



# Recovering facial pose with the EM algorithm

Kwang Nam Choi<sup>1</sup>, Marco Carcassoni, Edwin R. Hancock\*

*Department of Computer Science, University of York, York, YO10 5DD, UK*

Received 24 August 2000; received in revised form 28 December 2000; accepted 12 January 2001

## Abstract

This paper describes how 3D facial pose may be estimated by fitting a template to 2D feature locations. The fitting process is realised as projecting the control points of a 3D template onto the 2D feature locations under orthographic projection. The parameters of the orthographic projection are iteratively estimated using the EM algorithm. The method is evaluated on both contrived data with known ground-truth together with some more naturalistic imagery. These experiments reveal that under favourable conditions the algorithm can estimate facial pitch to within  $3^\circ$ . © 2002 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.

*Keywords:* Facial pose estimation; Facial feature detection; EM algorithm

## 1. Introduction

Facial pose estimation is a key task for many practical computer vision applications. Specific examples include visual surveillance, camera assisted user interfaces [1] and user identification or verification [2]. In essence, the problem revolves around the fitting of a generic 3D template to segmented facial features located in a 2D image. The complexity of the problem depends critically on whether or not the features are labelled, i.e. whether the model-data correspondences are known a priori. In the case of labelled data with known correspondences, then the complexity of the search-space is considerably reduced. For unlabelled data with unknown feature correspondences, then great care must be taken to render the registration process computationally tractable. Once the template has been fitted to the feature data, then 3D pose parameters may be used to manipulate the face. For instance Avidan and Shashua have used such information for view synthesis [3]. Viewed in this way pose

estimation may be regarded as an essential pre-requisite for detailed facial verification.

There have been many attempts at efficiently recovering the 3D facial pose. Most of these use domain specific cues to limit the search-space of the 3D model. Typically, the generic facial template must be translated, scaled and subjected to Eulerian rotation. One of the most powerful cues is to use the baseline of the eyes to estimate the gaze direction [4]. In this way the tilt-direction may be determined prior to rotation estimation. Based on the known ratio of the inter-eye separation and the distance to other axial features such as the tip of the nose or the lips, the rotation angle may also be estimated. In fact, the idea of using domain-specific cues to restrict the search-space is quite generic and has been used in a number of 3D object registration applications. One notable example is the fitting of 3D models to 2D images of vehicles [5].

### 1.1. Related literature

The facial pose estimation problem can be regarded as having two ingredients. The first of these is a means of locating facial features. The second is the use of these features to estimate pose angles. In this subsection we review the literature in these two areas.

\*Corresponding author. Tel.: +44-1904-43-3374; fax: +44-1904-43-2767.

*E-mail address:* erh@cs.york.ac.uk (E.R. Hancock).

<sup>1</sup> The author is now with Information and Telecommunication Research Institute, Chung-Ang University, Seoul 156-756, South Korea.

The first step that must be undertaken is to ascertain whether a face is present in a scene and to locate its salient features. The detection of faces can be effected using a number of different techniques. For instance Sung and Poggio [6] used example-based learning, Tsukamoto et al. [7] use a recognition by synthesis method, Ng and Gong [8] use a support vector machine, while Takacs and Wechsler use an iconic filter bank [9]. Once the face has been detected then feature localisation can be attempted. Considerable effort has been devoted to the detection of facial features. For instance Chow and Li [10] have used a Hough-based template fitting method which gives success rates greater than 80% for the eye and the mouth. Xie et al. [11] have reported a similar idea which uses energy minimisation methods for template fitting. Wu et al. [12] have used a relational model to extract facial features from colour images. Sobottka and Pitas [13] have used the topographic relief structure of colour and intensity images to identify the positions of the eye and the mouth. Takacs and Wechsler [9,14] have used trainable retinal filter banks implemented via self-organising feature maps (SOFM) for detecting faces and facial landmarks. Bala et al. [15] used genetic algorithms (GAs) for feature localisation and face identification. Pinto and Sossa [16] also use genetic algorithms for detecting the eyes and the mouth. Reisfeld and Yeshurun [17] have localised the eyes and mouth in a face using a generalised symmetry operator.

The current research literature on face representation and recognition can be divided into that which adopts a feature based approach and that which adopts a template based approach. This latter method includes the popular eigenspace approach [18]. Both approaches have difficulties with shape normalisation and photometric variation, which can have serious effects on the success rate for recognition. The template based approach was introduced by Yuille et al. [19] and extensively developed by Cootes et al. [20]. Yuille et al. [19] proposed a method for detecting and describing facial features using hand crafted deformable templates. Cootes et al. [20] have developed a method for automatically building models by learning patterns of variability from a training set of correctly annotated images. There are also many examples of the use of active contours for facial representation [21,22]. Concrete examples of the feature-based approach are provided by the eigenface method of Turk and Pentland [18] and the iconic method of Rao and Ballard [23] which uses basis functions for facial feature representation. Pentland and Moghaddam [24] have shown how the eigenface method can be extended to accommodate variable pose. Brunelli and Poggio [25] compared the performance of the feature and the template based approaches, and showed how the two methods could be combined to give enhanced performance. Craw et al. have used a “shape-free” face template representation together with principal components analysis to extract facial features. Von der Malsburg et al. use a higher-level relational description, referred to as

a bunch-graph, which uses columns of Gabor filter outputs to represent facial features [26]. Several authors have used colour and range-data to enhance the facial feature representation [27–30].

Once features are in hand then they may be used for facial pose estimation. For example Gee and Cipolla [31] estimated the direction of gaze from a single view of a face using the positions of the eyes. The method requires very few image measurements and can be run in real-time. Hattori et al. [30] have achieved facial pose estimation using a much richer set of facial features derived from colour and range images. Wiskott et al. [26] have shown how the bunch graph representation can be used to recover pose information using an elastic graph as a relational model. Several authors have couched the pose estimation problem as one of learning from examples. Mckenna et al. [32] have used principal component analysis (PCA) to examine the distribution of face pose without using facial features. Huang et al. [33,34] has used support vector machines (SVMs). Romdhani et al. [35] proposed a multi-view nonlinear active shape model [36] that uses 2D view-dependent contextual constraints without explicit reference to 3D object geometry of the head of a human figure.

## 1.2. Motivation

The observation underpinning this paper is that although specific constraints can be effectively used to restrict the search process, the underlying statistical methodology employed in the registration process is extremely limited. The aim of the work reported here is to exploit the framework of the expectation–maximisation algorithm of Dempster et al. [37] to recover the 3D pose parameters. The method is initialised using constraints provided by the location of the bilateral symmetry axis of the face and the orientation of the line connecting the two eyes. Our motivation in adopting the EM algorithm as a registration engine is provided by recent in-house work where Hancock and his co-workers have successfully matched both line-templates [38] and 3D perspective models [39]. Here we commence by constructing a generic 3D template of the facial features. The template is quite simple. It assumes that the eyes, chin and lip are approximately co-planar and that the tip of the nose resides at some significant height above the plane. The eyes are assumed to be symmetrically placed either side of the axis defined by the nose-tip and the lips.

Our aim is to develop a template registration process which can operate effectively when the set of feature correspondences is not known a priori. In other words, we are dealing with unlabelled feature points. This lack of correspondence information can be regarded as missing data. In keeping with the philosophy of the EM algorithm we construct a mixture model over the set of missing correspondences between the 2D facial features and the projections of the 3D template features. By assuming a Gaussian model for the registration errors, the template has freedom to deform

under both uncertainties in the positions of the feature points due to inaccuracies in the template model together with the intrinsic variability of natural faces. The parameters underpinning our model are the six degrees of freedom of the orthographic projection. These are the two translation parameters on the image plane, an overall object scale together with the three Euler angles for the bary-centric (object-centred) model rotation. We reduce the parametric complexity of the 3D template registration process by centring and aligning the template at a fixed point on the bilateral facial symmetry axis. This removes the two degrees of freedom associated with two template translation parameters on the image plane. In order to remove the degree of freedom associated with scale parameter, we normalise the distances between facial features. Specifically, we unitise the distance between the bridge of the nose and the centre of the chin. Then we adjust the remaining inter-feature distances accordingly.

The outline of this paper is as follows. In Section 2 we outline the geometry of our 3D facial template and explain how it is projected onto the 2D image data. Section 3 reviews the EM algorithm and explains how it may be used to estimate the parameters of orthographic projection. In Section 4 we describe the facial feature localisation method used in our pose estimation experiments. Section 5 contains details of the evaluation of the pose estimation method and provides an algorithm sensitivity analysis. Finally, Section 6 offers some conclusions.

**2. Geometric model**

Our basic aim is to align the control points in a 3D facial template against a set of 2D facial feature locations. The template is constructed as follows. We commence by assuming that the left and right eyes, the lips and the chin are coplanar. We describe how these features are detected in more detail in Section 4. These planar features are symmetric about the bilateral symmetry axis of the face which is defined by the centre-points of the lip and the chin. The tip of the nose (N in Fig. 1) is assumed to be elevated at some height  $h$  above the plane and to fall on the perpendicular plane through facial symmetry axis. The basic geometry of the template is shown in Fig. 1. The feature points used to define the eyes are the outermost extremities of the eyes (E1 and E2 in Fig. 1). We also use the the centre point at the bridge of the nose (B in Fig. 1), and the centre-points of the lip (L in Fig. 1) and chin (C in Fig. 1) as feature-points placed to define the bilateral symmetry axis of the face. We place the origin of co-ordinates for the facial template at the centre-point of the bridge of the nose (B in Fig. 1).

The projection of this 3D template onto the locations of the 2D facial feature points has six degrees of freedom. These correspond to the two translation parameters on the 2D image plane, the overall isotropic model scale together with the three Euler angles that define the 3D rotation of the

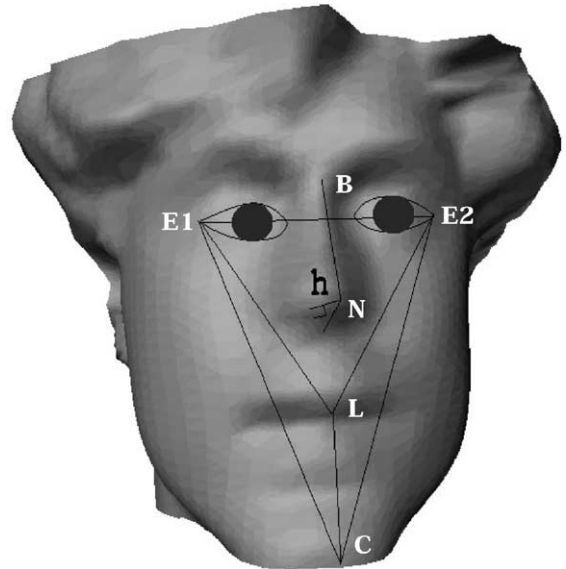


Fig. 1. The basic geometry of the face template.

model points. However, the parametric complexity of the projection can be simplified using constraints provided by the 2D geometry of the feature points. For instance the direction on the bilateral facial symmetry axis is easily computed by finding the perpendicular bisector of the line connecting the centres of the eyes.

The 3D template control points are represented by co-ordinate vectors  $\underline{v}_j = (x_j, y_j, z_j)^T$ , where the index  $j$  is drawn from the set of facial feature labels  $\mathcal{M}$ . The available facial features are represented by 2D co-ordinate vectors  $\underline{u}_i = (x_i, y_i)^T$  whose index  $i$  is drawn from the set of data-items  $\mathcal{D}$ . We represent the projection of the template control points into the image co-ordinate system in the following manner:

$$\underline{v}_j^{proj} = sUR_1(\phi)R_2(\psi)R_3(\theta)\underline{v}_j - X_0, \tag{1}$$

where  $s$  is the overall model scale parameter and  $X_0 = (x_0, y_0)^T$  is the translation of the origin in the image co-ordinate system. The matrices  $R_1(\phi)$ ,  $R_2(\psi)$  and  $R_3(\theta)$  represent Euler rotations of the model about its bary-centre at the origin. The  $3 \times 2$  projection matrix

$$U = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

selects the two  $x$ - $y$  components from the three  $x$ - $y$ - $z$  components of the transformed template control points. The Euler co-ordinate transformation is shown in Fig. 2.

The sequence of Euler rotations is defined as follows. The first step is to rotate the template about the normal to the facial-plane by an angle  $\theta$ . Recall that in our template, this plane is defined by the eyes, lip and chin. The net effect of this rotation about the  $z$ -axis is to tilt the head to the left or the right. In other words, it rotates the bilateral axis of facial

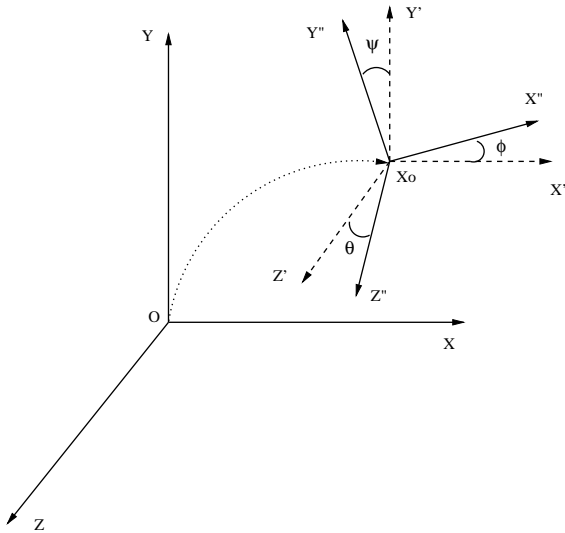


Fig. 2. The Euler co-ordinate transformation.

symmetry by an angle  $\theta$  in the image plane. The rotation matrix is given by

$$R_3(\theta) = \begin{pmatrix} \cos \theta & \sin \theta & 0 \\ -\sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{2}$$

The next step is to rotate by an angle  $\psi$  about the  $y$ -axis of the template. In our representation, the  $y$ -axis is parallel to the bilateral symmetry axis of the face and passes through the bary-centre of the template. The corresponding rotation matrix is given by

$$R_2(\psi) = \begin{pmatrix} \cos \psi & 0 & -\sin \psi \\ 0 & 1 & 0 \\ \sin \psi & 0 & \cos \psi \end{pmatrix}. \tag{3}$$

Finally, there is a rotation about the new template normal,  $x$ -axis, by an angle  $\phi$ . The matrix representation of this rotation is

$$R_1(\phi) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos \phi & \sin \phi \\ 0 & -\sin \phi & \cos \phi \end{pmatrix}. \tag{4}$$

The overall rotation is represented using the matrix

$$\Phi = \begin{pmatrix} \Phi_{11} & \Phi_{12} & \Phi_{13} \\ \Phi_{21} & \Phi_{22} & \Phi_{23} \\ \Phi_{31} & \Phi_{32} & \Phi_{33} \end{pmatrix}. \tag{5}$$

The three rotation angles are related to the elements of this matrix in the following manner:

$$\psi = \arcsin(-\Phi_{13}),$$

$$\begin{aligned} \phi &= \arcsin\left(\frac{\Phi_{12}}{\sqrt{1-\Phi_{13}^2}}\right), \\ \theta &= \arccos\left(\frac{\Phi_{33}}{\sqrt{1-\Phi_{13}^2}}\right). \end{aligned} \tag{6}$$

The parametric complexity of the projection can be reduced using some simple constraints provided by the geometry of the facial feature points on the 2D image plane. In the first instance, we can remove the translational degrees of freedom by placing the origin of the template co-ordinate system at a salient point. We take this point to be at the bridge of the nose (B in Fig. 1). This point is found by intersecting the line joining the eyes and the extension of lip-chin line. Once the origin has been established, we can scale the positions of the image features. We remove the scale parameter by normalising the inter-feature distances using the distance from the origin at the bridge of the nose (B in Fig. 1) to the centre point of the chin (C in Fig. 1). If  $X_0$  is the location of the bridge of the nose and  $s$  is the length of line  $BC$ , then we translate and scale the original feature points  $\underline{u}_i, \forall i \in \mathcal{D}$  to give the transformed set of data-point co-ordinate vectors

$$\underline{w}_i = \frac{1}{s}(\underline{u}_i - X_0).$$

Once this transformation has been performed then we aim to find the rotation matrix  $\Phi$  which best aligns them with the projected model-point locations  $\underline{v}_j^{proj} \cong U\Phi\underline{v}_j$ . the orthographic projection can be viewed as rotating the template by the angles  $\phi, \psi$  and  $\theta$ . In the next section, we explain how the resulting three degrees of freedom facial template may be registered by using the EM algorithm to iteratively estimate the parameters  $(\phi, \psi, \theta)^T$ . In our experimental section, we show results only for the angle  $\psi$ , because the change in this angle determines the most often encountered change of pose.

### 3. Registration process

In this Section we detail our model registration process and describe how the underlying set of transformation parameters can be recovered using the EM algorithm. The EM algorithm was first introduced by Dempster et al. as a means of finding maximum-likelihood solutions to problems involving incomplete data [37]. The algorithm has two stages. The expectation step involves estimating the a posteriori probabilities of the missing data using a mixture distribution defined over the current parameter values. The maximisation step involves computing new parameter values that optimise the expected value of the data log-likelihood. This two-stage process is iterated to convergence. Although the EM algorithm has been exploited in the recovery of object pose by Hornegger and Nieman [40], the main contribution of this paper is to demonstrate the effectiveness of the algorithm in matching a generic facial template to poorly localised feature-points.

### 3.1. Expected log-likelihood

Basic to our philosophy of exploiting the EM algorithm is the idea that every facial feature-point can, in principle, associate to each of the points in the 3D model template with some a posteriori probability. This modelling ingredient is naturally incorporated into the fitting process by developing a mixture model over the space of potential matching assignments which represent the “missing data” in our application. The expectation step of the EM algorithm provides an iterative framework for computing the a posteriori matching probabilities using Gaussian mixtures defined over a set of transformation parameters.

The EM algorithm commences by considering the conditional likelihood for the normalised 2D facial feature locations  $\underline{w}_i$  given the current set of transformation parameters,  $\Phi^{(n)}$ . The algorithm builds on the assumption that the individual data items are conditionally independent of one-another given the current parameter estimates, i.e.

$$p(\mathbf{w}|\Phi^{(n)}) = \prod_{i \in \mathcal{D}} p(\underline{w}_i|\Phi^{(n)}). \quad (7)$$

Each of the component densities appearing in the above factorisation is represented by a mixture distribution defined over a set of putative model-data associations

$$p(\underline{w}_i|\Phi^{(n)}) = \sum_{j \in \mathcal{M}} p(\underline{w}_i|\underline{v}_j, \Phi^{(n)})P(\underline{v}_j|\Phi^{(n)}). \quad (8)$$

The ingredients of the above mixture density are the component conditional measurement densities  $p(\underline{w}_i|\underline{v}_j, \Phi^{(n)})$  and the mixing proportions  $P(\underline{v}_j|\Phi^{(n)})$ . The conditional measurement densities represent the likelihood that the 2D facial feature location  $\underline{w}_i$  originates from the 3D template control point indexed  $j$  under the prevailing set of transformation parameters  $\Phi^{(n)}$ . We use the shorthand notation  $\alpha_j^{(n)} = P(\underline{v}_j|\Phi^{(n)})$  to denote the mixing proportions. These quantities provide a natural mechanism for assessing the significance of the individual template control points in explaining the current data-likelihood.

Conventionally, maximum-likelihood parameters are estimated using the complete log-likelihood for the available data

$$L(\Phi^{(n)}, \mathbf{w}) = \sum_{i \in \mathcal{D}} \ln p(\underline{w}_i|\Phi^{(n)}). \quad (9)$$

In the case where the conditional measurement densities are univariate Gaussian, then maximising the complete likelihood function corresponds to solving a system of least-squares equations for the transformation parameters. By contrast, the expectation step of the EM algorithm is aimed at estimating the log-likelihood function when the data under consideration is incomplete. In our 3D template-matching example this incompleteness is a consequence of the fact that we do not know how to associate feature tokens in the image and their counterparts in the

3D face template. In other words we need to compute the expected value of the data log-likelihood over the space of potential correspondence matches. In fact, it was Dempster et al. [37], who observed that maximising the weighted log-likelihood was equivalent to maximising the conditional expectation of the log-likelihood for a new parameter set given an old parameter set. For our matching problem, maximisation of the expectation of the conditional likelihood, i.e.  $E[L(\Phi^{(n+1)}, \mathbf{w})|\Phi^{(n)}, \mathbf{w}]$ , is equivalent to maximising the weighted log-likelihood function

$$Q(\Phi^{(n+1)}|\Phi^{(n)}) = \sum_{i \in \mathcal{D}} \sum_{j \in \mathcal{M}} P(\underline{v}_j|\underline{w}_i, \Phi^{(n)}) \ln p(\underline{w}_i|\underline{v}_j, \Phi^{(n+1)}). \quad (10)$$

### 3.2. Expectation

The a posteriori probabilities  $P(\underline{v}_j|\underline{w}_i, \Phi^{(n)})$  play the role of matching weights in the expected likelihood. We interpret these weights as representing the probability of match between the facial feature point indexed  $i$  and the template control-point indexed  $j$ . In other words, they represent model-datum affinities. Using the Bayes rule, we can re-write the a posteriori matching probabilities in terms of the components of the conditional measurement densities appearing in the mixture model in Eq. (8)

$$P(\underline{v}_j|\underline{w}_i, \Phi^{(n)}) = \frac{\alpha_j^{(n)} p(\underline{w}_i|\underline{v}_j, \Phi^{(n)})}{\sum_{j' \in \mathcal{M}} \alpha_{j'}^{(n)} p(\underline{w}_i|\underline{v}_{j'}, \Phi^{(n)})}. \quad (11)$$

The mixing proportions are computed by averaging the a posteriori probabilities over the set of facial feature points, i.e.

$$\alpha_j^{(n+1)} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} P(\underline{v}_j|\underline{w}_i, \Phi^{(n)}). \quad (12)$$

In order to proceed with the development of the facial template registration process we require a model for the conditional measurement densities, i.e.  $p(\underline{w}_i|\underline{v}_j, \Phi^{(n)})$ . Here we assume that the required model can be specified in terms of a multivariate Gaussian distribution. The random variables appearing in these distributions are the error residuals for the 2D position predictions of the  $j$ th template point delivered by the current estimated transformation parameters. Accordingly we write

$$p(\underline{w}_i|\underline{v}_j, \Phi^{(n)}) = \frac{1}{(2\pi)^{3/2}\sigma} \times \exp \left[ -\frac{1}{2\sigma^2} (\underline{w}_i - U\Phi^{(n)}\underline{v}_j)^T (\underline{w}_i - U\Phi^{(n)}\underline{v}_j) \right]. \quad (13)$$

According to this model, we assume that the vector of error-residuals  $\varepsilon_{i,j}(\Phi^{(n)}) = \underline{w}_i - U\Phi^{(n)}\underline{v}_j$  between the components of the projected position of the  $j$ th template point position  $U\Phi^{(n)}\underline{v}_j$  and the transformed location of the  $i$ th transformed image feature-point, i.e.  $\underline{w}_i = \frac{1}{s}(\underline{u}_i - X_0)$ , is governed

by an isotropic Gaussian distribution. The quantity  $\sigma$  is the variance of the error-residuals and is given by

$$\sigma^2 = \frac{\sum_{i \in \mathcal{D}} \sum_{j \in \mathcal{M}} P(\underline{v}_j | \underline{w}_i, \Phi^{(n)}) (\underline{w}_i - U\Phi^{(n)}\underline{v}_j)^T (\underline{w}_i - U\Phi^{(n)}\underline{v}_j)}{\sum_{i \in \mathcal{D}} \sum_{j \in \mathcal{M}} P(\underline{v}_j | \underline{w}_i, \Phi^{(n)})}. \quad (14)$$

With these ingredients, and using the shorthand notation  $q_{i,j}^{(n)} = P(\underline{v}_j | \underline{w}_i, \Phi^{(n)})$  for the a posteriori matching probabilities, the expectation step of the EM algorithm simply reduces to computing the weighted squared error criterion

$$\begin{aligned} Q'(\Phi^{(n+1)} | \Phi^{(n)}) \\ = -\frac{1}{2} \sum_{i \in \mathcal{D}} \sum_{j \in \mathcal{M}} q_{i,j}^{(n)} (\underline{w}_i - U\Phi^{(n+1)}\underline{v}_j)^T (\underline{w}_i - U\Phi^{(n+1)}\underline{v}_j). \end{aligned} \quad (15)$$

In other words, the a posteriori probabilities  $q_{i,j}^{(n)}$  effectively regulate the contributions to the likelihood function. Matches for which there is little evidence contribute insignificantly, while those which are in good registration dominate.

### 3.3. Maximisation

The maximisation step aims to locate the updated parameter-vector  $\Phi^{(n+1)}$  that optimises the quantity  $Q(\Phi^{(n+1)} | \Phi^{(n)})$ , i.e.

$$\Phi^{(n+1)} = \arg \max_{\Phi} Q'(\Phi | \Phi^{(n)}). \quad (16)$$

Since we recover the translation and scale parameters by normalising the feature-points, the goal in the maximisation step of the algorithm is to recover the Euler angles  $\psi$ ,  $\phi$  and  $\theta$  which define the elements of the matrix  $\Phi$ . To do this we compute the derivatives of  $Q'(\Phi | \Phi^{(n)})$  with respect to the elements of the parameter matrix  $\Phi$ . By setting the resulting derivatives equal to zero we obtain a family of linear equations for the elements of the matrix  $\Phi$ . These may be solved using matrix inversion. To do this we must introduce some matrix notation. Let  $V = (\underline{v}_1 | \underline{v}_2 | \dots | \underline{v}_{|\mathcal{M}|})$  be a  $3 \times |\mathcal{M}|$  matrix with the model-point position vectors as columns. Similarly, let  $W = (\underline{w}_1 | \underline{w}_2 | \dots | \underline{w}_{|\mathcal{D}|})$  be a  $2 \times |\mathcal{D}|$  matrix with the data-point position vectors as columns. We also introduce a  $|\mathcal{D}| \times |\mathcal{M}|$  weight-matrix  $Q^{(n)}$  whose elements  $q_{i,j}^{(n)}$  are the a posteriori correspondence probabilities evaluated at iteration  $n$ . Finally, we let  $R^{(n)}$  be the  $|\mathcal{D}| \times |\mathcal{D}|$  matrix whose diagonal elements  $R_{i,i}^{(n)} = \sum_{j \in \mathcal{M}} q_{i,j}^{(n)}$ . With this notation, the updated rotation matrix is

$$\Phi^{(n+1)} = U^{-1}(V(Q^{(n)})^T W)(W(R^{(n)})^T W^T)^{-1}. \quad (17)$$

Once the updated rotation matrix is to hand then it may be used to project the model-points onto the image. The rotation angles  $\psi$ ,  $\phi$  and  $\theta$  may be computed using Eq. (6).

This allows us to recover a set of improved transformation parameters at iteration  $n+1$ . Once these are computed, the a posteriori measurement probabilities may be updated using the method outlined in the previous subsection.

### 3.4. Algorithm summary

To conclude this section we summarise the steps involved in the template fitting process. We commence by standardising the co-ordinates of the feature-points; this involves translating them to an origin located at the bridge of the nose and scaling them so that there is unit length between the chin and the origin of co-ordinates. Next, we set the initial parameter values. We do this by setting  $\phi = \theta = 0$ . The initial value of the angle  $\psi$  is set equal to the angle between the y-axis and the line connecting the bridge of the nose and the chin (i.e. the bilateral symmetry axis of the face).

Once the algorithm has been initialised, in this way then we iterate the EM algorithm to recover the three Euler angles. In each M-step we perform matrix inversion to recover estimates of the three Euler angles. These parameters are used to project the 3D facial template onto the image. The distances between the  $x$ - $y$  co-ordinates of the projected template points and the facial feature points are used to compute the residuals  $\varepsilon_{i,j}(\Phi^{(n)}) = \underline{w}_i - U\Phi^{(n)}\underline{v}_j$ . Once these residuals are to hand then the updated a posteriori probabilities can be computed in the E-step using the Bayes rule appearing in Eq. (11) and the Gaussian probability distribution appearing in Eq. (13). These two steps are iterated until convergence. Usually, we find 4 or 5 iterations are needed before the mean squared residuals stabilise.

## 4. Facial feature localisation

Before we experiment with our pose recovery algorithm, we pause to detail the facial feature localisation method used in our experiments. This is based on a Fourier domain matched filter technique. The basic aim is to develop a set of matched frequency domain filters that can be used to characterise each of eight different facial features. These are the left and right eyes, the left and right-eyebrows, the hairline, the nose, the mouth and the chin. The frequency domain matched filters are extracted from training images using inverse Fourier analysis. We provide an experimental evaluation of the method on the University of Berne face database. Here we explore the most effective choice of training data so that the filters can be effectively applied when the facial pose varies. We also evaluate the effectiveness of the method when facial occlusion due to spectacles is present.

### 4.1. Frequency domain matched filter design

Our feature detection filters are obtained using a frequency domain matched filtering technique. This is a well known

technique. However, we employ a pre-processing step in which we blur the marked features using a Gaussian filter.

In order to extract frequency domain matched filters, we exploit the duality between convolution in the spatial domain and multiplication in the frequency domain. We commence from a set of hand-labelled training images in which the desired feature points are labelled. In practice this is performed by manually assigning a binary representation to the training images. Feature points are marked with the high-bit and the remaining background areas are marked with the low-bit. In this raw form the binary training-data are unsuitable for Fourier domain analysis since point-features are associated with high frequency components. The solution to this problem is to blur the feature-point images with a Gaussian smoothing kernel. The filter is

$$G_{\sigma}(x, y) = \frac{1}{2\pi\sigma^2} \exp\left[-\frac{x^2 + y^2}{2\sigma^2}\right].$$

In other words, the training images are pre-processed with a low-pass filter.

Suppose that  $I_n$  is a blurred training image and that  $O_n$  represents the blurred hand-segmentation of the raw facial features, where  $n = 1, \dots, N$  is an index which runs over the  $N$  images available for training. We wish to find the convolution filter  $C_n$  that best matches the input and output images from the training set. In other words we seek the filter for which

$$O_n = I_n * C_n, \quad (18)$$

where  $*$  denotes the convolution operation in the spatial domain.

From the duality theorem, the Fourier transform of the convolution of the signal  $I_n$  with a filter  $C_n$  is equal to the product of the Fourier transform of signal  $I_n$  denoted by  $\mathcal{F}(I_n)$  and the filter characteristic  $\mathcal{F}(C_n)$

$$\mathcal{F}(O_n) = \mathcal{F}(I_n) \times \mathcal{F}(C_n), \quad (19)$$

where  $\mathcal{F}(\cdot)$  represents the Fourier transform.

In other words, the convolution filter  $C_n$  may be obtained from the inverse Fourier transform in the following way:

$$C_n = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(O_n)}{\mathcal{F}(I_n)}\right). \quad (20)$$

Each image in the training set yields a different convolution filter. In order to combine the results over the  $N$  examples in the training-set we average the resulting convolution filters, i.e.

$$C = \frac{1}{N} \sum_{n=0}^N C_n. \quad (21)$$

One of the potential problems with this Fourier inversion process is the existence of unsampled frequencies in the raw training data. In other there are frequency components for which the sampled Fourier transform of the image,  $\mathcal{F}(I_n)$

is zero. This means that the corresponding filter coefficients are undefined. This problem is commonly circumvented by setting the undefined coefficients themselves to zero. Because we have access to a large training set of images, we adopt a different strategy in which it is the sample average that is assigned rather than zero.

Our frequency domain filters have been trained using the University of Berne face-data-base. The database contains five different poses for each of 20 subjects. The viewing poses correspond to the cases when the face is front-on, tilted-up, tilted-down, turned to the left and turned to the right. The subjects are a mixture of European and oriental, both with and without spectacles. Here we aim to develop filters for recognising eight salient facial features. These are the left and right eyes, hair, left and right eyebrow, the nose, the mouth and the chin. We have sub-divided the images into the following classes. The first four classes distinguish the ethnic origins of the subjects (European or oriental) and whether or not the subjects are wearing spectacles. In addition we sub-divide the images according to the five facial poses described above. Once the subjects have been divided into classes, we segregate the data into disjoint sets for training and evaluation. In the case of pose, we have investigated the effects of choosing different views in the training phase. For each image in the training data we hand-segment and separately label the eight facial features. The eight feature localisation filters are averaged over the selected poses for the subjects in the training set.

The training procedure is as follows. For each face in the database we prepare a separate image in which each of the different target features is marked. These images are each convolved with a Gaussian filter to blur the marked feature points. Next, the matched filter is computed using Eq. (19). The filters obtained for the different training images are then averaged to give a single matched filter. This procedure is repeated for each of the target features in turn.

Once filters for facial feature extraction are to hand, the next task is to localise candidate features. We realize this as a two-stage process. Each image in turn is separately filtered with the appropriate set of eight feature characterisation filters. The relevant features are marked as the positions of global maximum convolution response for the corresponding filter. This process is stabilised by post-smoothing the filtered images with a Gaussian kernel whose size is of the same order as the target features.

#### 4.2. Examples

Fig. 3 illustrates the filter training and feature localisation. In Figs. 3(a)–(c) we show the original training image, the Gaussian blurring of the hand-labelled feature, and the convolution filter found. A different image (test image) is shown in Fig. 3(d). We convolve this image with the filter from Fig. 3(c). The response given by Eq. (19) is shown in Fig. 3(f). Figs. 3(g)–(i) show the steps used to localise the desired feature point.

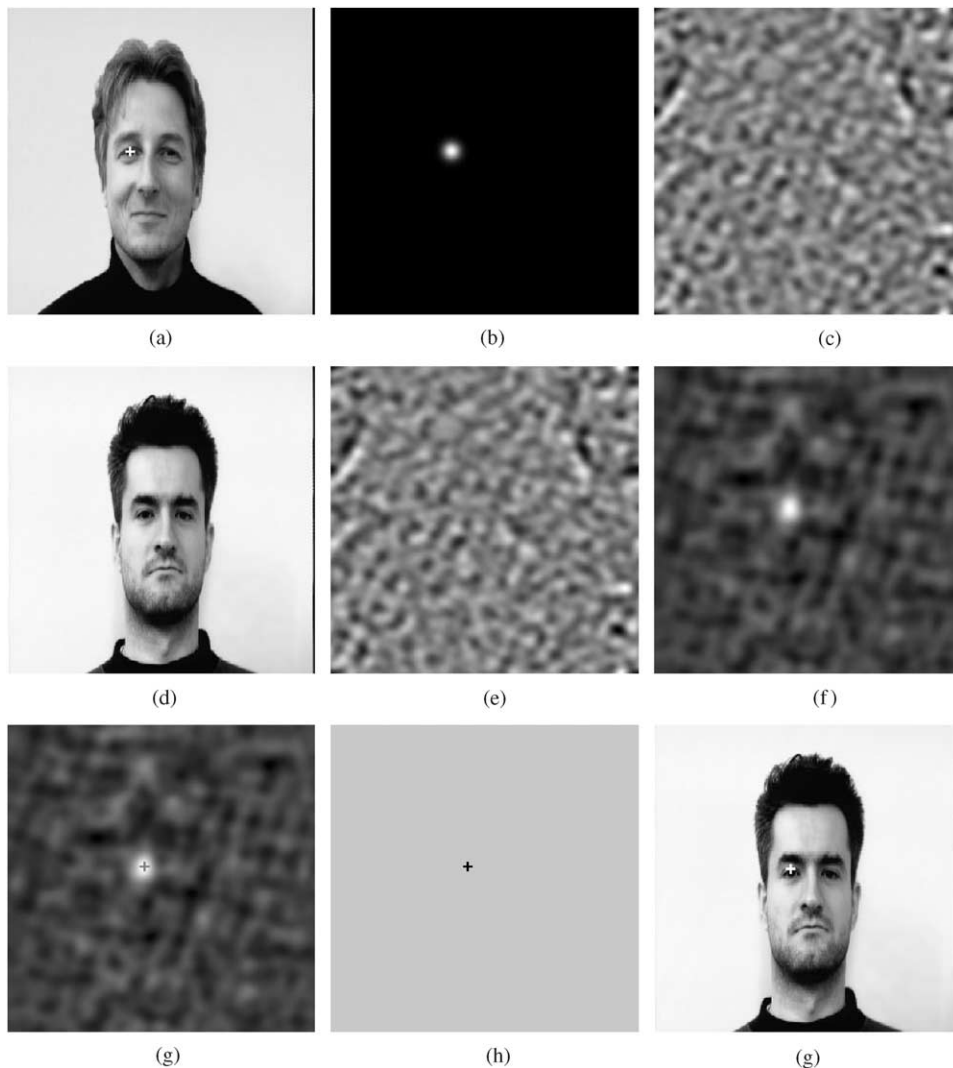


Fig. 3. Facial feature localisation with together with their frequency domain matched filters. (a) Manually marked point for training. (b) Gaussian blurred response. (c) Frequency domain filter obtained by training. (d) Image used for evaluation. (e) Frequency domain filter used for feature detection: same as (c). (f) Filter response. (g) Localised maximum point with response. (h) The point for the feature. (i) Marked point in the image.

We now present results to investigate some of the systematic variations of the frequency domain matched filters with pose and degree of blurring. We commence by investigating the effect of varying facial pose. Fig. 4 shows examples of the Gaussian-blurred responses and the corresponding frequency domain matched filters for the right eye when the head changes pose. In each case column (a) shows the original image, column (b) shows the Gaussian blur of the hand labelled feature point and column (c) shows the matched filter. It is clear that the matched filter changes significantly with pose. As a result, we may expect to have to use separate frequency domain filters for different poses.

Next, we illustrate the effect of varying the width of the Gaussian filter used to blur the hand-labelled features. For each filter in turn we have measured the localisation error for the detected feature points. Figs. 5 and 6 shows the localisation error as the width of the Gaussian filter is increased. This figure illustrates that the localisation error is a minimum when the width of the Gaussian kernel is 3 pixels. In this experiment we use 10 people with five poses for training. The same number of test faces are used to obtain the localisation error.

Finally, we investigate the effect of averaging the frequency domain matched filters over different subjects. In

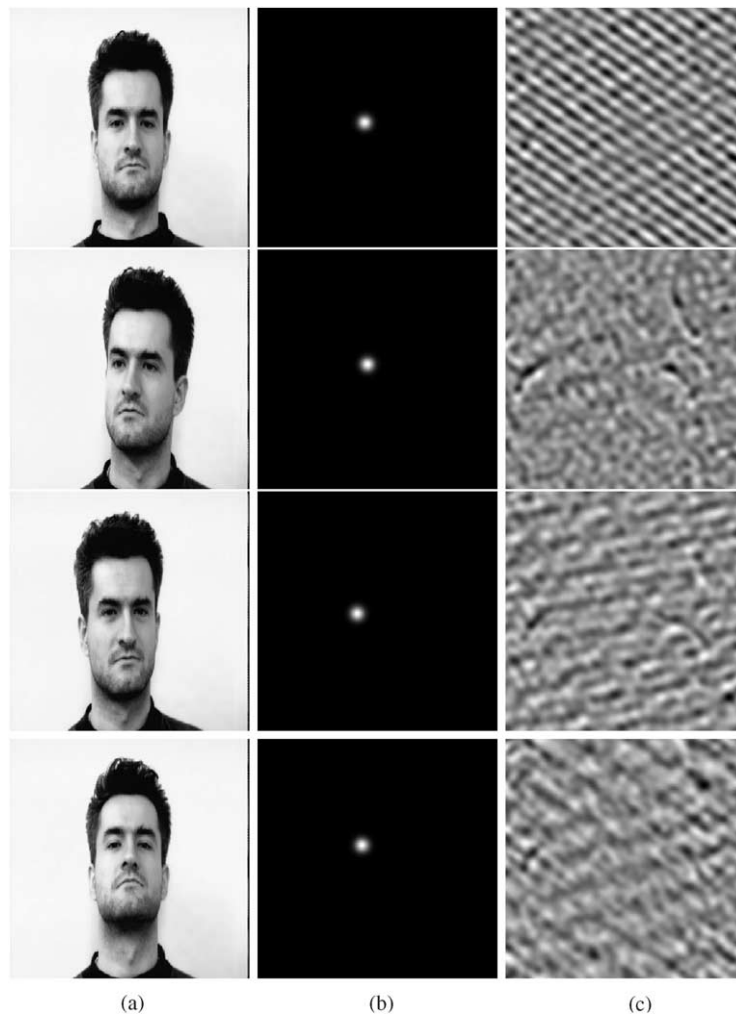


Fig. 4. Gaussian kernel responses the right eye for each pose and the corresponding frequency domain matched filters. (a) Facial poses. (b) Gaussian kernel responses. (c) Frequency domain filters.

Fig. 7 we show the resulting feature point localisation error as a function of the number of individuals used in training. The main effect to note is that the localisation error decreases as the number of subjects increases. In this experiment 10 people with five poses each are used for evaluation. The localisation error drops rapidly once more than four individuals are used for training. This suggests that it is not necessary to use many facial images for training in order to obtain good localisation accuracy.

#### 4.3. Real world examples

The aims of our experiments are twofold. Firstly, we demonstrate that the matched filter technique is effective at localising facial features over a sample of subjects having diverse appearance and pose. The second aim is to evaluate

the accuracy of feature localisation to feature-type, to choice of training examples, to ethnic group and, finally, to facial occlusion caused by spectacles.

To commence we illustrate some examples of the feature localisation process on images drawn from the face-database. Fig. 8 illustrates the feature localisation process when the frequency domain matched filters are trained over all five poses. The localisation accuracy is good and the method works well provided that there is no facial occlusion by spectacles. To take this experimentation further, we have investigated the ability of the filters to generalise over varying poses. Fig. 9 shows the training images used to compute the eight matched filters for the different facial features. The marked points are the manually located positions of the feature centres in the training data. The training set contains both European and oriental subjects

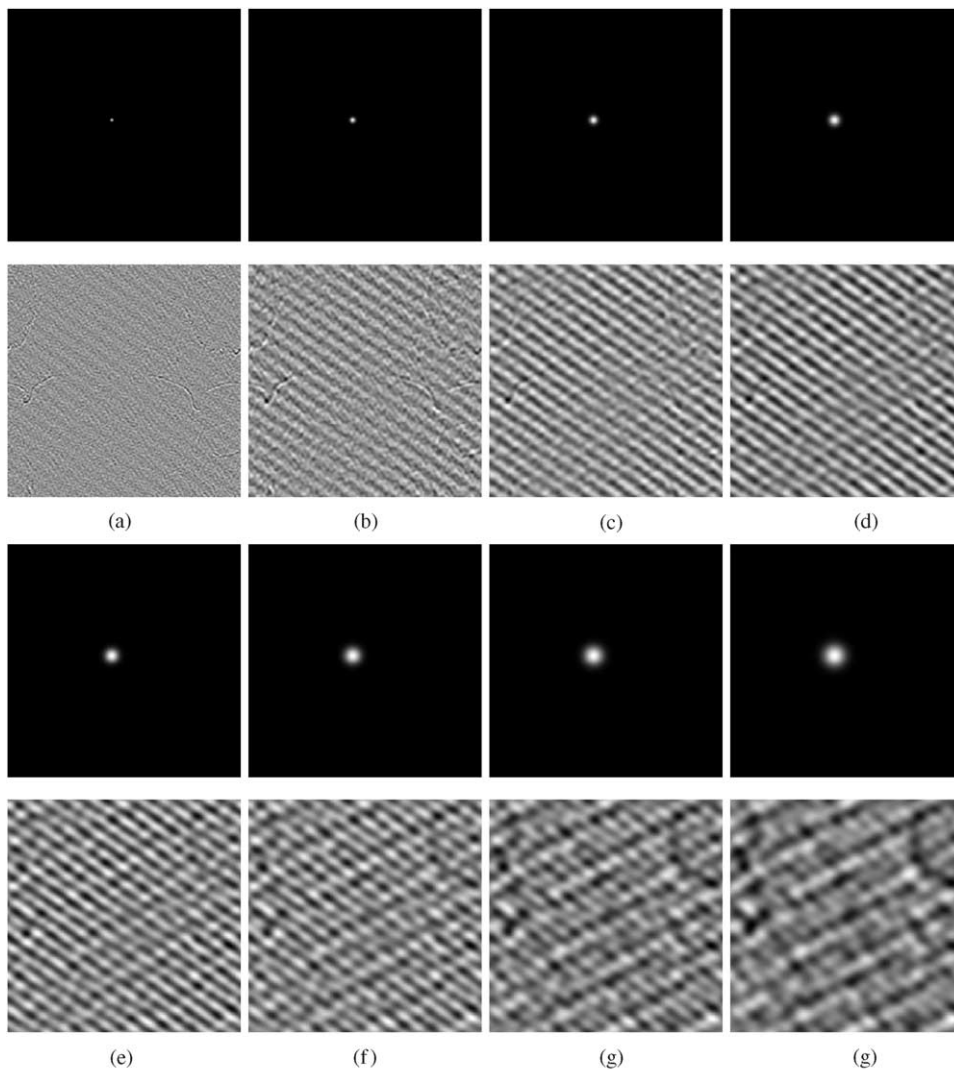


Fig. 5. Variation of the scale for Gaussian kernel response for the right eye and the corresponding frequency domain matched filters. (a)  $\sigma = 1$ . (b)  $\sigma = 2$ . (c)  $\sigma = 3$ . (d)  $\sigma = 4$ . (e)  $\sigma = 5$ . (f)  $\sigma = 6$ . (g)  $\sigma = 7$ . (h)  $\sigma = 8$ .

viewed in the frontal, left and tilted-down poses. There are also examples in which the subjects wear spectacles. The frequency domain feature localisation filters are averaged over the different poses and different subjects. Fig. 10 illustrates the results of applying the resulting feature filters to the two poses not used in the training procedure, i.e. the tilted-up and right poses. The resulting localisation is best for the eye and eyebrow features and poorest for the nose, lip and chin features. This conclusion is supported in a more quantitative manner in Fig. 11, which shows the feature localisation error as a function of feature-type. Here we show the result of choosing different combinations of views in the training and evaluation steps. From this plot

it is clear that using frontal, right-tilted and down-tilted views in training gives the smallest localisation error. It is also apparent that the localisation error decreases with the number of training examples.

We have also investigated the sensitivity of the method to facial occlusion by spectacles. Fig. 12 compares the effect of including examples of spectacle wearers in the training set. The images used for subsequent evaluation consist of both Europeans and oriental subjects; there are examples of spectacle wearers and non-spectacle wearers. In the figure the plus points show the result obtained when spectacle-wearers are included in training, the square points are the case when they are excluded. The diamond points show the result ob-

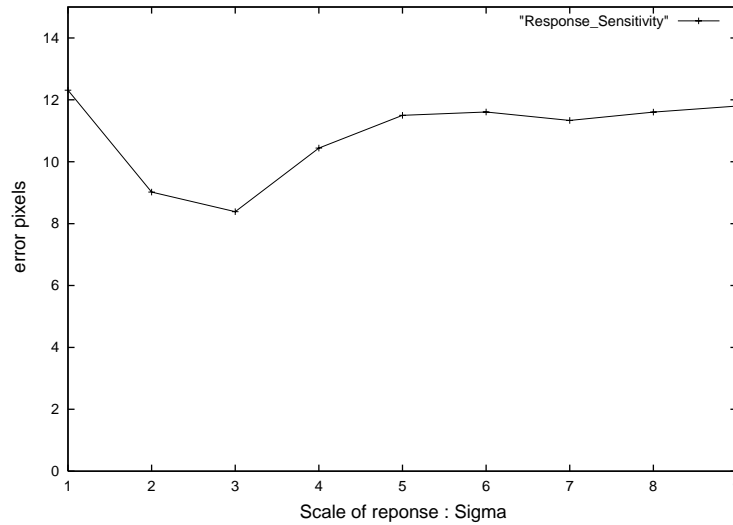


Fig. 6. The localisation accuracy with varying scale of the Gaussian blur  $\sigma$  for responses.

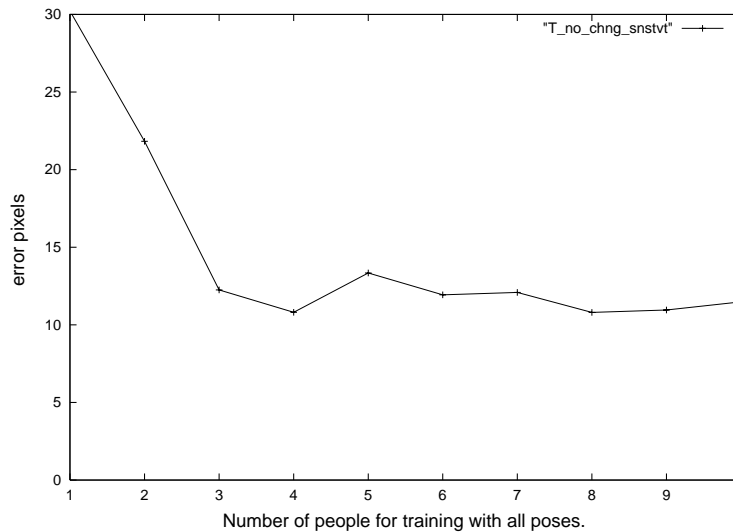


Fig. 7. The feature localisation error changes when the number of people increases for training. In this experiment 10 people with five poses each are used for evaluation.

tained when the two training sets are amalgamated. We can conclude that the amalgamated training set gives the best set of filters. From the plot it is clear that if we attempt to recognise normal facial features when the filters have been trained on examples in which there are spectacles, then we encounter problems with the localisation of the features on the facial symmetry axis (i.e. the nose, lip and chin). The main effect of including the non-spectacle examples in the training set is to improve the localisation for the features on the symmetry axis of the face, i.e. the nose, lip and chin. Comparing Figs. 11 and 12 it seems to be the case that the localisation error due to occlusion is smaller than the er-

ror that results from a bad choice of training poses. Fig. 13 provides some illustrative examples of the images used to compile the graph in Fig. 12 when the filter is trained using the amalgamated set of samples. These results shown in Fig. 13 represent an improvement in localisation error over those shown in Figs. 8 and 10.

### 5. Pose estimation experiments

The evaluation of our pose estimation procedure involves experiments on both contrived and natural imagery. The

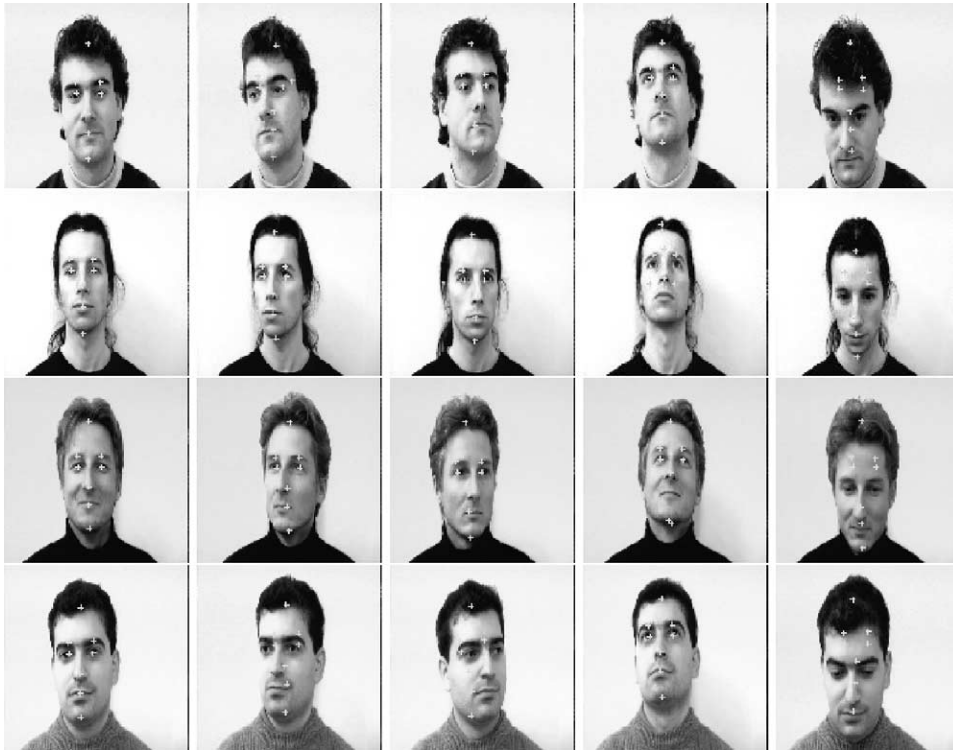


Fig. 8. Located feature points marked with a '+' symbol.



Fig. 9. Training images.

contrived data is provided by various camera views of a plaster bust. Here the ground-truth pose angle is measured in the laboratory and the facial feature points are marked by hand. The natural data is provided by 21 different cam-

era views for each of eight different individuals. Here we experiment with both hand-segmented, together with automatically segmented, feature points. The segmentation process uses Fourier-domain matched filters, as described in



Fig. 10. Located features.

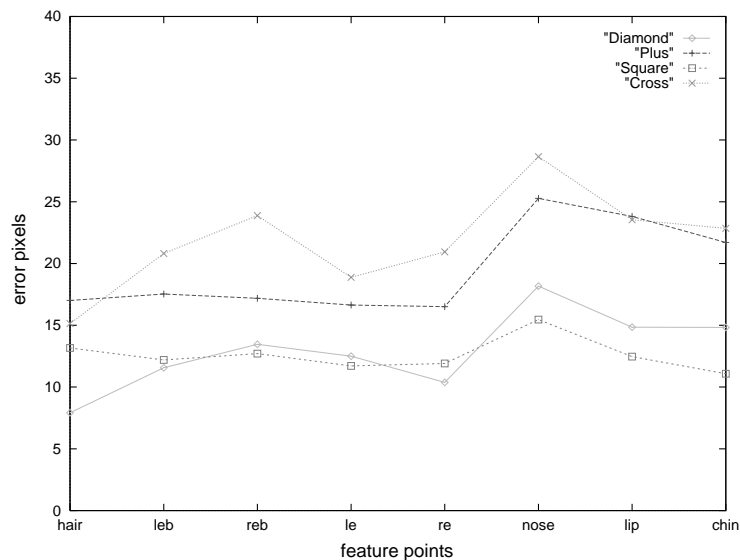


Fig. 11. Effect of training pose on localisation error; diamond points (forward, right, down), plus (left, up), square (forward, left, up), cross (right, down).

the previous section, to characterise each of eight facial features (left and right eyebrows, left and right eye centres, hairline, nose, lips and chin). When averaged over the eight feature types, the feature localisation error is about 5 pixels. However, for certain features (e.g. the eye centres) the localisation error is about 3 pixels. These sensitivity systematics are summarised in Fig. 14, which shows the localisation error as a function of facial pose and camera direction (fronto-parallel, oblique from above, oblique from below).

We commence our study by considering the contrived data. Fig. 15 shows a series of views of a plaster bust. There are three camera directions. These are approximately fronto-parallel, oblique from above and oblique from below. Under each of the views we list the ground truth rotation angle for the bust. This angle is measured with a protractor

attached to the base of the bust. Zero rotation angle corresponds to the case when the nose points straight towards the camera. Also listed below the different views is the pose angle computed using our EM algorithm. For pose angles of up to  $40^\circ$  in both the clockwise and counterclockwise senses, there is good agreement between the ground-truth and recovered angles.

This feature of the data is illustrated more directly in Fig. 16. Here we show the difference between the ground-truth and recovered pose angles as a function of the ground-truth angle. There are three features of this plot that deserve further comment. Firstly, for moderate rotation angles the average error is approximately  $3^\circ$ . Secondly, the error increases dramatically for rotation angles greater than  $40^\circ$ . Finally, there appears to be a positive bias to the

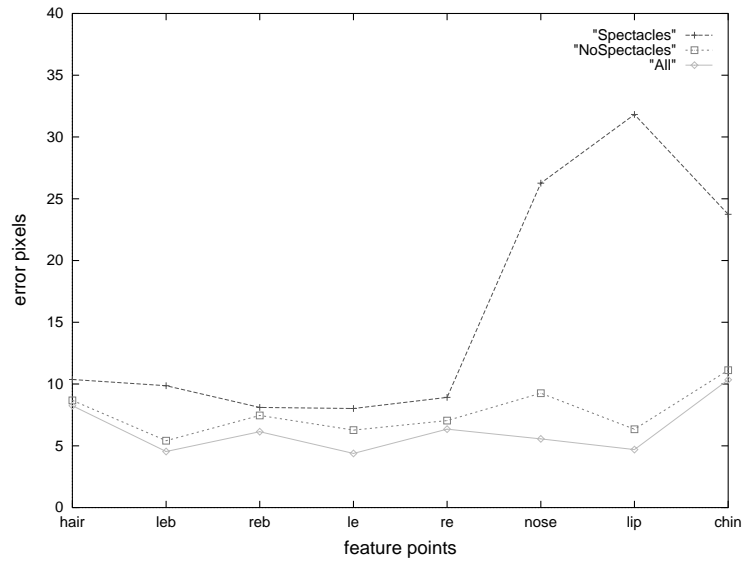


Fig. 12. The effect of spectacle occlusion on localisation error; the plus points are the result if spectacles are included in the training set while the square points are the result if they are excluded. The diamond points show the result obtained with the amalgamated training set.



Fig. 13. Located feature points using the filter trained with the amalgamated set of samples.

computed error. This is attributable to the fact that we initialise our face-template in a fronto-parallel configuration at zero rotation angle. In other words, the model must always make a positive rotation on to the data. Local optima or

premature convergence in the fitting process may therefore bias the method to under-estimate the rotation angle.

To illustrate the iterative qualities of the algorithm, Fig. 17 shows the feature template converging on the fea-

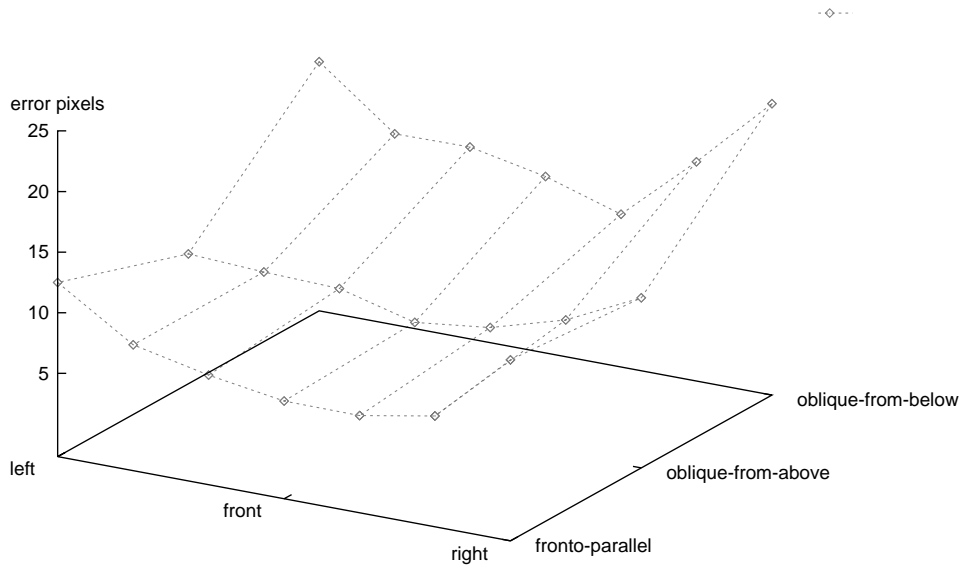


Fig. 14. The positional accuracy of automatically segmented points.



Fig. 15. A series of views of a plaster bust in which the camera direction is approximately fronto-parallel, oblique from above and oblique from below. The ground-truth angles are denoted by “T” and the estimated angles by “E”. (a) T:  $-30$  E:  $-29.2$ . (b) T:  $-10$  E:  $-9.8$ . (c) T:  $+10$  E:  $+14.8$ . (d) T:  $+30$  E:  $+31.7$ . (e) T:  $-30$  E:  $-32.7$ . (f) T:  $-10$  E:  $-9.6$ . (g) T:  $+10$  E:  $+12.3$ . (h) T:  $+30$  E:  $+34.4$ . (i) T:  $-30$  E:  $-34.4$ . (j) T:  $-10$  E:  $-11.2$ . (k) T:  $+10$  E:  $+12.9$ . (l) T:  $+30$  E:  $+35.4$ .

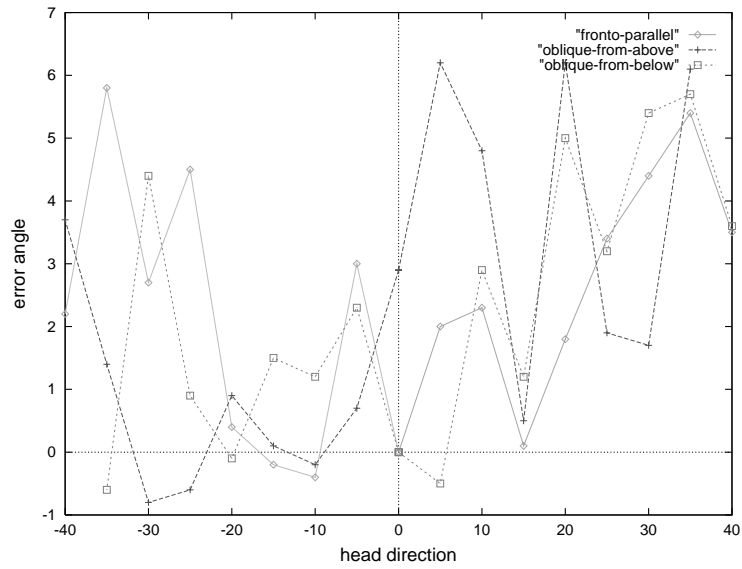


Fig. 16. The difference between the ground-truth and recovered pose angles as a function of the ground-truth angle.



Fig. 17. The 3D feature template converging on the feature points. (a) Initial step. (b) Final step. (c) Initial step. (d) Final step.



Fig. 18. Examples of the estimated rotation angles for various facial poses. (a), (b), (c) The same rotation angles for the oblique from below, fronto-parallel, and oblique from above. (d), (e), (f) The same rotation angles for the oblique from below, fronto-parallel, and oblique from above. (a) E:  $-36.7$ . (b) E:  $-36.7$ . (c) E:  $-38.1$ . (d) E:  $-13.4$ . (e) E:  $-10.0$ . (f) E:  $-25.8$ .

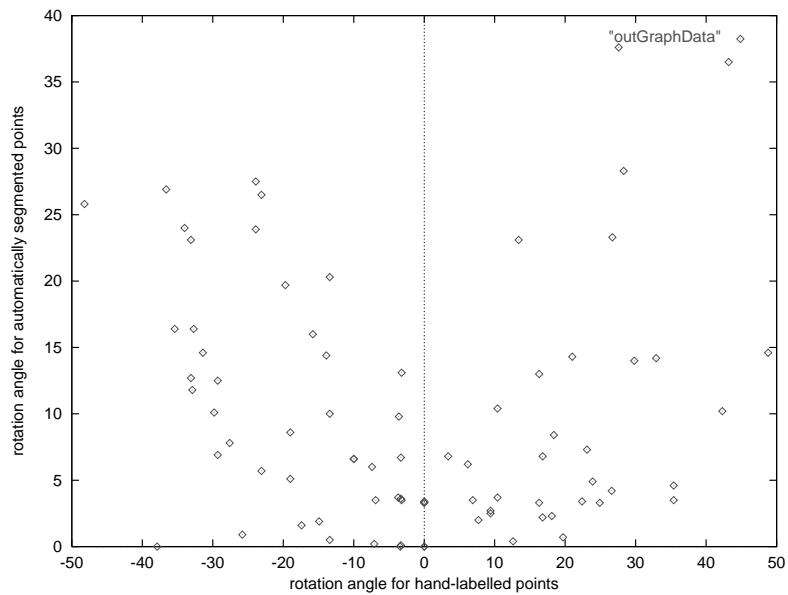


Fig. 19. The difference in computed rotation angle for the hand-segmented and automatically segmented points.

ture points. The two examples are for the plaster-bust and the natural image. The first image shows the initial template alignment using constraints on the position of the origin co-ordinates and the direction of the bilateral symmetry axis. The second image shows the final position of the template after convergence of the EM algorithm.

Turning our attention to the real-world data, Fig. 18 shows a sequence of views with the computed rotation angles appended. In this figure, the feature points are hand segmented. Here our experiments have focussed on how the template registration method degrades when automatically segmented, rather than hand-segmented, feature points are used. Fig. 19 shows a plot of the difference in computed rotation angle for the hand-segmented and automatically segmented feature points. Each entry in the plot is averaged over eight different individuals. The main feature to note from the plot is that the error increases with the rotation angle. However, for moderate rotation angles, the error is only about  $3^\circ$ .

Although the plaster bust allows us to evaluate the sensitivity of the pose estimation method when hand-labeled features are available, it does not allow us understand the effects of poor feature localisation in realistic face images. However, ground-truth pose angles are difficult to obtain for natural face images. For this reason, we present a simulation study using a VRML model to establish ground-truth. The model is constructed using both range data and binocular intensity images. A mesh is fitted to the depth information returned by the range sensor. The intensity image data is then texture-mapped onto the mesh. Fig. 20 shows the original frontal intensity view, the mesh extracted from the range data and the synthetic texture-mapped views of the VRML model obtained using 3DStudioMax. Fig. 21 shows views of the VRML model under rotations by  $10^\circ$  increments, between  $-30^\circ$  and  $30^\circ$ . Our aim is to align a 3D template to feature points located in this data. The feature points are located using the matched filter technique described in Section 5. The 3D facial template used for alignment has been constructed using depth information provided by the range data for the face. The template control points have been hand located at the positions of the six facial features in the range data.

In Fig. 22 we plot the error in the angle  $\psi$ . There are three curves on this plot. The dotted curve is the result obtained when we match the generic 3D template to the 2D feature points. For reference, the dashed curve is the result obtained when we match the ground-truth 3D template extracted from the range data. The solid curve is the result obtained when we align a range template from which the nose is removed. All three methods return angular errors of less than a degree. The error increases with increasing facial rotation. In the case of the generic template the angular errors are about twice as large as those obtained with the range-template. For the noseless range template the error is approximately flat with rotation angle, and much poorer than that ob-

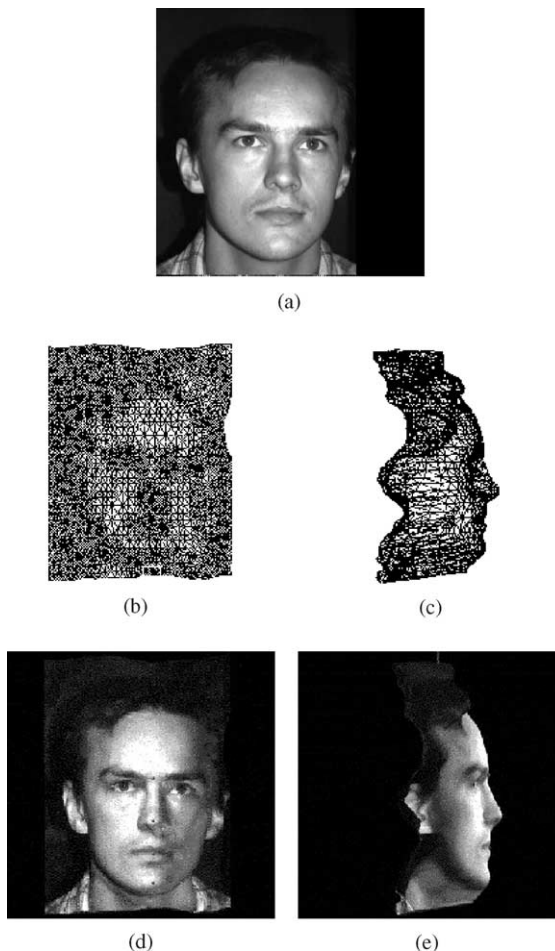


Fig. 20. 3D face images derived from the original 2D face image. (a) A frontal facial view (b), (c) Meshes of the front and left views. (d), (e) The front and left side face images computed using the VRML model (virtual reality modelling language).

tained with either the range-data template or the generic template.

## 6. Conclusions

The main contribution of this paper is to present a statistical framework for iteratively registering 3D facial templates against 2D feature points. The iterative procedure is based on the EM algorithm and allows the parameters of orthographic projection between the 3D model and the 2D data to be estimated. An analysis on ground-truthed data reveals that the method is capable of recovering the rotation angle of the head to within  $3^\circ$  provided that the overall rotation does not exceed  $40^\circ$ . The main limitation of the method is the need for accurately located feature points.

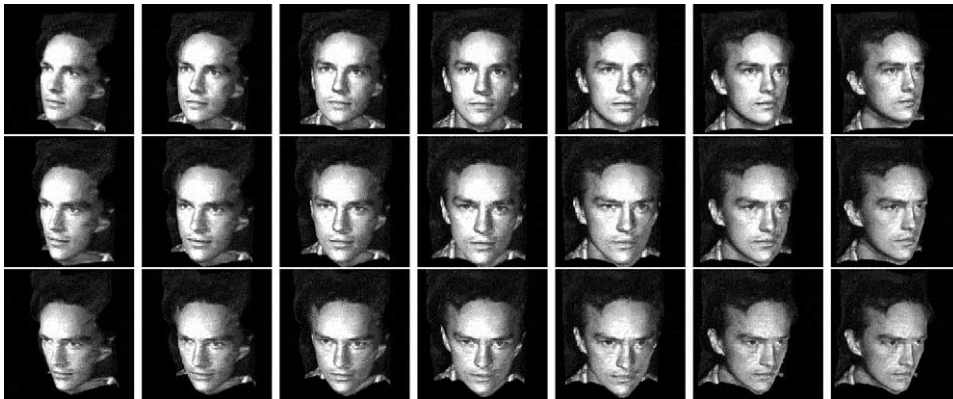


Fig. 21. Views from various view angles of 3D face in which the camera direction is approximately fronto-parallel, oblique from above and oblique from below. The faces are artificially generated after rotating the original image using VRML.

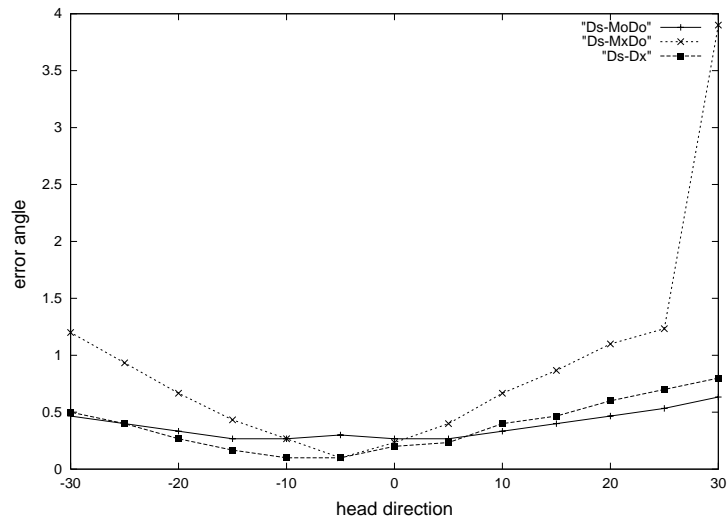


Fig. 22. Ground-truth and recovered pose angles as a function of the ground-truth angle for the images from Fig. 21. ‘MoDo’, represents the plot when both model and data have nose height information. ‘MxDo’ when only data has it. ‘Ds-Dx’ is the plot when the data do not include any nose information.

## References

- [1] B. Moghaddam, A. Pentland, Probabilistic visual learning for object detection, Proceedings of the Fifth International Conference on Computer Vision, 1995, pp. 786–793.
- [2] C. Kotropoulos, I. Pitas, S. Fischer, B. Duc, Face authentication using morphological dynamic link architecture, Lecture Notes in Computer Science, vol. 1206, Springer, Berlin, 1997, pp. 169–176.
- [3] S. Avidan, A. Shashua, Novel view synthesis by cascading trilinear tensors, IEEE Trans. Visual. Comput. Graph. 4 (1998) 293–306.
- [4] A. Gee, R. Cipolla, Determining the gaze of faces in images, Image Vision Comput. 12 (10) (1994) 639–647.
- [5] G. Sullivan, K. Baker, A. Worrall, C. Attwood, P. Remagnino, Model-based vehicle detection and classification using orthographic approximations, Image Vision Comput. 15 (1997) 649–654.
- [6] K. Sung, T. Poggio, Example based learning for view-based human face detection, Technical Report, AI Memo 1521, CBCL Paper 112, MIT, 1995.
- [7] A. Tsukamoto, C. Lee, S. Tsuji, Detection and pose estimation of human face with synthesized image models, in: International Conference on Pattern Recognition, vol. 94, 1994, pp. A:754–757.
- [8] J. Ng, S. Gong, Multi-view face detection and pose estimation using a composite support vector machine across the view sphere, in RATFG99, 1999.
- [9] B. Takacs, H. Wechsler, Detection of faces and facial landmarks using iconic filter banks, Pattern Recognition 30 (1997) 1623–1636.

- [10] G. Chow, X. Li, Towards a system for automatic facial feature detection, *Pattern Recognition* 26 (1993) 1739–1755.
- [11] X. Xie, R. Sudhakar, H. Zhuang, In improving eye feature-extraction using deformable templates, *Pattern Recognition* 27 (1994) 791–799.
- [12] H. Wu, Q. Chen, M. Yachida, Facial feature extraction and face verification, in: *International Conference on Pattern Recognition*, vol. 96, 1996, pp. 480–488.
- [13] K. Sobottka, I. Pitas, Extraction of facial regions and features using color and shape information, in: *International Conference on Pattern Recognition*, vol. 96, 1996, pp. C:421–425.
- [14] B. Takacs, H. Wechsler, Locating facial features using softm, in: *International Conference on Pattern Recognition*, vol. 94, 1994, pp. B:55–60.
- [15] J. Bala, H. Wechsler, H. Vafaie, J. Huang, K. DeJong, Visual routine for eye detection using hybrid genetic architectures, in: *International Conference on Pattern Recognition*, vol. 96, 1996, pp. 606–610.
- [16] R. Pinto-Elias, J. Sossa-Azuela, Automatic facial feature detection and location, in: *International Conference on Pattern Recognition*, vol. 98, 1998, pp. 1360–1364.
- [17] D. Reisfeld, Y.Y., Robust detection of facial features by generalized symmetry, in: *International Conference on Pattern Recognition*, vol. 92, 1992, pp. 117–120.
- [18] M. Turk, A. Pentland, Eigenfaces for recognition, *J. Cognit. Neurosci.* 3 (1991) 71–86.
- [19] A. Yuille, D. Cohen, P. Hallinan, Feature extraction from faces using deformable templates, *Int. J. Comput. Vision* 8 (1992) 99–112.
- [20] T. Cootes, C. Taylor, D. Cooper, J. Graham, Active shape models—their training and application, *Comput. vision Image Understand.* 61 (1995) 38–59.
- [21] I. Craw, H. Ellis, J. Lishman, Automatic extraction of face features, *Pattern Recognition Lett.* 5 (1987) 183–187.
- [22] I. Craw, D. Tock, A. Bennett, Finding face features, in: *European Conference on Computer Vision*, vol. 92, 1992, pp. 92–96.
- [23] R. Rao, D. Ballard, Natural basis functions and topographic memory for face recognition, in: *Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI, 1995*, pp. 10–17.
- [24] A. Pentland, B. Moghaddam, T. Starner, View-based and modular eigenspaces for face recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1994*, pp. 84–91.
- [25] R. Brunelli, T. Poggio, Face recognition: features versus templates, *IEEE Trans. Pattern Anal. Mach. Intell.* 15 (1993) 1042–1052.
- [26] L. Wiskott, J. Fellous, N. Kruger, C. Malsburg, Face recognition by elastic bunch graph matching, Technical Report IRINI 96-08, Ruhr-Universitat Bochum, 1996.
- [27] J. Lee, E. Miliotis, Matching range images of human faces, in: *International Conference on Computer Vision*, vol. 90, 1990, pp. 722–726.
- [28] G. Gordon, Face recognition based on depth and curvature features, *Comput. Vision Pattern Recognition* 92 (1992) 808–810.
- [29] Y. Yacoob, L. Davis, Labeling of human face components from range data, *Computer Vision, Graph. Image Process.* 60 (1994) 168–178.
- [30] K. Hattori, S. Matsumori, Y. Sato, Estimating pose of human face based on symmetry plane using range and intensity image, in: *International Conference on Pattern Recognition*, vol. 98, 1998, pp. 1183–1187.
- [31] A. Gee, R. Cipolla, Estimating gaze from a single view of a face, in: *International Conference on Pattern Recognition*, vol. 94, 1994, pp. A:758–760.
- [32] S. McKenna, S. Gong, J. Collins, Face tracking and pose representation, in: *British Machine Vision Conference*, vol. 96, 1996, pp. 755–764.
- [33] J. Huang, D. Li, X. Shao, H. Wechsler, Pose discrimination and eye detection using support vector machines, *Face Recognition, From Theory to Applications—NATO ASI Series, Series F: Computer and Systems Sciences*, vol. 163, Springer, Berlin, 1997, pp. 528–535.
- [34] J. Huang, X. Shao, H. Wechsler, Face pose discrimination using support vector machines, in: *International Conference on Pattern Recognition*, vol. 98, 1998, pp. 154–156.
- [35] S. Romdhani, S. Gong, A. Psarrou, Multi-view nonlinear active shape model using kernel pca, in: *British Machine Vision Conference*, vol. 99, 1999, pp. 483–492.
- [36] T. Cootes, C. Taylor, A. Lanitis, Active shape models: evaluation of a multi-resolution method for improving image search, in: *Proceedings of the British Machine Vision Conference, 1994*, pp. 327–336.
- [37] A. Dempster, N. Laird, D. Rubin, Maximum-likelihood from incomplete data via the em algorithm, *Royal Statistical Society Ser. B (methodological)*, vol. 39, 1977, pp. 1–38.
- [38] S. Moss, E. Hancock, Registering incomplete radar images with the em algorithm, image and vision computing, *Image and Vision Comput.* vol. 15, 1997, pp. 637–648.
- [39] A. Cross, E. Hancock, Graph matching with dual step em algorithm, *IEEE Trans. Pattern Recognition Mach. Anal.*, vol. 20, 1998, 1236–1253.
- [40] J. Hornegger, H. Niemann, Statistical learning localisation and identification of objects, *Proceedings of the Fifth International Conference on Computer Vision, 1995*, pp. 914–919.

**About the Author**—KWANG NAM CHOI is from Korea. He holds a masters degree in Computer Science. He is currently undertaking research on face recognition and analysis as part of a DPhil degree in Computer Vision in the Department of Computer Science at the University of York.

**About the Author**—MARCO CARCASSONI received his Laurea degree from the University of Padua in Computer Science and Engineering in 1997. Since 1998 he has been working towards a DPhil degree in Computer Vision in the Department of Computer Science at the University of York. He is currently employed as a Research Associate working on an EPSRC project concerned with image retrieval. He has published some 10 papers in image analysis, computer vision and pattern recognition. He is a member of the IEEE.

**About the Author**—EDWIN HANCOCK studied Physics as an undergraduate at the University of Durham and graduated with honours in 1977. He remained at Durham to complete a Ph.D. in the area of high energy physics in 1981. Following this he worked for ten years as a researcher in the fields of high-energy nuclear physics and pattern recognition at the Rutherford-Appleton Laboratory (now the Central Research Laboratory of the Research Councils). During this period he also held adjunct teaching posts at the University of Surrey and the Open University. In 1991 he moved to the University of York as a lecturer in the Department of Computer Science. He was promoted to Senior Lecturer in 1997 and to Reader in 1998. In 1998 he was appointed to a Chair in Computer Vision.

Professor Hancock now leads a group of some 15 faculty, research staff and Ph.D. students working in the areas of computer vision and pattern recognition. His main research interests are in the use of optimisation and probabilistic methods for high and intermediate level vision. He is also interested in the methodology of structural and statistical pattern recognition. He is currently working on graph-matching, shape-from-X, image data-bases and statistical learning theory. His work has found applications in areas such as radar terrain analysis, seismic section analysis, remote sensing and medical imaging. Professor Hancock has published some 60 journal papers and 200 refereed conference publications. He was awarded the Pattern Recognition Society medal in 1991 for the best paper to be published in the journal *Pattern Recognition*. The journal also awarded him an outstanding paper award in 1997.

Professor Hancock has been a member of the Editorial Boards of the journals *IEEE Transactions on Pattern Analysis and Machine Intelligence*, and, *Pattern Recognition*. He has also been a guest editor for special editions of the journals *Image and Vision Computing* and *Pattern Recognition*, and he is currently a guest editor of a special edition of *IEEE Transactions on Pattern Analysis and Machine Intelligence* devoted to energy minimisation methods in computer vision. He has been on the programme committees for numerous national and international meetings. In 1997 he established a new series of international meetings on energy minimisation methods in computer vision and pattern recognition. He was awarded a Fellowship of the International Association for Pattern Recognition in 2000.