows two examples of coding and reconstructing face image sequences; the images in these sequences are new images of individuals who were in the training set. By using this method of coding, appearance variations caused by differences between individuals, changes in expression, and changes in head orientation, are reproduced accurately in the reconstruction. The total number of parameters needed for coding and reconstructing face images using this method, is 99 (16 shape parameters, four pose parameters, and 79 shape-free gray-level parameters).

Fig. 16. Examples of tracking and reconstruction of face image sequences (top row: originals, bottom rows: reconstructions).

We have also tested the approach by coding and reconstructing images of new individuals. The reconstructions obtained (see Fig. 17) are promising implying that the model can generalize its knowledge about the appearance of faces to unseen examples. However, it would be desirable to train the gray-level model of shape-free appearance using more examples in order to improve the quality of reconstructions for images of unseen individuals.

Fig. 17. Reconstruction of faces images of new individuals (top row: originals, bottom row: reconstructions).

For the face images used in training the model, no individual was wearing spectacles. As a result the model cannot reproduce faces wearing spectacles. This is demonstrated in Fig. 18, where the reconstruction of a partially occluded face is the corresponding occlusion-free face. However, if the occlusions are very severe the reconstructed faces can be significantly distorted (see the fourth example in Fig. 18).

Fig. 18. Reconstructing occluded face images (top row: originals, bottom row: reconstructions).

11 Recovering 3D Pose

11.1 Method

Once the shape model has been fitted to a new face, the recovered shape model parameters can be used to determine the 3D pose of the face. As shown in Fig. 5, the first (b_1) and third (b_3) shape model parameters are responsible for controlling the apparent changes in shape due to turning and nodding the head. 3D pose recovery can be based on the numerical values of these parameters, calculated for a new face outline. To calibrate the system, we captured a series of face images in which an individual was asked to rotate his head in both the vertical and horizontal direction, from -20 degrees to +20 degrees by looking at a number of grid points on a wall; each grid point corresponded to a known combination of horizontal and vertical rotation angles. We automatically fitted the shape model to these images and recorded the values of the first and third shape parameters. Plots of b1 and b3 against angles are shown in Fig. 19 and demonstrate approximately linear relationships for the range of angles considered. When a new face is presented the shape model is fitted, the resulting numerical values for b1 and b3 are recorded and the 3D pose angles are calculated based on the calibration graphs in Fig. 19.

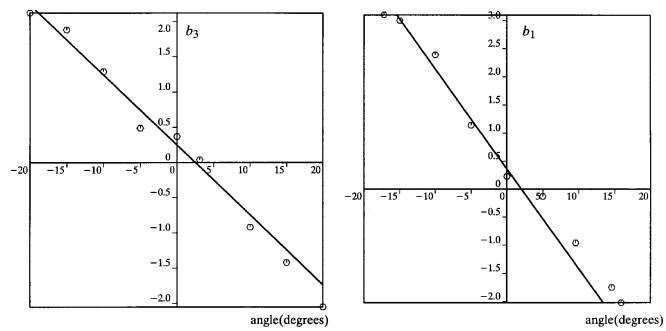


Fig. 19 Graphs of b_3 against the horizontal rotation angle (left), and b_1 against the vertical rotation angle.

11.2 Experimental Results

We tested the accuracy of 3D pose recovery on 30 new test images obtained in a similar manner to the calibration set. In the test set we included images of previously unseen individuals. The results obtained are summarized in Fig. 20 and show robust recovery of 3D pose angles. Fig. 21 shows examples of test images and the 3D pose angles computed.

12 Person Identification

12.1 Method

Once the shape model has been fitted to a new image, a full set of appearance parameters can be computed (see Fig. 11). Some of these are associated with differences between individuals, while others model changes in 3D pose, facial ex-

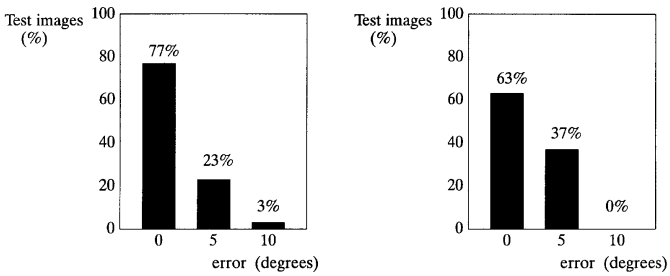
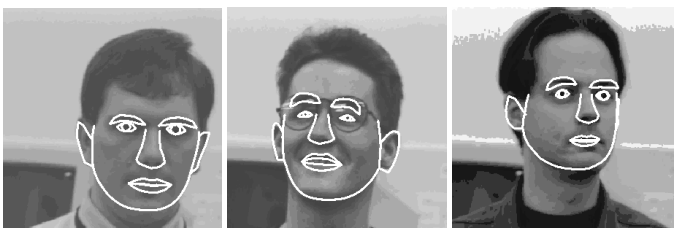


Fig. 20. Results for the calculation of the horizontal (left hand graph) and vertical (right hand graph) angles on test images.



Face rotated –20° in the horizontal plane and 5° in the vertical plane

Face rotated 20° in the horizontal plane and 10° in the vertical plane

Face rotated –10° in the horizontal plane and –10° in the vertical plane

Fig. 21. Examples of 3D pose recovery on test images.

pression, or lighting. We have developed a classification system for identifying the individual appearing in an image, irrespective of 3D pose, expression or lighting. The classifier was trained by computing the appearance parameters for all training images (10 images for each of the 30 individuals) in our database and establishing the distribution of appearance parameters for each individual. Simple discriminant analysis techniques [16], [28] were applied in order to enhance the effect of the inter-class (between person) variation parameters using the Mahalanobis distance(D_i) measure.

$$D_i^2 = \left(\boldsymbol{b} - \overline{\boldsymbol{b}_i}\right)^T \boldsymbol{C}^{-1} \left(\boldsymbol{b} - \overline{\boldsymbol{b}_i}\right) \quad (4)$$

Where **b** is a multivariate observation (a set of appearance parameters in this application), $\overline{\mathbf{b}_i}$ is the centroid of a multivariate distribution for a particular class i and **C** is the common covariance matrix for all classes. By using the common covariance matrix rather than covariance matrices computed for individual classes, we ensure that we get a good estimate despite the limited number of training examples available. A given multivariate observation is assigned to the class that minimizes the Mahalanobis distance between the observation and the centroid of that class. Since Mahalanobis distance uses covariance information, it has the ability to suppress the effect of parameters responsible for within-class variation. We have described elsewhere [26] how the interclass variation parameters can be explicitly isolated, using canonical discriminant analysis.

When a new face image is presented the shape model is fitted and the appearance parameters computed. The face is assigned to the class for which the Mahalanobis distance is minimized. Classification can be performed using the three types of appearance parameter (shape, local gray-level, and shape-free gray-level) individually, or in combination.

12.2 Experimental Results

We performed person identification experiments on the test and difficult test sets from our face database using various combinations of appearance parameters. The classification results are shown in Table 2. The timings quoted are for a SunSparc 20 workstation. The two methods involving gray-level information achieve much better results, than the method based on shape alone. These results are consistent with those reported by other researchers [5], [14], suggesting that the gray-level appearance of faces is much more important than shape for identification purposes. However, shape has an important role since classification rates improve significantly when shape information is combined with gray-level information. Gray-level profile models proved to be the least sensitive to occlusion. Any of the combinations of the measures produced good results. For real-time applications the method combining shape and gray-level profile information is the least computationally expensive, but if optimal classification accuracy is required the shape-free gray-level model parameters should also be included. It is very important to note that when all three methods are combined together the results obtained are significantly better than any other combination, both for the test and difficult test images. This implies that each method conveys important and unique classification information.

TABLE 2
RESULTS FOR THE FACE IDENTIFICATION EXPERIMENTS

		Normal test set (300 images)		Difficult test set (90 images)	
Method	Classification time	Correct classifications	Correct class within best three	Correct classifications	Correct class within best three
Shape model	2 sec.	50.3%	66.6%	15.6%	31.11%
Shape-free gray model	6 sec.	78.7%	87.33%	31.1%	53.3%
Local gray-level models	2 sec.	77.33%	89.7%	28.9%	57.8%
Shape + Shape – free model	6 sec.	85.3	93.3	34.4	57.7
Shape + Local models	2 sec	80.0%	90.3 %	34.4 %	66.7 %
All three methods	6 sec	92.0%	97.0%	48.9%	74.4%

For access control applications, it is important that face identification systems have the ability to reject faces which do not look similar to any of the database entries. We performed a preliminary experiment in which images were rejected if the minimum Mahalanobis distance was not within acceptable limits. The results of this experiment are shown in Fig. 22 and Fig. 23. When the threshold was set to 0.8, the classification rates for both the test and difficult test sets were 100 percent. In this case about 55 percent of the test images and 97 percent of the difficult test images were rejected since they were not sufficiently similar to any of the individuals in the training set, which is not surprising bearing in mind the appearance variations between training and test images in our database.

13 Gender Recognition

13.1 Method

Gender recognition can also be attempted in our framework, since some appearance parameters reflect inter-gender differences in facial appearance (for example, the second mode of variation in Fig. 9). We have investigated the use of shape model parameters and shape-free gray-level parameters both individually and in combination. During training we established the distributions of these parameters for the male and female subjects in the training set of face images. The classification procedure was similar to the one used for face identification.

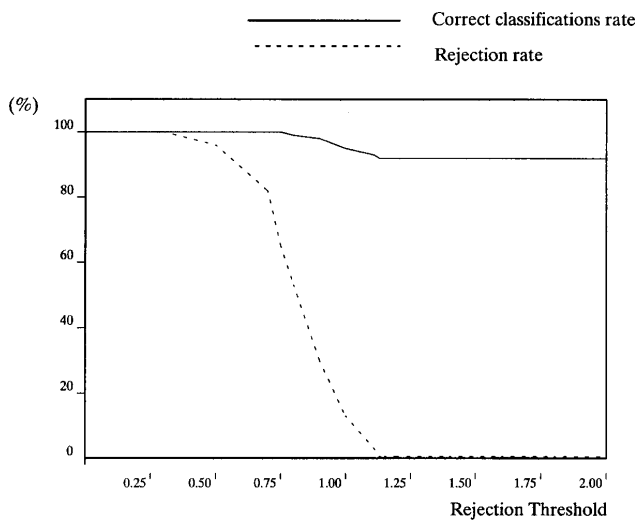


Fig. 22. Rejection and classification ratios for the normal test set.

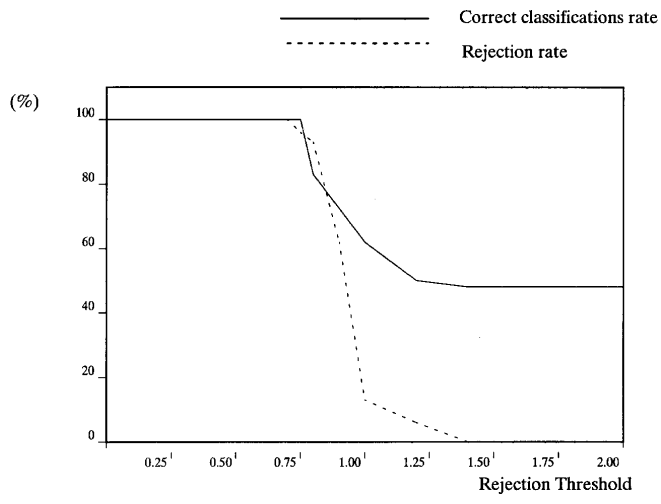


Fig. 23. Rejection and classification ratios for the difficult test set.

13.2 Experimental Results

In order to perform the experiment on images of unseen persons, we have trained the system using the training images from 20 individuals in our database and tested it using the test images for the remaining 10 individuals (100 test images). The results obtained are summarized in Table 3, and show that gray-level information is more important than shape information for gender recognition. The peak classification rate was 94 percent.

TABLE 3
Results for Gender Recognition

Method	Correct Classifications
Shape model	72%
Shape-free gray-level	94%
Shape + Shape-free model	86%

14 Expression Recognition

14.1 Method

Expression recognition, particularly from static images, is a difficult but interesting problem. It is known that even human observers often fail to agree in expression classification experiments [15]. We have addressed the problem by establishing the distribution of appearance parameters over a selected training set for each expression category so that the appearance parameters calculated for a new face image could be used for determining the expression. For this experiment we used shape and shape-free appearance parameters both individually and in combination.

14.2 Experimental Results

We asked five observers to classify the expression of each of our training and test face images using the seven psychologically recognized categories [15] shown in Fig. 24. For our subsequent experiments we used the images for which at least four of the observes agreed (139/300 training and 118/300 test images). Fig. 25 shows the reconstruction obtained from the centroid of the distribution for each expression; the reconstructed mean expressions are realistic. Because the face database used was not originally intended to be used for expression recognition experiments, it does not contain an adequate number of examples for all expressions. Few of the selected training images were categorized as afraid or disgusted; as a result the reconstructed mean expressions for afraid and disgusted look more like the subjects in the database who displayed those emotions.

The results for automatic classification are shown in Table 4. We also asked two more observers to classify the expressions in each of the 118 test images with which we

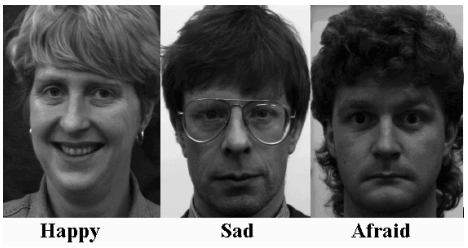


Fig. 24. Faces displaying the seven expressions used in the expression recognition experiment.

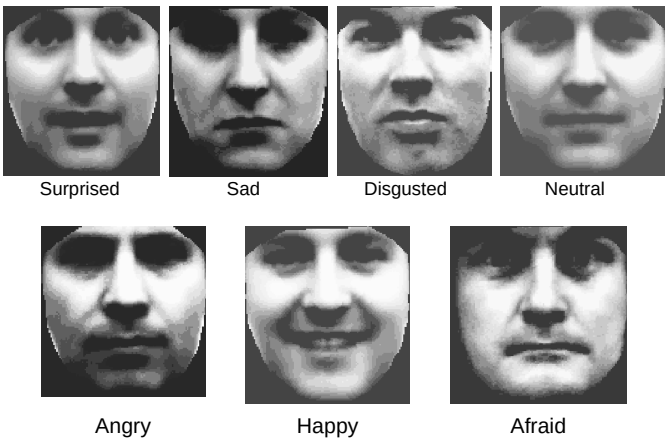


Fig. 25. The reconstructed mean expressions for our database.

tested our classification system. They achieved 80 percent and 83 percent correct classification. The peak classification rate of 74 percent obtained from the automatic system compares well with these results, implying that our methodology has the potential to be employed successfully in this application.

In the future, we plan to extend our preliminary work on expression recognition by performing experiments using a more suitable database. We are in the process of collecting a new database using actors to animate all seven fundamental expressions, so that we have an adequate number of examples for all expressions.

15 Conclusions

We have presented a system which can be used for locating facial features, coding, and reconstruction, recovering 3D pose, recognizing gender and expression, and identifying the individual in an image. Our results for most of the applications are promising, even though considerable variation in 3D pose, lighting, and expression were allowed, demonstrating the potential for use in real life applications. The distinctive feature of our system is that it can cope successfully with almost all aspects of face image processing, within a unified framework.

TABLE 4
RESULTS FOR EXPRESSION RECOGNITION

Method	Correct Classifications
Shape model	53%
Shape-free gray-level	74%
Shape + Shape-free model	70%

The face interpretation procedures described are fully automatic; errors for the classification experiments may be caused either by failure in locating landmarks accurately, or by failure of the classification algorithm. We do not distinguish between the two cases, since we believe that locating facial characteristics automatically is an important aspect of an integrated system.

Studies performed by other researchers have shown that certain facial areas are more important than others for specific applications. For example, the eyes and mouth are the most important areas for identification [4], [5], [30], and for expression recognition, the left side of the face is considered to be more important than the right [32]. We intend to perform classification experiments using subparts of the shape-free appearance of faces in order to establish the parts with the highest discriminating power for specific classification tasks.

Although the lighting conditions were not altered systematically during the capture of training images, there was some variability. As a result, test images captured with different lighting intensity and/or direction were recognized correctly. Ideally, we need to collect a very large training and test set, in which pose, expression, and lighting conditions are varied systematically for each of a large number of individuals. This would allow us to build a more complete model and characterize the performance of our system more thoroughly.

Our approach is generic, and can be easily adopted for different applications. For example, a similar method has been used for automatic interpretation of hand gestures in image sequences [1], [25]. For this application, flexible shape models were used for tracking hand and finger movements and classifying the gesture in each frame.

Acknowledgments

We would like to thank all members of our department for their help and advice, especially Dr. Andrew Hill, Dr. Peter Sozou, and Mr. Dave Cooper. We are grateful to all those who kindly volunteered to provide face images for our experiments. Discussions with Prof. J.F.W. Deakin and Dr. J.F. Whittaker from the Department of Psychiatry, Manchester Royal Infirmary, UK, were particularly useful for the expression recognition experiments. Also, we are very grateful to the observers who helped us with the expression recognition experiments.

Most of the work described in this paper was supported by a University of Manchester Research Studentship and an ORS award to Andreas Lanitis.

REFERENCES

[1] T.N. Ahmed, "A Model Based Hand Gesture Recognition System," MSc Thesis, Dept. of Computer Science, Univ. of Manchester, 1994.

[2] M. Bichsel and A. Pentland, "Automatic Interpretation of Human Head Movements," MIT Media Laboratory, Vision and Modeling Group, Technical Report No. 186, 1993.

[3] F.L Bookstein, "Principal Warps: Thin-Plate Splines and the Decomposition of Deformations," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 11, no. 6, pp. 567-585, 1989.

[4] V. Bruce, *Recognising Faces*. London : LEA Publishers, 1991.

[5] R. Brunelli and T. Poggio, "Face Recognition: Features versus Templates." *IEEE Trans. Pattern Analysis and Machine Intelligence,"* vol. 15, no. 10, pp. 1,042-1,052, 1993.

[6] A.F. Clark and M. Kokuer, "A Model-Based Codec With Potential For Deaf Communication," *Proc. 12th Int'l Conf. Pattern Recognition*, vol. 3, pp. 195-197. Los Alamitos, Calif.: IEEE CS Press, 1994.

[7] T.F. Cootes, C.J. Taylor, A. Lanitis, D.H. Cooper, and J. Graham, "Building and Using Flexible Models Incorporating Grey-Level Information," *Proc. Fourth Int'l Conf. Computer Vision*, pp. 242-246. Los Alamitos, Calif.: IEEE CS Press, 1993.

[8] T.F. Cootes, A. Hill, C.J. Taylor, and J. Haslam, "The Use of Active Shape Models For Locating Structures in Medical Images," *Image and Vision Computing*, vol. 12, no. 6, pp. 355-366, 1994.

[9] T.F. Cootes, C.J. Taylor, and A. Lanitis, "Multi-Resolution Search Using Active Shape Models," *Proc. 12th Int'l Conf. Pattern Recognition*, vol. 1, pp. 610-612. Los Alamitos, Calif.: IEEE CS Press, 1994.

[10] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham, "Active Shape Models—Their Training And Application," *Computer Vision Graphics and Image Understanding*, vol. 61, no. 1, pp. 38-59, 1995.

[11] G.W. Cottrell and M.K. Fleming, "Categorisation of Faces Using Unsupervised Feature Extraction," *Proc. Int'l Conf. Neural Networks*, San Diego, vol. 2, pp. 65-70, 1990.

[12] G.W. Cottrell and J. Metcalfe, "EMPATH: Face, Gender and Emotion Recognition Using Holons," *Advances in Neural Information Processing Systems*, vol. 3, R.P. Lippman, J. Moody, and D.S. Touretzky, eds., pp. 564-571. San Mateo, Calif.: Kaufmann, 1991.

[13] I. Craw, D. Tock, and A. Bennett, "Finding Face Features," *Proc. European Conf. Computer Vision*, G. Sandini, ed. Springer-Verlag, 1992.

[14] I. Craw and P. Cameron, "Face Recognition by Computer," *Proc. British Machine Vision Conference 1992*, pp. 489-507, David Hogg and Roger Boyle, eds. Springer Verlag, 1992.

[15] P. Ekman and W. Friesen, *Unmasking the Face, A Guide to Recognising Emotions From Facial Expressions.* Prentice Hall, 1975.

[16] J.H. Friedman, "Requlised Discriminant Analysis," *J. American Statistical Assn.*, vol. 84, no. 405, pp. 165-175, 1989.

[17] J. Fiser, I. Biederman, and E. Cooper, "To What Extent Can Matching Algorithms Based on Direct Outputs of Spatial Filters Account For Human Shape Recognition?," *Filters and Recognition*, Mar. 1993.

[18] A. Gee and R. Cipolla, "Estimating Gaze From a Single View of a Face," *Proc. 12th Int'l Conf. Pattern Recognition*, vol. 1, pp. 758-760. Los Alamitos, Calif.: IEEE CS Press, 1994.

[19] M. Kass, A. Witkin, D. Terzopoulos, "Snakes, Active Contour Models," *First Int'l Conf. Computer Vision*, pp. 259-268. Los Alamitos, Calif.: IEEE CS Press, 1987.

[20] M. Kirby and L. Sirovich, "Application of the Karhumen-Loeve Procedure for the Characterization of Human Faces," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 12, no. 1, pp. 103-108, 1990.

[21] M. Lades, J.C. Vorbruggen, J. Buhmann, J. Lange, C. Malsburg, R.P. Wurtz, and W. Konen, "Distortion Invariant Object Recognition in the Dynamic Link Architecture," *IEEE Trans. Computers*, vol. 42, no. 3, pp. 300-311, 1993.

[22] A. Lanitis, C.J. Taylor, and T.F. Cootes, "Automatic Tracking, Coding and Reconstruction of Human Faces, Using Flexible Appearance Models," *Electronics Letters*, vol. 30, no. 19, pp. 1,578-1,579, 1994.

[23] A. Lanitis, C.J. Taylor, and T.F. Cootes, "Automatic Identification of Human Faces Using Flexible Appearance Models," *Image and Vision Computing*, vol. 13, no. 5, pp. 393-401, 1995.

[24] A. Lanitis, C.J. Taylor, and T.F. Cootes, "A Unified Approach To Coding and Interpreting Face Images," *Proc. 5th Int'l Conf. Computer Vision*, pp. 368-373, Cambridge Mass. Los Alamitos, Calif.: IEEE CS Press, 1995.

[25] A. Lanitis, C.J. Taylor, T.F. Cootes, and T. Ahmed, "Automatic Interpretation of Human Faces and Hand Gestures Using Flexible Templates," *Proc. Int'l Workshop Face and Gesture Recognition*, pp. 98-103, M. Bichsel, ed., Zurich, Switzerland, 1995.

[26] A. Lanitis, C.J. Taylor, T. Ahmed, and T.F.Cootes, "Classifying Variable Objects Using a Flexible Shape Model," *Proc. Fifth Int'l Conf. Image Processing and its Applications*, pp. 70-74, Edinburg, UK, 1995.

[27] H. Li, P. Roivainen, and R. Forchheimer, "3-D Motion Estimation in Model-Based Facial Image Coding," *IEEE Trans. Pattern Analysis and Machine Intelligence* , vol. 15, no. 6, pp. 545-555, 1993.

[28] B.F.J. Manly, *Multivariate Statistical Methods, A Primer.* Chapman and Hall, 1986.

[29] Y. Moses, Y. Adini, and S. Ulman, "Face Recognition: The Problem of Compensating For Changes in Illumination Direction," *Proc. European Conf. Computer Vision*, vol. 1, pp. 286-296, J. Eklundh, ed. Springer Verlag, 1994.

[30] A. Pentland, B. Moghaddam, T. Starner, O. Oliyide, and M. Turk, "View-Based and Modular Eigenspaces for Face Recognition," M.I.T Media Laboratory, Perceptual Computing Section, Technical Report No. 245, 1994.

[31] M. Rydfalk, "CANDICE, A Parametrized Face," Technical Report, Linkoping Univ., Dept. of Electrical Eng., S-581 83 Linkoping, Sweden, 1987.

[32] H.A. Sackeim, R. Gur, and M. Saucy, "Emotions Are Expressed More Intensively on The Left Side of The Face," Science, vol. 202, pp. 434-435, 1978.

[33] D. Terzopoulos and K. Waters, "Analysis and Synthesis of Facial Image Sequences Using Physical and Anatomical Models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, no. 6, pp. 569-579, 1993.

[34] D. Tock and I. Craw, "Blink Rate Monitoring For a Driver Awareness System," *Proc. British Machine Vision Conf. 1992*, pp. 518-527, D. Hogg and R. Boyle, eds. Springer Verlag, 1992.

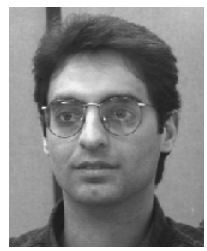
[35] A. Tsukamoto, C. Lee, and S. Tsuji, "Detection and Pose Estimation of Human Faces with Synthesized Image Models," *Proc. 12th Int'l Conf. Pattern Recognition*, vol. 1, pp. 754-757. Los Alamitos, Calif.: IEEE CS Press, 1994.

[36] M. Turk and A. Pentland, "Eigenfaces for Recognition," *J. Cognitive Neuroscience*, vol. 3, no. 1, pp. 71-86, 1991.

[37] J.B. Waite and W.J. Welsh, "An Application of Active Contour Models to Head Boundary Location," *Proc. First British Machine Vision Conference*, pp. 407-412, 1990.

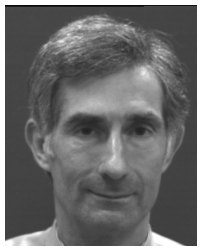
[38] Y. Yacoob and L. Davis, "Recognising Facial Expressions by Spatio-Temporal Analysis," *Proc. 12th Int'l Conf. Pattern Recognition*, vol. 1, pp. 747-749. Los Alamitos, Calif.: IEEE CS Press, 1994.

[39] A.L. Yuille, D.S. Cohen, and P. Halliman, "Feature Extraction From Faces Using Deformable Templates," *Int'l J. Computer Vision* vol. 8, pp. 104-109, 1992.

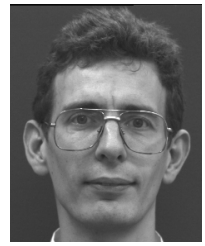


Andreas Lanitis received a BEng degree, with honors, in electronic and electrical engineering, in 1991, and a PhD degree in image processing and computer vision, in 1995, both from the University of Manchester, England. Subsequently, he was employed as a research associate in the Department of Medical Biophysics, University of Manchester. Currently he is assistant professor in the Department of Computer Science, Cyprus College, Cyprus.

His research interests include face image interpretation and coding, forensic applications of face image processing, modeling and recognizing variable objects and pattern recognition.



Chris J. Taylor received his BSc in physics, and a PhD in computer image analysis from the University of Manchester, England, in 1967 and 1972, respectively. He is currently a professor of medical biophysics and computer science, and director of the Wolfson Image Analysis Unit at the University of Manchester. Dr. Taylor's long term interest is in practical applications of computer vision in both medicine and industry. From 1974 to 1986, he was involved in developing high performance hardware architectures and model-based algorithms on which commercial systems for medical image analysis, environmental monitoring, and industrial inspection were based. His current research interests are in the areas of model-based vision (with particular emphasis on modeling variable objects), medical image interpretation, user-configurable inspection systems, and vision-based human-computer interaction, including face and gesture recognition.



Timothy F. Cootes received the BSc degree with honors in mathematics and physics from Exeter University, England, in 1986, and a PhD in Eng. from Sheffield City Polytechnic, in 1991. He obtained a postdoctoral fellowship from SERC in 1993, and an advanced fellowship from EPSRC in 1995. His research interests include statistical models of shape and appearance variation, and their applications to industrial and medical computer vision problems.