Estimating 3D Facial Pose using the EM Algorithm

Kwang Nam Choi, Marco Carcassoni and Edwin R. Hancock
Department of Computer Science
University of York, York, YO10 5DD, UK

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 the 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 degrees.

1 Introduction

Facial pose estimation is a key task for many practical computer vision applications. Specific examples include visual surveillance, camera assisted user interfaces [9] and user identification verification [7]. In essence, the problem revolves around the fitting of a generic 3D template to labelled facial features located in a 2D image. Once the template has been fitted to the feature data, then 3D pose parameters may be used to manipulate the face. Viewed in this way pose estimation may be regarded as an essential pre-requisite to detailed facial verification.

There have been many attempts at efficiently recovering the 3D facial pose [4, 5, 8]. 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 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, then 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 [12].

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, Laird and Rubin [3] to learn the 3D pose parameters subject to 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 work where it has been successfully used to match line-templates [10],

BMVC 1998 doi:10.5244/C.12.9

Figure 1: The basic geometry of the face template.

shape templates [11] and 3D perspective models [2]. Here we commence by constructing a generic 3D template of the facial features. The template is quite simple. It assumes that the eyes 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. 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 centering and aligning the template at a fixed point on the bilateral facial symmetry axis. This removes the three degrees of freedom associated with two template translation parameters on the image plane together with an Euler rotation of the template symmetry axis in the bary-centred co-ordinate system.

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. Experiments and sensitivity analysis are presented in section 4. Finally, Section 5 offers some conclusions and outlines our future plans.

2 Geometric Model

Our basic aim is to register 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. These planar features are symmetric about the axis defined by the centre-points of the lip and the chin. The tip of the nose is assumed to 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 Figure 1.

The projection of the 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 model points. However, the complexity of the projection can be simplified using constraints provided by the 2D geometry of the labelled 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. An alternative is to connect the centres of the lips and chin.

The 3D template control points are represented by co-ordinate vectors $\underline{\mathbf{v}}_j = (x_j, y_j, z_j)^T$, 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 $\underline{\mathbf{w}}_i = (x_i, y_i)^T$ 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

$$\underline{\mathbf{u}}_j(\Phi) = sUR_\phi S_\psi T_\theta \underline{\mathbf{v}}_j - X_o \tag{1}$$

Here s is the overall model scale parameter and $X_o = (x_o, y_o)^T$ is the translation of the origin in the image co-ordinate system. The matrices R_ϕ, S_θ and T_ψ represent Euler rotations of the model about its bary-centre. The 3x2 matrix U selects the two x-y components from the three x-y-z components of the transformed template control points.

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 θ. Recall that in our template, this plane is defined by the eyes, lip and chin. The net effect of this rotation is tilt the head to the left or the right. In other words, it rotates the bilateral axis of facial symmetry by an angle θ in the image plane. The rotation matrix is given by

$$T_\theta = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{pmatrix} \tag{2}$$

The next step is to rotate by an angle ψ about the z-axis of the template. In our representation, the z-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

$$S_\psi = \begin{pmatrix} \cos\psi & \sin\psi & 0 \\ -\sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{3}$$

Finally, there is a rotation about the new template normal by an angle ϕ. The matrix representation of this rotation is

$$R_\phi = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\phi & \sin\phi \\ 0 & -\sin\phi & \cos\phi \end{pmatrix} \tag{4}$$

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. Here we place the origin at a fixed distance along the projection of the chin-lip line. This point corresponds to the perpendicular

projection of the nose onto the bilateral symmetry axis of the face. Once the origin has been established, the angle θ, i.e. the rotation of the symmetry axis about the z-axis defined by the nose, can be estimated from the orientation of bilateral symmetry axis on the image plane. Once these constraints have been exploited, the orthographic projection can be viewed as slanting and tilting the planar component of the template by angles ψ and ϕ. The net effect is just to subject the eye-lip-chin plane to affine shear. In other words we have only to recover the slant and tilt parameters ϕ and ψ together with the overall scale s. 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 parameter vector $\Phi = (s, \phi, \psi)^T$.

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, Laird and Rubin as a means of fitting incomplete data [3]. The algorithm has two stages. The expectation step involves estimating a mixture distribution using current parameter values. The maximisation step involves computing new parameter values that optimise the expected value of the weighted data likelihood. This two-stage process is iterated to convergence. Although the EM algorithm has been exploited in the matching of 2D shape models [11] and in recovery of object pose by Hornegger and Nieman [6], the main contribution of the this paper is to demonstrate the effectiveness of the algorithm in matching generic facial templates to poorly localised feature-points.

3.1 Expectation

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 2D facial feature locations $\mathbf{w}_i$ 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.

$$p(\mathbf{w}|\Phi^{(n)}) = \prod_{i \in \mathcal{D}} p(\mathbf{w}_i|\Phi^{(n)}) \tag{5}$$

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

$$p(\mathbf{w}_i|\Phi^{(n)}) = \sum_{j \in \mathcal{M}} p(\mathbf{w}_i|\mathbf{v}_j, \Phi^{(n)}) P(\mathbf{v}_j|\Phi^{(n)}) \tag{6}$$

The ingredients of the above mixture density are the component conditional measurement densities $p(\underline{\mathbf{w}}_i|\underline{\mathbf{v}}_j, \Phi^{(n)})$ and the mixing proportions $P(\underline{\mathbf{v}}_j|\Phi^{(n)})$. The conditional measurement densities represent the likelihood that the 2D facial feature location $\underline{\mathbf{w}}_i$ originates from the 3D template control point indexed j under the prevailing set of transformation parameters $\Phi^{(n)}$. We use the shorthand notation $\alpha_j^{(n)} = P(\underline{\mathbf{v}}_j|\Phi^{(n)})$ 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

$$L(\Phi^{(n)}, \mathbf{w}) = \sum_{i \in \mathcal{D}} \ln p(\underline{\mathbf{w}}_i|\Phi^{(n)}) \tag{7}$$

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 3D face template. In other words we need to average the log-likelihood over the space of potential correspondence matches. In fact, it was Dempster, Laird and Rubin [3] 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

$$Q(\Phi^{(n+1)}|\Phi^{(n)}) = \sum_{i \in \mathcal{D}} \sum_{j \in \mathcal{M}} P(\underline{\mathbf{v}}_j|\underline{\mathbf{w}}_i, \Phi^{(n)}) \ln p(\underline{\mathbf{w}}_i|\underline{\mathbf{v}}_j, \Phi^{(n+1)}) \tag{8}$$

The *a posteriori* probabilities $P(\underline{\mathbf{v}}_j|\underline{\mathbf{w}}_i, \Phi^{(n)})$ 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 equation (6)

$$P(\underline{\mathbf{v}}_j|\underline{\mathbf{w}}_i, \Phi^{(n)}) = \frac{\alpha_j^{(n)} p(\underline{\mathbf{w}}_i|\underline{\mathbf{v}}_j, \Phi^{(n)})}{\sum_{j' \in \mathcal{M}} \alpha_{j'}^{(n)} p(\underline{\mathbf{w}}_i|\underline{\mathbf{v}}_{j'}, \Phi^{(n)})} \tag{9}$$

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

$$\alpha_j^{(n+1)} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} P(\underline{\mathbf{v}}_j|\underline{\mathbf{w}}_i, \Phi^{(n)}) \tag{10}$$

In order to proceed with the development of the facial template registration process we require a model for the conditional measurement densities, i.e. $p(\underline{\mathbf{w}}_i|\underline{\mathbf{v}}_j, \Phi^{(n)})$. 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 jth template point delivered by the current estimated transformation parameters. Accordingly we write

$$p(\underline{\mathbf{w}}_i|\underline{\mathbf{v}}_j, \Phi^{(n)}) = \frac{1}{(2\pi)^{\frac{3}{2}}\sqrt{|\Sigma|}} \exp\left[-\frac{1}{2}\epsilon_{i,j}(\Phi^{(n)})^T\Sigma^{-1}\epsilon_{i,j}(\Phi^{(n)})\right] \tag{11}$$

In the above expression Σ is the variance-covariance matrix for the vector of error-residuals $\epsilon_{i,j}(\Phi^{(n)}) = \underline{\mathbf{w}}_i - \underline{\mathbf{u}}_j(\Phi^{(n)})$ between the components of the predicted template point positions $\underline{\mathbf{u}}_j(\Phi^{(n)})$ and the facial feature locations in the data, i.e. $\underline{\mathbf{w}}_i$. Formally, the matrix is related to the expectation of the outer-product of the error-residuals i.e. $\Sigma = E[\epsilon_{i,j}(\Phi^{(n)})\epsilon_{i,j}(\Phi^{(n)})^T]$. With these ingredients, and using the shorthand notation $q_{i,j}^{(n)} = P(\underline{\mathbf{v}}_j|\underline{\mathbf{w}}_i, \Phi^{(n)})$ for the *a posteriori* matching probabilities, the expectation step of the EM algorithm simply reduces to computing the weighted squared error criterion

$$Q'(\Phi^{(n+1)}|\Phi^{(n)}) = -\frac{1}{2}\sum_{i\in\mathcal{D}}\sum_{j\in\mathcal{M}} q_{i,j}^{(n)}\epsilon_{i,j}(\Phi^{(n)})^T\tilde{\Sigma}^{-1}\epsilon_{i,j}(\Phi^{(n)}) \tag{12}$$

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.

3.2 Maximisation

The maximisation step aims to locate the updated the parameter-vector $\Phi^{(n+1)}$ that optimises the quantity $Q(\Phi^{(n+1)}|\Phi^{(n)})$, i.e. $\Phi^{(n+1)} = \arg\max_{\Phi} Q'(\Phi|\Phi^{(n)})$. We solve the implied weighted least-squares minimisation problem using the Levenberg-Marquardt technique. This non-linear optimisation technique offers a compromise between the steepest gradient and inverse Hessian methods. The former is used when close to the optimum while the latter is used far from it. In other words, when close to the optimum, parameter updating takes place with step-size proportional to the gradient $\nabla_{\Phi}Q'(\Phi|\Phi^{(n)})$. When far from the optimum the optimisation procedure uses second-order information residing in the Hessian, H, of $Q'(\Phi|\Phi^{(n)})$; the corresponding step-size for the parameter vector Φ is $H^{-1}\nabla_{\Phi}Q(\Phi|\Phi^{(n)})$.

4 Experiments

The evaluation of our pose estimation procedure involves experiments on both contrived and natural imagery. The 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 15 different camera views for each of eight different individuals. Here we experiment with both hand-labelled together with automatically segmented and labelled feature points. The segmentation process uses Fourier-domain matched filters to characterise each of eight facial features (left and right eyebrows, left and right eye centres, hairline, nose, lips and chin) [1]. When averaged over the eight feature types, the feature localisation error is about 5 pixels. However,

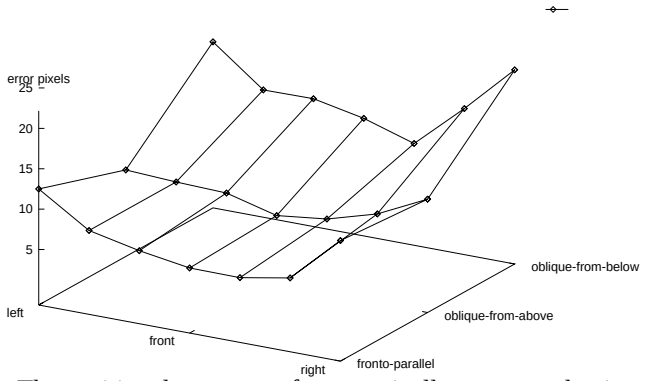

Figure 2: The positional accuracy of automatically segmented points.

for certain features (e.g. the eye centres) the localisation error is about 3 pixels. These sensitivity systematics are summarised in Figure 2 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. Figure 3 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 degrees 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 Figure 4. 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 degrees. Secondly, the error increases dramatically for rotation angles greater than 4 degrees. Finally, there appears to be a positive bias to the 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, Figure 5 shows the feature template converging on the labelled feature 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.

Finally, we focus on how the template registration method degrades when automatically segmented, rather than hand labelled feature points, are used. Figure 6 shows a

(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

Figure 3: 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”.

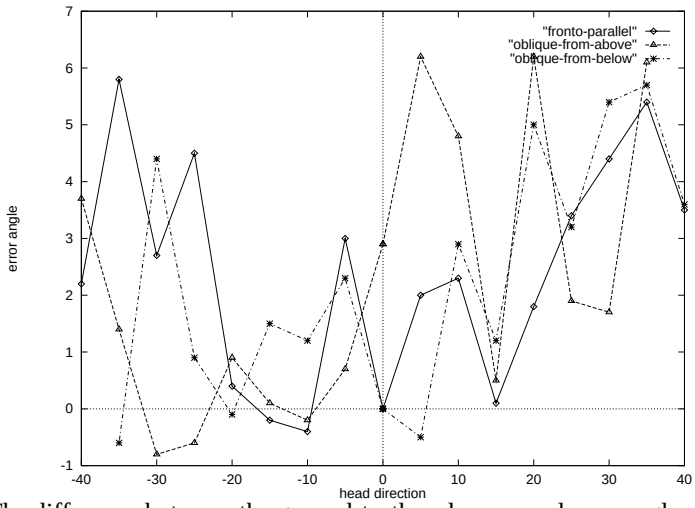

Figure 4: The difference between the ground-truth and recovered pose angles as a function of the ground-truth angle.

(a) initial step (b) final step (c) initial step (d) final step

Figure 5: The 3D feature template converging on the labelled feature points.

plot of the difference in computed rotation angle for the hand-labelled and automatically segmented 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 degrees.

5 Conclusions

The main contribution in 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 degrees provided that the overall rotation does not exceed 40 degrees. The main limitation of the method is the need for accurately located feature points. Our next steps will be focussed on improving the robustness of localisation process. Here we aim to couple the feature segmentation and pose estimation steps of the algorithm.

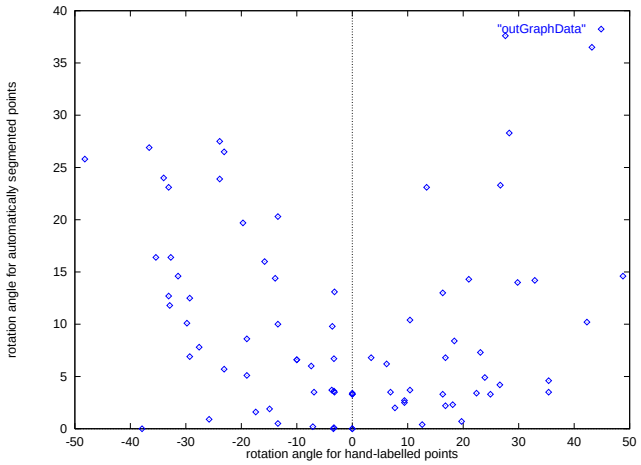

Figure 6: The difference in computed rotation angle for the hand-labelled and automatically segmented points.

References

[1] Choi K.N., Cross A.D.J. and Hancock E.R., "Localising Facial Features with Matched Filters", *First International Conference on Audio- and Video-based Biometric Person Authentication, LNCS, 1206*, pp. 11-20, 1997.

[2] Cross A.D.J. and Hancock E.R., "Recovering Perspective Pose with a Dual Step EM Algorithm, *Advances in Neural Information Processing Systems 10, Editted by M. Jordan, M. Kearns and S. Solla*, MIT Press to appear, 1998.

[3] Dempster A.P., Laird N.M. and Rubin D.B., "Maximum-likelihood from incomplete data via the EM algorithm", *J. Royal Statistical Soc. Ser. B (methodological)*, **39**, pp 1-38, 1977.

[4] Gee A.H. and Cipolla R., "Determining the Gaze of Faces in Images", *Image and Vision Computing*, **12**, pp. 639–647, 1994.

[5] Gee A.H. and Cipolla R., "Fast Visual Tracking by Temporal Consensus", *Image and Vision Computing*, **14**, pp. 105–114, 1996.

[6] Hornegger J. and Niemann H., "Statistical Learning Localisation and Identification of Objects" *Proceedings Fifth International Conference on Computer Vision*, pp. 914–919, 1995.

[7] Kotropoulos C., Pitas I., Fischer S. and Duc B., "Face Authentication Using Morphological Dynamic Link Architecture", *LNCS 1206*, pp. 169-176, 1997.

[8] Lanitis A., Taylor C.J. and Cootes T.F., "Automatic Interpretation and Coding of Face Images using Flexible Models", *IEEE PAMI*, **19**, pp. 743–756, 1997.

[9] Moghaddam B. and Pentland A., "Probabilistic Visual Learning for Object Detection", *Proceedings of the Fifth International Conference on Computer Vision*, pp. 786–793, 1995.

[10] Moss S. and Hancock E.R., "Registering Incomplete Radar Images with the EM Algorithm", *Image and Vision Computing*, **15**, pp. 637–648. 1997.

[11] Revow M., Williams C.K.I. and Hinton G.E., "Using Generative Models for Handwritten Digit Recognition", *IEEE PAMI*, **18**, pp. 592–606 , 1996.

[12] Sullivan G.D., Baker K.D., Worrall A.D., Attwood C.I. and Remagnino P.M., "Model-based vehicle detection and classification using orthographic approximations", *Image and Vision Computing*, **15**, pp. 649-654, 1997.