Pose-Invariant Face Matching Using MRF Energy Minimization Framework

Shervin Rahimzadeh Arashloo and Josef Kittler

Center for Vision, Speech and Signal Processing
University of Surrey
Guildford GU2 7XH, United Kingdom

Abstract. A pose-invariant face verification system based on an image matching method is presented. The method uses the normalized energy of the established match between images as a measure of goodness-of-match. The method can tolerate moderate global spatial transformations between the gallery and the test images and alleviates the need for geometric and photometric normalization of facial images. It requires no training on non-frontal face images. A number of innovations, such as a dynamic block size and block shape adaptation, as well as label pruning and error prewhitening measures have been introduced to increase the effectiveness of the approach. The experimental evaluation of the method is performed on the rotation shots of the XM2VTS database and promising results are obtained.

1 Introduction

In spite of the impressive progress in face recognition technology, many problems still remain unsolved. The two most challenging requirements for a recognition system to operate in real world conditions are invariance to facial image pose and illumination [23]. A variety of methods have been proposed to deal with the problem of pose changes. One of the earlier attempts to overcome the pose variation problem is the work of Beymer [1] in which images of different views of subjects were stored in a database and every input image was first aligned with the relevant reference images from the database and then a similarity measure was computed for recognition. There are also other works which take advantage of multiple images corresponding to different poses in the gallery *e.g.* the method by Singh et al. [17] which constructs composite images using semi-profile and frontal views. Other work by Pentland *et al.* [13] and Wiskott *et al.* [21] can be considered as relatively robust feature-based methods which could tolerate moderate pose variations. There are also 2D learning-based algorithms that try to synthesize a virtual frontal view in the 2D domain. The active appearance model [3] is a well known example of this category. Other techniques which use 3D methods to construct a novel view of the face image form another category. One of the most successful methods in this category is the 3D morphable model proposed by Blanz and Vetter [2]. In parallel with the methods which try to synthesize a novel view of the face, using either one or multiple images from the

D. Cremers et al. (Eds.): EMMCVPR 2009, LNCS 5681, pp. 56–69, 2009.

gallery, there are other methods which try to learn the most discriminant information between classes across different poses. The work by Kim and Kittler [7] and Kanade and Yamada [6] are some examples of this category. In another work by Kim and Kittler [8] authors have tried to make use of different pose-invariant face recognition experts in a multiple classifier fusion system framework.

In short, pose-invariant face recognition systems fall broadly into three categories. The first group are those which try to synthesize novel views, either in 2D or 3D. The methods, which try to infer the most discriminatory information across different poses between distinct classes, constitute the second category. There are also some methods which make use of multiple images of different poses in the database. The algorithms in the last group fail when only one image per subject is available in the database. The main drawback of the methods which synthesize novel views is the imperfection of the synthesizing process in addition to the requirement for prior labeling of landmarks which is usually carried out manually. An observation regarding the 3D methods is that although these techniques perform slightly better than the 2D alternatives, they still suffer from unresolved problems. The most important one is that in 3D-based geometric normalization methods, the recovered shape and texture are completely determined by the 3D morphable face model fitted to the query 2D face image which has the capacity to reconstruct only the information captured during statistical learning. As a result, these approaches can not recover atypical features that have not been available in the training set. Moreover, the high computational complexity of 3D methods in comparison with 2D algorithms makes them unsuitable for real-time applications.

In this work we propose a face recognition system which operates on 2D images. The images are first matched densely and then a similarity criterion defined as the normalized energy of the match is used to judge the goodness-of-match. The method excludes the need for geometric pre-processing of images by encapsulating a matching stage as part of the method. The underlying idea in this work is not new. Similar approaches have been employed especially in general object recognition systems. In a general object recognition setting, it is commonly believed that in the presence of varying illumination, partial occlusion, change of viewing angle, cluttered background, change of scale *etc.*, graph-based techniques perform better in comparison to other alternatives. In a graph matching approach, the concepts of interest are assumed to be built up from simple neighboring primitives. The primitives are coded as the nodes of a graph while the edges convey the neighborhood structure and contextual dependencies. In the area of face recognition and authentication, this approach has been previously attempted by a number of researchers. In [11] the authors have used a dynamic link architecture to construct model and scene graphs. In a later work [21], the authors extended their previous work in [11] by performing the matching twice. In the first stage, the location and size of the face are estimated. The second matching is performed to find the exact location of fiducial points on the test image. Measuring the similarity of the test image to the models of the database was performed using only the node attributes without taking into account the structure (distortion) of the underlying graph explicitly. In [20] an extension to the previous method in [21] was proposed to identify special characteristics of the unknown facial image. In another similar work [10], a graph matching scheme was proposed in which instead of Gabor wavelet filter outputs, multi-scale morphological operators were employed as node attributes. In order to take into account different discriminatory capabilities of nodes, a weighting scheme was employed. The work presented in [18] is very similar to [10] but the authors here have tried to estimate the node weights by reformulating Fisher's discriminant ratio as a quadratic optimization problem which is then solved by combining statistical pattern recognition methods and support vector machines. There are also some other approaches in the context of facial image analysis which use graph theoretic methods to recognize expressions, *e.g.* in [19].

2 Contributions

In this work a method for verification of facial images under varying pose, based upon the method in [15] is presented. The method takes advantage of an image matching method [16] for establishing correspondences between images and formulates the similarity criterion between objects as a combination of the normalized distortion energy of the match as well as texture similarities.

The contributions of the present work and modifications to the matching method in [16] can be outlined as below.

- In order to cope better with matching under different viewing angles, a dynamically deformable block matching method is proposed. In the new generalized block matching scheme, blocks are neither of the same size, nor the same shape. Blocks are deformed according to a global projective transformation estimated between the two images. Accordingly, a square block on the model image is matched to a patch of pixels in the scene image whose shape and area are determined based on the global transformation. The new matching scheme allows much denser sampling of the areas of the face which have undergone contraction and coarser sampling in the areas of expansion as a result of changes in head pose. It has been found that much better matches can be established using the new method.
- In the case of a pan movement of the subject's head, only a half of the face is used for recognition. It is shown experimentally that the visible half of the face contains much more useful shape information and is superior to the whole face.
- The data term has been truncated to achieve more robustness against matching of outliers or occlusions.
- Since the matching method should be able to match facial images of different subjects under varying pose, in order to achieve more flexibility in the deformation, the binary hard constraints are replaced by quadratic penalty functions.
- In order to cope better with illumination changes, the data term has been computed using edge maps. The data term is defined as a combination of normalized vertical and horizontal edge magnitudes.

– The method in [16] used distance transforms [4] to compute messages in linear time. Here, additional speed gain is achieved by pruning unlikely labels at the node level during optimization.

The paper is organized as follows: In Section 3, the image matching method in [16] is overviewed. In Section 4, a new deformable block matching method is introduced. The method incorporates a label-pruning heuristic to speed up the matching process. A similarity criterion for assessing the quality of a match is presented in Section 5. The results of an experimental evaluation of the method on the rotation shots of XM2VTS database are presented and discussed in Section 6. Section 7 concludes the paper.

3 Image Matching

There are different methods for image matching proposed in the literature. In fact, for recognition purposes, the following properties are desirable in an image matching method:

The method should support large displacements to allow the matching of images taken at different viewing angles and scales. It is also important to achieve good solutions in a reasonable time, *i.e.* the method adopted should be efficient enough so that it can be used for recognition purposes in a large database of images. Both objectives can be realized by taking advantage of recent optimization techniques for MRFs [16].

The efficiency of the image matching technique in [16] is based on the fact that disparities in two directions are modeled by two fields interacting together rather than coding them in a single MRF. Additional efficiency was gained through the application of a fast energy minimization technique [9] and updating messages using distance transforms [4]. In the following we briefly review the method used for image matching followed by an overview of the object recognition method in [15].

3.1 Preliminaries

Many computer vision problems can be formulated in an energy minimization framework where the objective function takes the following form:

$$E(X|\theta) = \sum_{s\in\nu} \theta_s(x_s) + \sum_{(s,t)\in\epsilon} \theta_{st}(x_s, x_t) \tag{1}$$

ν corresponds to sites and ϵ to edges. x_s denotes the label of site $s \in \nu$. θ defines the parameters of the energy: θ_s denotes unary data penalty functions whereas θ_{st} denotes pairwise potentials. It is worth noting that in this formulation only cliques of size up to two are considered.

The minimum energy in equation (1) corresponds to the maximum probability of a Gibbs distribution. According to the Hammersley-Clifford theorem, the configuration of a set of sites with respect to the neighborhood system adopted, is an MRF if and only if it is a Gibbs random field with respect to the same neighborhood system. Thus, the solution on an MRF can be considered as the configuration of a Gibbs distribution with maximum probability or inversely as the configuration with minimum posterior energy.

3.2 Decomposed Model

The method proposed in [16] formulates the image matching as a labeling problem on MRFs with the label set $L_{reg} = \{(x_{s^1}, x_{s^2}) | x_{s^1}, x_{s^2} \in L\}$ where x_{s^1} and x_{s^2} denote displacements in horizontal and vertical directions. In fact this technique models the deformation in horizontal and vertical directions by two MRFs interacting together. The edge set of this model is comprised of two separate edge sets(inter-layer and intra-layer edges). The edge potential functions on each of these layers are assumed to be identical (intra-layer edges) while inter-layer edges encode the data term. For the intra-layer edges the following crisp continuity terms are adopted:

$$\theta_{st}(x_s, x_t) = \begin{cases} 0, & x_s = x_t, \\ c_r, & |x_s - x_t| = 1, \\ \infty, & |x_s - x_t| > 1. \end{cases} \tag{2}$$

In order to achieve more flexibility in deformation, hard continuity terms are replaced by quadratic penalty function:

$$\theta_{st}(x_s, x_t) = c(x_s - x_t)^2 \tag{3}$$

where c is a normalizing constant. In our experiments, setting c to 5×10^{-3} was found to give good results. It should be noted that this value depends on the range of input data (normalized to [-1,1] in our case) and also determines the elasticity of the model. In [16], by restricting the neighboring blocks (blocks are of size 4×4) to have relative displacements of no more than one pixel, the scale changes were limited to [.75,1.25] of the model image size whereas by replacing the hard constraints by a quadratic term a much greater range of scales can be accomodated.

The inter-layer edges encode the data term, *i.e.* the cost of assigning label x_{s^1} in layer one and label x_{s^2} in layer two to two isomorphic nodes of the graph. The data term has been constructed using *block model.* In the block model, the pixels are grouped into non-overlapping blocks which correspond to nodes of the graph. The data term for the block model is defined as below:

$$\theta_{s^1 s^2}(x_{s^1}, x_{s^2}) = \frac{1}{\sigma^2} \mathrm{Dis}(I^1_{\boldsymbol{s}}, I^2_{\boldsymbol{s}+(x_{s^1}, x_{s^2})}), s^1 \in \nu^1, s^2 \in \nu^2 \tag{4}$$

where $I^1_{\boldsymbol{s}}$ is a block on image I^1 and the corresponding block on image I^2 is denoted by $I^2_{\boldsymbol{s}+(x_{s^1}, x_{s^2})}$, which is the block with the coordinates $\boldsymbol{s} + (x_{s^1}, x_{s^2})$, where $\boldsymbol{s}$ is the vector pointing to the position of block $I^1_{\boldsymbol{s}}$. Dis(.,.) is a dissimilarity measure which is defined as the sum of squared differences over the pixels of corresponding blocks. Since edge maps are less affected by unwanted illuminations changes, in order to achieve robustness against changes in illumination, we use horizontal and vertical edge maps instead of grey scale images. Horizontal and vertical edges are normalized to the range [-1,1] and combined to form the data term. The data term then becomes:

$$\theta_{s^1 s^2}(x_{s^1}, x_{s^2}) = \frac{1}{\sigma^2}[\mathrm{Dis}(I^{1h}_{\boldsymbol{s}}, I^{2h}_{\boldsymbol{s}+(x_{s^1},x_{s^2})}) + \mathrm{Dis}(I^{1v}_{\boldsymbol{s}}, I^{2v}_{\boldsymbol{s}+(x_{s^1},x_{s^2})})], s^1 \in \nu^1, s^2 \in \nu^2 \quad (5)$$

where $I^{1h}_{\boldsymbol{s}}$ and $I^{2h}_{\boldsymbol{s}+(x_{s^1},x_{s^2})}$ denote a block in the horizontal edge map of the first image and its corresponding block in the horizontal edge map of the target image respectively. $I^{1v}_{\boldsymbol{s}}$ and $I^{2v}_{\boldsymbol{s}+(x_{s^1},x_{s^2})}$ are defined in a similar way.

Since we are interested in comparing configurational arrangements of the entities of the model and scene images, it is desirable to rely more on common features of the two images and bypass the atypical features which appear only in one image. This can be achieved by ignoring the weak edges and setting those below a threshold to zero and also by truncating the data term. By truncating the data term, the matching becomes more robust to outliers and occlusions.

4 Matching with Deformable Blocks

In [16] it has been assumed that for a block in the model image, there exists a block with the same size which has undergone some translational motion. This assumptions ignores any global geometric transformation between the template and the target which is one of the omnipresent factors when matching objects viewed from different angles. Obviously, those parts of the object closer to the sensing device appear larger than the parts further away. In order to handle this effect it seems appropriate to have much more dense sampling (smaller blocks) in the areas of contraction while coarser sampling (larger blocks) would be sufficient in areas of expansion.

In this work the variation in block sizes is controlled by a global projective transformation. Although in order to estimate a global geometric transformation between two images dense matching is not required and transformation can be estimated using a variety of techniques, we have used the same matching scheme [16] and RANSAC to exclude mismatches and estimated a projective transformation. In the second step of matching, each block on model image is warped according to the estimated global transformation and then the corresponding patch on the scene image is sought. The advantages of this method are two fold. First, as mentioned previously, it supports a more realistic sampling of signals subject to a global transformation. Second, as the global transformation controls and predicts the relative placement of corresponding blocks, the size of the neighborhood (search area) that has to be searched for correspondences can significantly be reduced. This minimizes the computational cost of matching in the second stage.

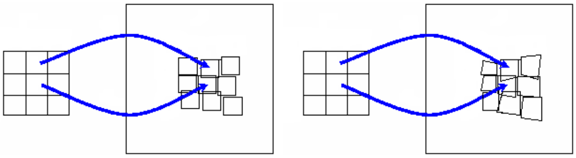

Fig. 1. Left: blocks in [16], Right: blocks in the new deformable block scheme

Considering T:

$$T = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & 1 \end{pmatrix} \tag{6}$$

as the estimated projective transformation between the images, the 2D spatial mapping of blocks then can be interpreted as a combination of projective mapping and translational motion:

$$x_{s^1} = (\frac{ax + by + c}{gx + hy + 1}) + \hat{x}_{s^1}, x_{s^2} = (\frac{dx + ey + f}{gx + hy + 1}) + \hat{x}_{s^2} \tag{7}$$

where x_{s^1} and x_{s^2} stand for horizontal and vertical displacements and x and y are coordinates of the block center. $\hat{x}_{s^1}$ and $\hat{x}_{s^2}$ are labels which are inferred in the second stage of matching. Since the projective transformation captures the dominant part of motion, the potential range of $\hat{x}_{s^1}$ and $\hat{x}_{s^2}$ can be reduced during second matching, thus reducing the computational cost. Another advantage of the deformable-block matching method is its enhanced robustness against outliers in matching. In practice, the matching is not perfect and there might be parts of the model image which are not matched correctly to the unknown image. By reducing the search region in the second stage of matching and allowing the estimated global spatial transformation to carry the dominant part of the motion, this shortcoming is partly corrected in the new matching scheme.

4.1 Pruning Unlikely Labels

Inference on the constructed MRF is performed using the sequential tree-reweighted message passing method [9] which is built upon the max-product belief propagation of Pearl [12]. In an ideal case, if the algorithm finds the exact solution, choosing the solution would be based on choosing the label which minimizes the cost at each node. Although the label with the minimum cost at each node might not correspond to the best solution when the number of iterations is limited (because the inference is not exact and because of the existence of multiple minima), it is unlikely for a label with a high cost at a node in an intermediate iteration of the algorithm to correspond to the optimal solution at the end of optimization. Based on this observation, one can prune out labels which are unlikely to be optimal at each node (labels with larger costs) and meet only admissible labels at each node during optimization. Pruning unlikely labels reduces the configurational search space, hence speeds up the method. In practice the following heuristic pruning scheme is found to result in reasonable solutions:

After n_1 iterations, prune out up to n_2 least probable labels at each node based on their corresponding costs ensuring that there are at least n_3 labels left at each node.

The choice of n_1, n_2 and n_3 depends on the difficulty of a specific problem. The easier the problem the smaller n_1 and n_3 and larger n_2. Although the pruning might sometimes lead to better results compared to the original method, in a limited number of iterations, it may sometimes introduce a trade off between speed and accuracy.

Using the matching method described, a model image is matched to the test image. Figure (2) shows an example of warping a gallery image (near frontal) to the test image (non-frontal) using the deformable block matching scheme.

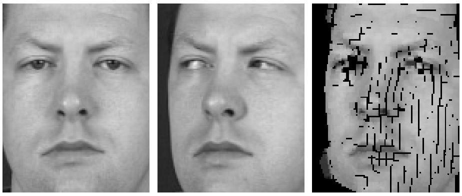

Fig. 2. Left: model image, Middle: scene image, Right: warped model image

Figure (3) shows an example in which the original method in [16] fails to find a correct match especially around the mouth and nose region of the model image. Using the new matching method one can get better matches in the problematic areas. Figure (3) shows the improvement in matching around the mouth and nose of the subject.

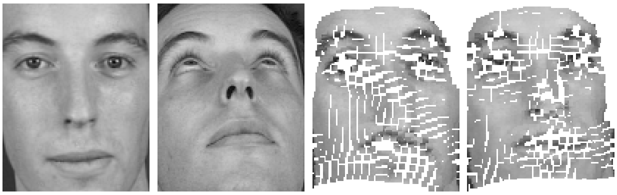

Fig. 3. Left to Right: template, target, result of warping the template using the method in [16], result obtained using deformable-block method

5 Classification

In order to measure the similarity, the two stages are cascaded: first matching the model image image to the unknown image and then computing a similarity/cost function invariant to unwanted global spatial transformation and illumination variations. More explicitly the problem can be described as follows: let I be the image of an ideal subject. Let J be the image of an unknown subject under analysis which depends on its geometrical parameters such as scale s, displacements d_x and d_y, rotation ϕ and perspective effects p, so that $J = J(s, d_x, d_y, \phi, p)$. Let $\mathrm{Dis}(I, J)$ be a dissimilarity function between the ideal image I and unknown image J. The problem is then formulated as calculating:

$$d = \min_{s,d_x,d_y,\phi,p} \mathrm{Dis}(I, J(s, d_x, d_y, \phi, p)) \tag{8}$$

In a hypothesis verification (two class) problem the decision rule is:

$$\text{Assign } J \text{ to class } \omega_r \textit{ iff } \min_{s,d_x,d_y,\phi,p} \mathrm{Dis}(I_r, J(s, d_x, d_y, \phi, p)) < thresh_r$$

where I_r is the template for the r^{th} class and $thresh_r$ is the dissimilarity threshold for the r^{th} class. In the context of recognition using MRFs, a cost function corresponding to the unary and pairwise terms is defined and optimized which is then used in the decision rule. However, the energy obtained in this way has been found not to have enough discriminatory capacity for classification. In [15] factors which unfavorably affect the energy function are identified as pose, non-rigidity of the pattern and last but not least statistical dependencies between residual displacements of neighboring sites and the limited cardinality of the potential functions. In order to remove the effect of the rigid motion, the distortion associated with the global spatial transformation was subtracted from distortion vectors thus achieving global spatial transformation invariance. In order to take into account non-rigidity of patterns, a number of different exemplars of each class were matched one to another and the average distortion was considered as a class-specific model of deformations. The problems of inherent correlation between residual displacements of neighboring sites and of the limited cardinality of the cliques defining the potential functions were partly compensated for by modeling these interactions using covariance matrices which convey correlation information between different sites even at a larger range. In this work we estimate covariance matrices for the full face in the case of tilt movement of head and for the half face in the case of pan movement of the head of the subject.

The structural differences between a pair of images is hence formulated in terms of the Mahalanobis distance:

$$\begin{aligned} D_{Mahalanobis}(I_i, J) = (\bar{e}_v - \bar{\mu}_{iv})^t {\textstyle\sum}_v^{-1} (\bar{e}_v - \bar{\mu}_{iv}) + \\ (\bar{e}_h - \bar{\mu}_{ih})^t {\textstyle\sum}_h^{-1} (\bar{e}_h - \bar{\mu}_{ih}) \end{aligned} \tag{9}$$

where I_i is a template of class i, $\bar{\mu}_{iv}$ and $\bar{\mu}_{ih}$ are the average distortions for this class in vertical and horizontal directions respectively pursued in a raster scan

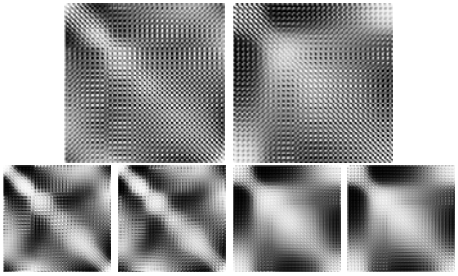

Fig. 4. Covariance matrices of distortions: up row left: full face covariance matrix for vertical direction, up row right: full face covariance matrix for horizontal direction, bottom row from left to right: half face covariance matrices for left half of face in vertical direction, right half of face in vertical direction, left half of face in horizontal direction, right half of face in horizontal direction

fashion. $\bar{e}_v$ and $\bar{e}_h$ are the local distortion vectors obtained after matching I_i to J. In order to obtain the local distortions, after the second stage of matching, another projective transformation is fitted to the set of corresponding points and the effect of rigid motion is subtracted from the distortion field. $\sum_v^{-1}$ and $\sum_h^{-1}$ represent inverse covariance matrices for distortions in vertical and horizontal directions respectively.

5.1 Textural Content

The spatial distortion measure should be complemented by a measure of quality of the match conveyed by the data to refine the cost of match. However, the data term should not be sensitive to unwanted changes in lighting conditions during image capture. Thus, it is essential to use illumination-invariant representation of images for comparison. Local binary patterns have been found to be effective texture descriptors as long as the intensity order of the pixels in a neighborhood is preserved. The textural content of the two images represented as the output of an LBP operator are compared using normalized correlation:

$$NC = \frac{\sum_i \sum_j \sum_h \sum_v (b_{I(i,j)}(h,v) b_{J(i,j)}(h,v))}{\sqrt{\sum_i \sum_j \sum_h \sum_v (b_{I(i,j)}(h,v)^2 b_{J(i,j)}(h,v)^2)}} \tag{10}$$

where $b_{I(i,j)}(h,v)$ is the pixel with horizontal and vertical indices h and v respectively in the block with horizontal and vertical indices i and j in image I. $b_{J(i,j)}(h,v)$ is defined in a similar way.

The final distance measure between a class model (I_i) and the unknown image (J) can be interpreted as a weighted measure of shape and texture distances:

$$D(I_i, J) = w_1(1 - NC) + (1 - w_1)((1 - w_2)D_{Mahal_H} + w_2 D_{Mahal_V}) \tag{11}$$

where D_{Mahal_H} and D_{Mahal_V} correspond to Mahalanobis distance between the two shapes using horizontal and vertical distortions.

6 Experimental Setup

To test our approach to pose-invariant face verification we performed experiments on the XM2VTS data set. This is a multi-modal database which contains color images plus video and sound sequences of 295 subjects. For these experiments we make use of 8 near frontal images for training and 8 rotated head shots per subject, as test images.

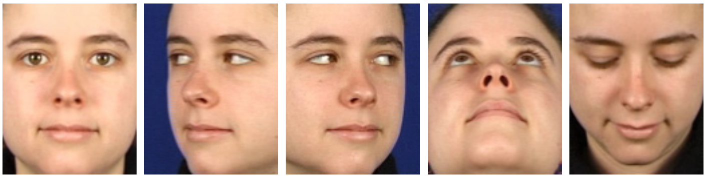

Fig. 5. Example images of frontal and rotated faces in XM2VTS corpus

Some examples of rotated and near frontal images used are shown in Figure 5. The XM2VTS database is divided into a training set of 200 clients, an evaluation set of the same 200 clients plus 25 impostors and a test set of the same 200 clients plus 70 different impostors. In our rotated head shot experiments we used a modified form of the XM2VTS Lausanne test protocol. This retains the same partitioning of subject identities into valid clients and impostors. It differs in that for clients the 2 frontal test images are replaced by 8 rotated head shots (down, left, right, up) and for impostors the 8 frontal images are replaced by 8 rotated head shots. Figure 5 illustrates the severity of the pose deviation from the frontal.

6.1 Results

Two error measures adopted for assessing the performance of the verification system are false acceptance and false rejection rates defined as:

$$FA = EI/I * 100\%, \qquad FR = EC/C * 100\% \tag{12}$$

where I is the number of imposter claims, EI the number of imposter acceptances, C the number of client claims and EC the number of client rejections.

The performance of a verification system is often stated in *Equal Error Rate* (EER) in which the FA and FR are equal and the threshold for the acceptance or rejection of a claimant is set using the true identities of test subjects. In our experiments we use all 8 near frontal images of clients as training images for estimating the covariance matrices. The results of the verification test on images under pan movement of the head using full face and half face shape information are reported in Tables 1 and 2 and the results using texture in Table 4 in which F.F., H.F and Tex. stand for full face, half face and texture respectively.

Table 1. EER for Pan Using Full-Face Shape Information

Horizontal Distortions		Vertical Distortions		Horizontal and Vertical Distortions	
Euc. dist.	Mahal. dist.	Euc. dist.	Mahal. dist.	Euc.n dist.	Mahal. dist.
53.92	31.24	25.13	17.09	25.13	16.86

Table 2. EER for Pan Using Half-Face Shape Information

Horizontal Distortions		Vertical Distortions		Horizontal and Vertical Distortions	
Euc. dist.	Mahal. dist.	Euc. dist.	Mahal. dist.	Euc.n dist.	Mahal. dist.
34.44	31.49	17.44	13.86	17.27	13.77

Table 3. EER for Tilt Using Shape Information

Horizontal Distortions		Vertical Distortions		Horizontal and Vertical Distortions	
Euc. dist.	Mahal. dist.	Euc. dist.	Mahal. dist.	Euc.n dist.	Mahal. dist.
25.80	25.25	42.58	29.37	25.80	24.9

Table 4. EER Using Texture/ Shape and Texture

Pan Movement			Tilt Movement	
F.F. Tex.	H.F. Tex.	F.F. Tex. and H.F. Shape	Tex.	Tex. and Shape
13.25	17.14	9.62	26.5	24.37

The results using the fusion of texture and shape scores are also reported in Table 4.

According to the results in Tables 1, 2 and 4 the following conclusions can be drawn. First, the effect of modeling statistical dependencies in deformation is evident by comparing the results obtained using Euclidean and Mahalanobis distances. Second, in the case of pan movement, distortions in vertical direction are much more discriminative than the distortions in the horizontal direction. Third, using half face image, the configurational information of face images is better represented. This is because the more distant part of the rotated face in the image is partly self-occluded and does not contain useful shape information. The second reason is the inability of a simple projective transformation to model 3D rigid motion of a non-planar object. The results of fusing texture and half face shape are reported in Table 4.

Table 3 reports the results for the tilt movement using shape information. In this case we have used full face matching in our experiments. The effectiveness of the covariance information is again evident. As in the case of pan movement, the distortions perpendicular to the direction of head movement are more discriminative. The result of fusing shape and texture scores is reported in Table 4. From the results it can be concluded that, the recognition of faces subject to pan movement ($EER = 9.62$) is more accurate than that of tilt ($EER = 24.37$). In the case of tilt motion, the self occlusion can not be compensated for by exploiting symmetry. Inevitably this decreases the quality of the match. The results obtained, are considerably better than the state-of-the-art results obtained on the XM2VTS database using AAMs in [5]. In [5] only a subset of the same database was used to test the method whereas here the obtained results are on the whole database. Also the results are significantly better than those obtained in [8] reporting error rates of 30% for the same session and 58% for different sessions on a subset consisting of 125 subjects of the same database. Also, the results compares favorably with the results in [14] which has used only 100 subjects out of 295 in the same database and achieves an error of 23% in recognition.

7 Conclusion

A pose-invariant face recognition system based on an image matching method was presented. The method uses the normalized energy of the established match between images as a criterion for assessing goodness-of-match. The method can tolerate moderate global spatial transformations between the model and the scene object and alleviates the need for geometric normalization of facial images which is commonly required in face recognition. The experimental evaluation of the method was performed on the very challenging rotation shots of the XM2VTS database and promising results were obtained.

References

1. Beymer, D.J.: Face recognition under varying pose. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, June 1994, pp. 756–761 (1994)
2. Blanz, V., Vetter, T.: Face recognition based on fitting a 3D morphable model. IEEE Trans. Pattern Analysis and Machine Intelligence 25(9), 1063–1074 (2003)
3. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Trans. Pattern Analysis and Machine Intelligence 23(6), 681–685 (2001)
4. Felzenszwalb, P.F., Huttenlocher, D.P., Kleinberg, J.M.: Fast algorithms for large-state-space HMMs with applications to web usage analysis. In: NIPS (2003)
5. Guillemaut, J.-Y., Kittler, J., Sadeghi, M., Christmas, W.: General pose face recognition using frontal face model. In: Proceedings of the 11th Iberoamerican Congress in Pattern Recognition, November 2006, pp. 79–98 (2006)
6. Kanade, T., Yamada, A.: Multi-subregion based probabilistic approach toward pose-invariant face recognition. In: Proc. IEEE International Symposium on Computational Intelligence in Robotics and Automation, vol. 2, pp. 954–959 (2003)
7. Kim, T.K., Kittler, J.: Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model Image. IEEE Trans. Pattern Analysis and Machine Intelligence 27(3), 318–327 (2005)
8. Kim, T.K., Kittler, J.: Design and fusion of pose-invariant face-identification experts. IEEE Trans. Circuits and Systems for Video Technology 16(9), 1096–1106 (2006)

9. Kolmogorov, V.: Convergent tree-reweighted message passing for energy minimization. IEEE Trans. on Pattern Recognition and Machine Intelligence 28(10), 1568–1583 (2006)
10. Kotropoulos, C., Tefas, A., Pitas, I.: Frontal face authentication using morphological elastic graph matching. IEEE Trans. on Image Processing 9(4), 555–560 (2000)
11. Lades, M., Vorbruggen, J.C., Buhmann, J., Lange, J., Von der Malsburg, C., Wurtz, R.P., Konen, W.: Distortion invariant object recognition in the dynamic link architecture. IEEE Trans. on Computers 42(3), 300–311 (1993)
12. Pearl, J.: Probabilistic reasoning in intelligent systems: Networks of plausible inference. Morgan Kaufmann, San Francisco (1988)
13. Pentland, A., Moghaddam, B., Starner, T.: View-based and modular eigenspaces for face recognition. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, June 1994, pp. 84–91 (1994)
14. Prince, J.D., Elder, J.H., Warrel, J., Felisberti, F.M.: Tie factor analysis for face recognition across large pose differences. IEEE Trans. Pattern Analysis and Machine Intelligence 30(6), 970–984 (2008)
15. Rahimzadeh Arashloo, Sh., Kittler, J.: On matching criteria for non-rigid object recognition, research report, Centre for Vision, Speech and Signal Processing, University of Surrey (July 2008)
16. Shekhovtsov, A., Kovtun, I., Hlavac, V.: Efficient MRF deformation model for non-Rigid image matching. In: Conf. Computer Vision and Pattern Recognition (CVPR), pp. 1–6 (2007)
17. Singh, R., Vatsa, M., Ross, A., Noore, A.: A mosaicing scheme for pose-invariant face recognition. IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics 37(5) (October 2007)
18. Tefas, A., Kotropoulos, C., Pitas, I.: Using support vector machines to enhance the performance of elastic graph matching for frontal face authentication. IEEE Trans. on Pattern Analysis and Machine Intelligence 23(7), 735–746 (2001)
19. Wang, M., Iwai, Y., Yachida, M.: Expression recognition from time-sequential facial images by use of expression change model. In: Proc. Third IEEE International Conf. on Automatic Face and Gesture Recognition, April 1998, pp. 324–329 (1998)
20. Wiskott, L.: Phantom faces for face analysis. In: Proc. International Conf. on image Processing, October 26-29, vol. 3, pp. 308–311 (1997)
21. Wiskott, L., Fellous, J.M., Kuiger, N., Von Der Malsburg, C.: Face recognition by elastic bunch graph matching. IEEE Trans. Pattern Analysis and Machine Intelligence 19(7), 775–779 (1997)
22. Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G.: XM2VTSDB: The extended m2vts database. In: Proc. Audio- and Video-based Biometric Person Authentication Conf. (1999)
23. Zhao, W., Chellappa, R., Phillips, P.J.: Face recognition: A literature survey. ACM Computing Surveys 35(4), 399–458 (2003)