Energy Normalization for Pose-Invariant Face Recognition Based on MRF Model Image Matching

Shervin Rahimzadeh Arashloo and
Josef Kittler, *Member*, *IEEE*

Abstract—A pose-invariant face recognition system based on an image matching method formulated on MRFs is presented. The method uses the energy of the established match between a pair of images as a measure of goodness-of-match. The method can tolerate moderate global spatial transformations between the gallery and the test images and alleviate the need for geometric preprocessing of facial images by encapsulating a registration step as part of the system. It requires no training on nonfrontal 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 two publicly available databases. First, the method is tested on the rotation shots of the XM2VTS data set in a verification scenario. Next, the evaluation is conducted in an identification scenario on the CMU-PIE database. The method compares favorably with the existing 2D or 3D generative model-based methods on both databases in both identification and verification scenarios.

Index Terms—Markov random fields, structural image analysis, image matching, face recognition, pose invariance.

1 Introduction

While the recently reported evaluations [1] confirm that face recognition systems can achieve very high performance in restricted environments, their accuracy can be compromised in more realistic situations. Any changes in the imaging conditions resulting in the acquired image pose being nonfrontal can seriously degrade the performance of face recognition algorithms [1], [2]. The problem is manifested in a variety of applications, such as security and surveillance, online image search and tagging, analysis of personal photos, etc., in which no control can be imposed on the imaging view point. The difficulty arises from the fact that the between-class separability is eroded by the within-class variance attributed to pose changes.

A successful category of pattern classification methods is inspired by the premise that each pattern is built up from simple primitives. Concepts of interest are modeled by different configurations of symbolic interconnecting parts forming graphs and their essence is conveyed by a number of exemplars/prototypes. The statistical verifiability of these structures is typically modeled by Markov random fields (MRF). In a recognition scenario, the goodness of a match is often gauged in terms of the *maximum a posteriori probability* of the corresponding configuration of the underlying graphical model (MRF). Or, conversely, the associated posterior energy is taken as the cost of matching. The current work follows this general approach for recognition/verification of faces under arbitrary pose with the restriction that only frontal images are available as class exemplars. In comparison to the existing approaches, some of the distinguishing characteristics of the proposed method can be outlined as below:

- The proposed method circumvents the need for geometric preprocessing of face images (often done manually) by encapsulating an image matching technique as part of face recognition. As a result, it can cope with moderate translation in and out of plane rotation, scaling, and perspective effects. This is very important as residual misalignments remaining after geometric normalization of face images based on automatic face detection and localization can seriously degrade the performance of a face recognition system. The misalignment problem is particularly pertinent as the automatic detection of facial landmarks used for geometric normalization is especially challenged by pose, lighting, or expression changes.
- Nonfrontal images are not needed for training. This is particularly advantageous in terms of training time and generalization capabilities of the algorithm across different databases.
- In the proposed method, no strict assumption is made about the pose of the subject prior to matching, and hence the system is better suited to more realistic scenarios. In contrast to other solutions, this eliminates the dependency of a face recognition system on the accuracy of the pose estimation module.
- In order to reduce the problems introduced by self-occlusion in the case of a pan movement, only half of the face is used for matching and recognition. The decision of whether there a pan component is present or not is made by comparing the normalized energies of the full-face versus half-face matches.
- In comparison to the state-of-the-art approaches based on 3D models, the proposed approach operates on 2D images, which bypasses the need for 3D face training data and the vagaries of 3D face model to 2D face image fitting.
- We argue that from the point of view of object recognition, the matching energy function in MRF-based approaches has certain drawbacks and should not be used as a similarity criterion for hypothesis selection directly. The main shortcomings of the energy function (using at most pairwise potentials) are identified and a plausible energy normalization scheme is proposed and discussed. In fact, one could directly incorporate a global interaction potential into the underlying MRFs, e.g., as in [3], [4], [5], and optimize the energy including the higher order potential. However, in our case, because of the huge configurational space, which results in inefficient marginalization over the higher order cliques, we propose matching the images using at most pairwise potentials and then normalizing the underlying energy for recognition, which is more efficient. Clearly, the gained efficiency may come at the risk of the quality of match being partially compromised. However, the choice between viability and perfection seems to be rather stark and we have opted for the former.

A preliminary version of the current work, tested on subsampled images, appeared in [6]. This paper introduces a different representation for the face texture, which resulted in improvements in performance. More experiments have been conducted on full size face images, not only on the XM2VTS pose database but also on the PIE database. The analysis and the discussion of the experimental results is more comprehensive. The text has been rewritten to enhance the clarity and completeness of the presentation of the proposed method.

The paper is organized as follows: In Section 2, we briefly review the literature on pose-invariant face recognition methods. In Section 3, the image matching method in [7] along with some modifications is described. In Section 4, the proposed dynamic

- *The authors are with the Center for Vision, Speech and Signal Processing, Faculty of Engineering and Physical Sciences, University of Surrey, Guildford, Surrey, GU27XH UK.*
  E-mail: {sr00048, j.kittler}@surrey.ac.uk.

Manuscript received 9 Oct. 2009; revised 6 May 2010; accepted 22 Sept. 2010; published online 29 Nov. 2010.
Recommended for acceptance by M. Pelillo.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2009-10-0675.
Digital Object Identifier no. 10.1109/TPAMI.2010.209.

block size and shape adaptation scheme is presented. Section 5 discusses a heuristic to speed up the inference method used for optimization. In Section 6, we introduce a classification scheme taking advantage of the proposed energy normalization method. Section 7 presents the experimental setup and the results obtained on the two publicly available databases for face verification and recognition tasks, respectively. In Section 8, conclusions are drawn.

2 Related Work

A variety of different approaches have been proposed to cope with the pose variation problem in face recognition. The early attempts to generalize across pose are the multiview systems, which are direct extensions of the systems operating on frontal images (e.g., the works in [8], [9]). The main drawback of such methods is that they need multiple images of subjects in different poses in the gallery, which might not be available in some scenarios, and also the requirement of a large memory for storage.

Another class of approaches uses a single gallery image for recognition, although, for training, they may need multiple images corresponding to different poses or illumination conditions. These methods can be further classified into two distinct categories. The first class synthesizes virtual views in desired poses using either 2D learning-based methods or 3D model-based approaches, whereas the second class seeks to find an optimal decision rule for better discrimination between classes. As examples of the 2D learning-based methods for virtual view synthesis, one can consider the works in [10], [11], [12]. A commonly followed method for virtual view synthesis is to use 3D models, the best known in this category being the 3D morphable model [13], [14]. These methods still suffer from unresolved problems. Most importantly, the recovered shape and texture in 3D geometric normalization-based approaches are completely determined by the 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 cannot recover atypical features that have not been observed in the training set. Another drawback is the necessity for landmark labeling, which is carried out manually. The second set of methods, which use a single gallery image per subject, can be considered as discriminative approaches. In comparison to the first group, in the second class no virtual views are synthesized. Instead, the features used lie in a pose-invariant space in which the identity is assigned based on some distance designed in this space [15], [16], [17], [18].

There are some methods in the literature similar to the present work which are based on the idea of MRFs and graphical models. However, no attempt is made in these works [19], [20], [21], [22] to recognize faces under severe pose changes.

3 Image Matching

In the context of MRF modeling, individual primitives are modeled as nodes of the graph (also called sites), while edges/hyperedges encode conditional dependencies and the neighborhood structure. The goal is to assign each node/site a label from a predefined admissible discrete set. In the current work, nodes correspond to individual patches in the images and labels to 2D displacement vectors. Two kinds of edges are present. One encodes a smoothness prior on the neighboring nodes in a four-connected neighborhood system in a lattice, while the other carries information about discrepancies between the edge map of the model and those of the scene. The goal is to find an assignment of labels with minimum cost, which is considered as the configuration of a Gibbs distribution with maximum probability. When one only assumes cliques of size up to 2, the posterior energy has the following form:

$$E(X;\theta) = \sum_{s\in V}\theta_s(x_s) + \sum_{(s,t)\in E}\theta_{st}(x_s, x_t). \quad (1)$$

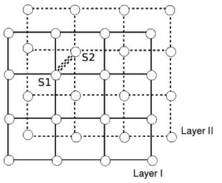

Fig. 1. The structures of MRFs used.

V is the set of sites and E represents the set of edges. θ parameterizes the energy and is dependent on the application in hand. In the following sections, we discuss how θ is defined for our task.

There are no constraints on selecting the matching method provided it is efficient and guarantees certain optimality conditions. While there are different image matching methods in the literature [23], [24], [25], we have adapted the one in [7] for our purposes because of its efficiency and the benefit of employing a successful optimization method [26] which outperforms others in a number of different tasks [27].

In [7], Shekhovtsov et al. formulate the image matching as a labeling problem on two interacting MRFs together. The label set in this case is defined as

$$L = \{(x_{s^1}, x_{s^2}) | x_{s^1} \in L_{s^1}, x_{s^2} \in L_{s^2}\}, \quad (2)$$

where L_{s^1} and L_{s^2} represent discrete label sets corresponding to disparities in horizontal and vertical directions. The edge set of this model is comprised of two separate edge sets as interlayer and intralayer edges, and the node set is comprised of nodes on two layers:

$$V = v^1 \cup v^2, E = e^1 \cup e^2. \quad (3)$$

The MRFs structure in the decomposed model is illustrated in Fig. 1. The efficiency of this method draws on the idea of *decomposition* of the 2D label set into two 1D distinct label sets, each representing the label set for one MRF. In the following, the smoothness and data term used in the model are discussed in detail.

3.1 Smoothness Prior

The intralayer edges in [7] encode a smoothness prior and are defined as

$$\theta_{st}(x_s, x_t) = \begin{cases} 0, & x_s = x_t, \\ c_r, & |x_s - x_t| = 1, \\ \infty, & |x_s - x_t| > 1. \end{cases} \quad (4)$$

In order to achieve more flexibility in deformation, in the current work we replace the crisp continuity terms by a quadratic penalty function:

$$\theta_{st}(x_s, x_t) = \rho(x_s - x_t)^2, \quad (5)$$

where ρ is a normalizing constant.

3.2 Data Term

The interlayer 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 a *block model*. In the block model, the pixels are grouped into nonoverlapping blocks. The data term for the block model is defined as

$$\theta_{s^1 s^2}(x_{s^1}, x_{s^2}) = \frac{1}{2\sigma^2}\mathrm{Dis}\big(I^1_{\bar{s}}, I^2_{\bar{s}+(x_{s^1},x_{s^2})}\big), \quad s^1 \in v^1, s^2 \in v^2, \quad (6)$$

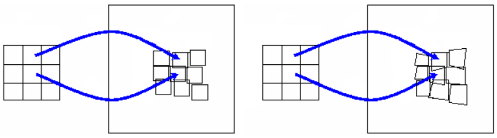

Fig. 2. Left: Blocks in [7]. Right: Blocks in the new deformable block scheme.

where $I^1_{\vec{s}}$ is a block in image I^1 and the corresponding block in image I^2 is denoted by $I^2_{\vec{s}+(x_{s^1},x_{s^2})}$, which is the block with the coordinates $\vec{s}+(x_{s^1},x_{s^2})$, where $\vec{s}$ is the vector pointing to the position of block $I^1_{\vec{s}}$. $\mathrm{Dis}(.,.)$ is a dissimilarity measure which is defined as the sum of squared differences of colors over the pixels of corresponding blocks and σ is the standard deviation of noise.

Since edge maps are less affected by unwanted illumination changes, we use horizontal and vertical edge maps instead of color or gray scale images. Horizontal and vertical edges are scaled to the range $[-1,1]$ and combined to form the data term. The data term is defined as

$$\begin{aligned}\theta_{s^1 s^2}(x_{s^1},x_{s^2}) = \frac{1}{2\sigma^2}\big[&\mathrm{Dis}\big(I^{1h}_{\vec{s}}, I^{2h}_{\vec{s}+(x_{s^1},x_{s^2})}\big)\\ &+\mathrm{Dis}\big(I^{1v}_{\vec{s}}, I^{2v}_{\vec{s}+(x_{s^1},x_{s^2})}\big)\big], s^1 \in v^1, s^2 \in v^2,\end{aligned} \tag{7}$$

where the same notation is used as in (6), except here we use edge maps instead of images. In our experiments, setting the value of the constant term ρ in the pairwise potentials to 5×10^{-3} was found to give reasonable results. The data term is then truncated to make the matching more robust to outliers, occlusions, and spurious structures.

4 Block Adaptation

The work in [7] assumes that for a block in the model image, there exists a block with the same size and shape which has only undergone some translational motion. This assumption ignores any global geometric transformation between the template and the target images. In order to consider the effects of a global transformation, it seems appropriate to have much more dense sampling in the areas of contraction, while coarser sampling would be sufficient in areas of expansion. In contrast to some other approaches (e.g., the work in [28]) which use training data to estimate the warp between the frontal and nonfrontal blocks, the adaptation proposed here does not require any training.

The block size and shape is controlled using a global projective transformation between two images. In order to estimate it, the method described earlier is used to find a set of corresponding points between the two images. Next, a global spatial transformation is estimated using the Levenberg-Marquardt method [29], along with RANSAC to exclude mismatches, if any. In the second round of matching, each block in the model image is warped according to the estimated transformation and the corresponding patch on the target image is sought. The proposed block adaptation method supports a more realistic sampling of signals subject to a global transformation while reducing the possibility of mismatches. Fig. 2 illustrates the block shape and size adaptation in comparison with the original method.

Considering T:

$$T = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & 1 \end{pmatrix}, \tag{8}$$

the 2D spatial mapping of blocks can be interpreted as a combination of projective mapping and translational motion

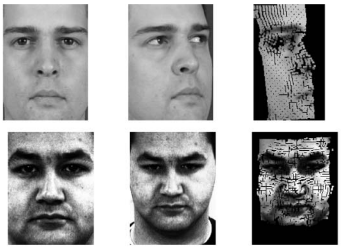

Fig. 3. In each row from left to right: template image, target image, and deformed template image. (In the first row, half of the template image is used in matching).

$$x_{s^1} = \left(\frac{ax+by+c}{gx+hy+1}\right) + \hat{x}_{s^1}, x_{s^2} = \left(\frac{dx+ey+f}{gx+hy+1}\right) + \hat{x}_{s^2}, \tag{9}$$

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 and making the method more robust against outliers.

5 Speeding up Inference by Label Pruning

Choosing the solution in the optimization method used [26] is based on choosing the label which minimizes the cost at each node subject to the contextual relationships. Although the label with the minimum cost at each node might not correspond to the best solution when the number of iterations is limited, 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, we prune out labels which are less likely to be optimal using the following heuristic:

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 task. While the inference using tree reweighted message passing is based upon linear programming *relaxation*, the heuristic pruning method acts as a hard propagator to speed up the process. Although label set pruning has been applied in other optimization approaches [30], it has not been considered in the context of linear programming relaxation using tree reweighting schemes. Applying the above pruning method, we achieved up to 30 percent speed-up on average.

Some examples of matching using the deformable block matching method are illustrated in Fig. 3.

6 Classification

In the context of recognition using MRFs, a cost function corresponding to the unary and pairwise relations is first optimized. Then, the result would normally be used as a basis for decision making. However, the energy obtained in this way does not have enough discriminatory information to support classification because of the following factors:

- The matching criterion, which partly gauges the geometric distortion, includes a global rigid transformation as well as local object shape deformations. For object recognition, only the latter is of importance.
- Restricting both the neighborhood system of sites in an MRF to a limited spatial range as well as clique cardinality is an essential prerequisite of efficient optimization.

However, this compromises the capacity to capture longer range interactions of object primitives.

- Measuring structural deformation as a function of the regularization term implicitly assumes a simple sum (euclidean distance) as a measure of similarity. This assumption completely ignores any statistical dependencies between deformations of different sites.
- Last but not least, the goodness of match tracked down by the data term in the matching criterion can be dramatically influenced by environmental changes, such as changes in illumination.

In the following, we suggest possible ways of normalizing the cost of matching so that it can serve as a more suitable similarity criterion.

6.1 Structural Dissimilarity

After matching two images, one expects small deformations for the objects of the same class, whereas large deformations are expected when the gallery and the test images do not belong to the same category. Based on this assumption, a shape dissimilarity measure is described next.

Expanding the pairwise interaction term of the energy function for the quadratic penalty function in the considered four-connected neighborhood system on one of the layers (layer one) yields

$$\begin{aligned}\sum_{(s^1,t^1)\in e^1} \theta_{s^1t^1}(x_{s^1}, x_{t^1}) &= 4\rho \sum_{s^1\in In^1} {x_{s^1}}^2 \\ &+ 3\rho \sum_{s^1\in B^1} {x_{s^1}}^2 + 2\rho \sum_{s^1\in C^1} {x_{s^1}}^2 - 2\rho \sum_{(s^1,t^1)\in v^1} x_{s^1} x_{t^1},\end{aligned} \tag{10}$$

where v^1 denotes the sites of layer one and In^1, B^1, and C^1 stand for internal nodes, nodes on the boundaries, and nodes on the corners of layer one, respectively. Considering the deformation in each layer as an interpolation surface which maps each grid location into its corresponding location on the target image, each of the first three terms on the right-hand side of (10) is a weighted measure of deformation of the interpolation surface, which can be considered as a physical measure of similarity between the two objects. In order to obtain local deformations, the effect of rigid motion is removed by fitting a projective transformation to the set of 2D corresponding points, using the Levenberg-Marquardt method [29] and subtracting it from the disparity vectors. Next, in order to take into account the nonrigid nature of faces, we estimate an average distortion for each class (subject) by matching gallery images of each class to each other and considering the average as the mean distortion of that class. The effective distortion energy for a test image can then be computed by matching it to one instance of the target class and subtracting the mean distortion of the target class to obtain the local distortion map. The distortion energy is then expressed as the squared euclidean distance between the structure of the unknown object and the mean distortion of the target class

$$\begin{aligned}E^{Euc}_{distortion} &= (X_1 - \bar{X}_1)^T (X_1 - \bar{X}_1) \\ &+ (X_2 - \bar{X}_2)^T (X_2 - \bar{X}_2).\end{aligned} \tag{11}$$

X_1 and X_2 are the residual raster scanned disparity vectors on the two layers after matching the model to the unknown object, respectively. Also, $\bar{X}_1$ and $\bar{X}_2$ denote the mean disparity vectors of the target class on each of the two layers.

6.1.1 Statistical Dependencies in Local Deformations

Considering the distortion energy, as in (11), ignores any correlations between local deformations. Statistical dependencies between different parts of a warped signal have been studied before in speech recognition [31]. The problem under consideration involved evaluating the similarity of an unknown pattern, representing a segment of a speech signal, to a set of reference prototypes on a frame-by-frame basis. The resulting criterion function value was then used to quantify goodness of match. In this context, it was observed that taking into account statistical dependencies between different frames of signal can improve the performance.

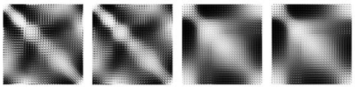

Fig. 4. Sample covariance matrices of local distortions (black: low correlation, white: high correlation). No correlation would have led to diagonal matrices.

The problem under investigation in this work is a 2D counterpart of the classical matching problem in speech. In order to take the effect of error correlations into account, we make use of covariance matrices. As stated earlier, by considering correlation properties of local distortions between all sites instead of only four-connected neighbors, one can also partly compensate for the absence of long-range interaction of sites in the underlying MRFs. Some sample covariance matrices for the vertical and horizontal directions estimated on the XM2VTS [32] database are visualized as images in Fig. 4. In the figure, the brighter areas reflect higher correlation while darker areas represent low correlation. If there were no correlations between deformations of different sites, the correlation coefficients would be nonzero only on the main diagonal. Clearly this is not the case. The structural differences between a pair of images then take the following form:

$$\begin{aligned}E^{Mah}_{distortion} &= (X_1 - \bar{X}_1)^T \Sigma_1^{-1} (X_1 - \bar{X}_1) \\ &+ (X_2 - \bar{X}_2)^T \Sigma_2^{-1} (X_2 - \bar{X}_2),\end{aligned} \tag{12}$$

where Σ_1^{-1} and Σ_2^{-1} represent inverse covariance matrices of the target class for residual distortions in layers 1 and 2, respectively.

6.2 Textural Content

The spatial distortion measure can be complemented by a measure of quality of the match conveyed by the data. In order to remove the effects of uneven illumination, we apply the photometric normalization method in [33] before measuring the texture similarities. The method in [33] is designed to decrease the effects of changes in illumination conditions, highlights, and local shadowing, while keeping the fundamental visual information. In the next step, in order to extract features, we use a local binary pattern operator [34]. The LBP operator is known to be one of the best performing texture descriptors, which, apart being efficient, is highly discriminative. We use uniform LBP patterns in a circular neighborhood. For face description, the face image is divided into different subregions in order to extract local histograms. Once local histograms are extracted from each window, a spatially enhanced histogram is constructed by concatenating local histograms to form a global face descriptor. For the extraction of LBP histograms, a regular division is used for the gallery images while, for nonfrontal test images, we use the information available from the matching to determine the corresponding patch in the test image. Extracted histograms from each region are normalized and concatenated into a single vector and compared using the χ^2 distance:

$$\chi^2(\eta, \xi) = \sum_{b,w} \frac{(\eta_{b,w} - \xi_{b,w})^2}{\eta_{b,w} + \xi_{b,w}}, \tag{13}$$

where η and ξ are the normalized histograms of gallery and test images and b and w stand for the bth bin of the histogram of the wth window in the images. Next, we combine shape and texture terms after normalization to obtain the final distance measure as a weighted measure of shape and texture distances

TABLE 1
The Effects of Block Adaptation and Covariance Estimation on Equal Error Rates Obtained on XM2VTS Corpus Using Shape Information

Pose	Euc.	Euc. with block adap.	Mah. with block adap.
Pan	9.1%	7.5%	5.24%
Tilt	18.5%	16.5%	13.8%

Euc.: euclidean distance, Mah.: Mahalanobis distance

$$D(I_i, J) = \alpha\chi^2 + (1-\alpha)E^{Mah}_{distortion} \tag{14}$$

for $\alpha \in [0, 1]$. $E^{Mah}_{distortion}$ corresponds to the structural distance in (12) and χ^2 represents the textural differences of the images being compared given in (13).

7 Experimental Evaluation

Upon the arrival of an unknown probe image, the method matches the probe image to the frontal gallery images of all classes and the similarity criterion in (14) is used in a nearest neighbor classifier for classification. The performance of the proposed methodology for pose-invariant face recognition is evaluated on two publicly available databases in two different scenarios, described next.

7.1 Verification Test on XM2VTS Database

In the XM2VTS data set [32] the evaluation protocol is based on 295 subjects consisting of 200 clients, 25 evaluation imposters, and 70 test imposters. Two error measures defined for a verification system are false acceptance and false rejection, given below:

$$FA = EI/I * 100\%, \quad FR = EC/C * 100\%, \tag{15}$$

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. Consistent with the definition of EER, parameter α in (14) is set using the true identities of test subjects.

7.1.1 Effects of the Proposed Modifications

Analyzing the effects of error correlation modeling using the covariance matrices shows an 8 percent improvement in error rate on the frontal images of XM2VTS database. On the rotation shots of the same database, the effects of block adaptation and correlation modeling on the error rates using different components of shape distance are reported in Table 1.

From the results it can be observed that block adaptation decreases the overall error obtained using euclidean shape distance by 3.6 percent. By employing the covariance matrices, a further 4.96 percent improvement in error rate is achieved. In total, block adaptation and covariance modeling reduce the *EER* by 4.28 percent, using only shape information.

For texture modeling, we use the Uniform LBP operator with radius 2 and use the smallest resolution available for constructing local histograms (4×4 blocks and their corresponding patches in the test images). The parameters of the illumination normalization method used [33] are set according to [35]. The overall average performance of the system improves with decreasing window sizes. This is understandable as severe pose changes in the image make different parts of the face undergo different appearance variations and, hence, more localized features can provide more discriminatory information. In the case of texture, block adaptation improves error rates by 1.28 percent.

TABLE 2
Comparison of Shape and Texture Information on the XM2VTS Corpus

	Ver.	Hor.	Ver.&Hor.	Texture
Pan	5.74%	10.29%	5.24%	1.0%
Tilt	15.75%	15.99%	13.8%	9.0%

7.1.2 Comparison of Shape and Texture Information

Table 2 provides a comparison between the discriminatory capability of different components of Mahalanobis distance and LBP histograms. Compared to the error rates obtained using shape, one observes that texture seems to be more discriminative. This can be explained from two points of views. First of all, shapes of the faces in the database are more similar. Second, a partial contradictory factor is the inadequacy of a planar transformation (i.e., projective) in modeling rigid motion of the head. Because the face is not planar, one can expect some errors as a result of the planarity assumption being made. Also from Table 2 it can be concluded that the verification of faces subject to pan movement is more accurate than that of tilt because, in the case of tilt motion, the self-occlusion problem cannot be compensated for by exploiting symmetry. Inevitably, this decreases the quality of the match, and hence the performance.

7.1.3 Comparison to a 3D Geometric Normalization-Based Method

In practical applications, the thresholds for acceptance or rejection of a claimant are set on the evaluation set. The performance measure in this case is the *Half Total Error Rate* (HTER), as below:

$$HTER_{FAE=FRE} = \frac{1}{2}(FA_{FAE=FRE} + FR_{FAE=FRE}), \tag{16}$$

where $FAE = FRE$ corresponds to the case when false acceptance and false rejection errors on the evaluation set are equal. In this, the parameter α in (14) is set on the evaluation set and used on the test set. In [36], the authors use a 3D morphable model for geometrically normalizing the rotated images and then use LBP histograms in the 2D geometrically normalized images. The results obtained in [36] and the proposed approach are compared in Table 3. For comparison, the EERs are also included in the same table. From the table, it is observed that the proposed method outperforms the geometric normalization approach using 3D morphable model in [36], both in terms of *EER* and *TER*.

7.2 Identification Test on the CMU PIE Database

7.2.1 Test on Images with Neutral Illumination

In this test, we use images captured under almost the same illumination conditions, with neutral expression, consisting of 884 images of 68 subjects viewed from 13 different angles. Frontal views of subjects (pose 27) are considered as gallery images, while all of the rest (12 different poses) are used as test images. We consider recognition results, using shape and texture separately. The weighting of shape and texture scores in (14) is the same over all poses and is done in such a way that the average overall performance of the system is maximized. From Table 4, the

TABLE 3
Comparison of Performance of the Proposed Method to the Method in [36] on the XM2VTS Database

Method	FAR	FRR	HTER	EER
3D pose correction [36]	0.59	23.25	11.92	7.12
The proposed approach	4.99	11.62	8.30	4.85

TABLE 4
Comparison of the Performance of the Proposed Approach to the State-of-the-Art Methods on the CMU-PIE Database

Pose	C02	C05	C07	C09	C11	C14	C22	C25	C29	C31	C34	C37
Horizontal deviation angle	-44°	-16°	0°	0°	32°	47°	-62°	-44°	17°	47°	66°	-31°
Vertical deviation angle	0°	0°	-13°	13°	0°	0°	1°	11°	0°	11°	1°	0°
eigenlight-fields Complex [37]	58	94	89	94	88	70	38	56	57	56	47	89
PDM [10]	72	100	na	na	94	62	na	na	98	na	20	97
AA-LBP [38]	95	100	100	100	100	91	na	89	100	80	73	100
3D morphable model [39]	76	99	99	99	93	87	50	75	97	78	49	94
Multi-subregion [17]	100	100	100	100	≈ 94	≈ 67	≈ 29	≈ 64	100	≈ 67	≈ 26	100
LLR [11]	na	98	98	98	89	na	na	na	100	na	na	82
Matching [7]	na	85	91	na	na	na	22	na	79	na	25	na
Hor. Mah. distortion	35	73	85	100	58	26	13	35	65	44	10	58
Ver. Mah. distortion	54	72	44	57	66	60	44	60	67	61	47	70
Hor. & Ver. Mah. distortions	66	75	85	100	70	61	48	66	72	64	52	76
Texture	94	97	95	100	86	88	76	94	88	85	76	100
Shape & Texture	95	98	98	100	89	91	79	95	91	88	83	100

following conclusions can be drawn: The horizontal distortion measure can be beneficial in poses where a large pan component is not present (poses C05, C07, C09, and C29). In contrast, the vertical distortion measure is more useful when the head movement contains a pan motion. It can be concluded that the two components complement each other and result in an average identification rate of nearly 70 percent for all poses in the database, using only shape information. In Table 4, we also present the results of fusing texture and shape scores and compare our results to some other approaches, using the original results reported in the literature. Some aspects of the specifications of the approaches are detailed in Table 5. The identification rates reported correspond to using frontal images (pose C27) as gallery images. It can be observed that the proposed technique outperforms most other approaches, and is less restrictive in terms of assumptions. In order to show the merits of the modifications to the matching method, we have included the results obtained using the matching method in [7] for a number of poses. These results are obtained using the method in [7] for matching and keeping all other texture and shape representations similar to the current work. From the results, it can be observed that the modifications improved the performance significantly, especially in extreme poses. The best performing method among other approaches in Table 4 is the method in [38], with an average overall performance of 93.45 percent. Interestingly, the proposed method achieves the same average performance (excluding pose C22), but one needs to take into account the following considerations: The method in [38] uses nonfrontal gallery images, as well as frontal images for training, whereas we do not use any nonfrontal training images. Also, the method in [38] uses 80 manually labeled landmark points, whereas the method proposed here does not need any manually annotated landmarks and only requires the face to be detected in a bounding box, which is much easier than providing landmarks automatically. Other advantages of the method proposed here is that it can also cope with moderate global spatial transformation (e.g., projective) between the images. Our work also outperforms the method in [17], on average by more than 15 percent, in spite of the fact that the approach in [17] uses nonfrontal images for training and uses 34 subjects (half of the database) for test. Also, compared to the algorithm in [11] our method achieves almost the same average performance without using nonfrontal training images, whereas the one in [11] uses such data for training. Considering the performance of other approaches and their specifications (reported in Table 5) in conclusion, our method compares very favorably with most of the existing approaches, with less restrictive assumptions and minimal injection of prior information. The main drawback of the algorithm is the computational complexity of the optimization stage, which is a common characteristic of MRF-based approaches. However, this issue can be partly addressed by employing a sparse MRF model instead of dense image matching methods, as done in the work in [40]. One can also take advantage of parallel processing hardware, such as GPUs, for further speeding-up of the algorithm.

7.2.2 Test on Images under Different Lighting Conditions

In order to determine the failing modes of the algorithm and to evaluate the degradation in system's performance under uneven illumination conditions, in this section we provide the results of a test on a subset of the PIE database consisting of images of 68 subjects captured in three different poses and three different lighting conditions. The images are captured under full profile (pose 22), 3/4 profile (pose 05), and full frontal (pose 27). In each pose, there are images captured with 21 different flashes for each subject, from which we randomly select three different flash conditions for our test. The same set of gallery images is used as in the previous section. Table 6 reports the average recognition rates over different illumination conditions for each pose. The results obtained under neutral illumination conditions are also included for a comparison. In comparison with the recognition rates under neutral illumination conditions, a drop in the system's performance is observed. This is caused by the matching being imperfect due to shadowing effects and also by the inadequacy of the photometric normalization method under severe illumination changes. One option which we aim to investigate in future experiments is to use the method on near infrared images, which are known to be less affected by illumination changes.

8 CONCLUSION

We addressed the pose-invariant face recognition problem within the framework of image matching using MRFs. Using the energy of

TABLE 5
Some Specifications of the Methods in Table 4 and Test Details

Method	Non-frontal training image	no. of landmark points used	no. of subjects used for test
eigenlight-fields Complex [37]	Y	39-54 depending on pose	34
PDM [10]	Y	62	68
AA-LBP [38]	Y	80	68
3D morphable model [39]	3D data	> 6	68
Multi-subregion [17]	Y	> 3	34
LLR [11]	Y	> 2	68
The proposed approach	N	None	68

TABLE 6
Comparison of Performance of the Proposed Method under Neutral Lighting and Variations in Lighting on PIE Database

Pose	05	22	27
Neutral illum.	98	79	na
Varying illum.	71.5	40.2	95.6

the established match between a pair of images, a measure for match quality is formulated and used for classification. 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 accuracy and computational efficiency of the approach. The experimental evaluation of the method performed on two publicly available databases confirmed the effectiveness of the approach. We plan to use a sparse MRF face model and to further accelerate the optimization process using GPUs in order to address the computational burden of the method in the future. Also, to enhance the system's robustness against illumination changes, we shall investigate the merit of the proposed approach on faces acquired with near infrared imaging.

REFERENCES

[1] W. Zhao, R. Chellappa, A. Rosenfeld, and P.J. Phillips, "Face Recognition: A Literature Survey," *Proc. ACM Computing Surveys,* pp. 399-458, 2003.

[2] M. Tistarelli, S. Li, and R. Chellappa, *Handbook of Remote Biometrics for Surveillance and Security.* Springer, 2009.

[3] T. Werner, "High-Arity Interactions, Polyhedral Relaxations, and Cutting Plane Algorithm for Soft Constraint Optimisation (MAP-MRF)," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* pp. 1-8, June 2008.

[4] N. Komodakis and N. Paragios, "Beyond Pairwise Energies: Efficient Optimization for Higher-Order MRFS," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* pp. 2985-2992, 2009.

[5] C. Rother, P. Kohli, W. Feng, and J. Jia, "Minimizing Sparse Higher Order Energy Functions of Discrete Variables," *Proc. IEEE CS Computer Vision and Pattern Recognition,* vol. 0, pp. 1382-1389, 2009.

[6] S.R. Arashloo and J. Kittler, "Pose-Invariant Face Matching Using mrf Energy Minimization Framework," *Proc. Int'l Conf. Energy Minimization Methods in Computer Vision and Pattern Recognition,* pp. 56-69, 2009.

[7] A. Shekhovtsov, I. Kovtun, and V. Hlavac, "Efficient MRF Deformation Model for Non-Rigid Image Matching," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* pp. 1-6, June 2007.

[8] A. Pentland, B. Moghaddam, and T. Starner, "View-Based and Modular Eigenspaces for Face Recognition," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* pp. 84-91, June 1994.

[9] D. Beymer, "Face Recognition Under Varying Pose," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* pp. 756-761, June 1994.

[10] D. Gonzalez-Jimenez and J. Alba-Castro, "Toward Pose-Invariant 2-D Face Recognition through Point Distribution Models and Facial Symmetry," *IEEE Trans. Information Forensics and Security,* vol. 2, no. 3, pp. 413-429, Sept. 2007.

[11] X. Chai, S. Shan, X. Chen, and W. Gao, "Locally Linear Regression for Pose-Invariant Face Recognition," *IEEE Trans. Image Processing,* vol. 16, no. 7, pp. 1716-1725, July 2007.

[12] T. Cootes, K. Walker, and C. Taylor, "View-Based Active Appearance Models," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition,* pp. 227-232, 2000.

[13] V. Blanz, S. Romdhani, and T. Vetter, "Face Identification Across Different Poses and Illuminations with a 3d Morphable Model," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition,* pp. 192-197, May 2002.

[14] X. Liu and T. Chen, "Pose-Robust Face Recognition Using Geometry Assisted Probabilistic Modeling," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 502-509, June 2005.

[15] T.-K. Kim and J. Kittler, "Locally Linear Discriminant Analysis for Multimodally Distributed Classes for Face Recognition with a Single Model Image," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 3, pp. 318-327, Mar. 2005.

[16] J. Huang, P. Yuen, W.-S. Chen, and J.H. Lai, "Choosing Parameters of Kernel Subspace lDA for Recognition of Face Images under Pose and Illumination Variations," *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics,* vol. 37, no. 4, pp. 847-862, Aug. 2007.

[17] T. Kanade and A. Yamada, "Multi-Subregion-Based Probabilistic Approach toward Pose-Invariant Face Recognition," *Proc. Int'l Symp. Computational Intelligence in Robotics and Automation,* vol. 2, pp. 954-959, July 2003.

[18] C. Castillo and D. Jacobs, "Using Stereo Matching for 2-D Face Recognition Across Pose," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* pp. 1-8, June 2007.

[19] R. Wang, Z. Lei, M. Ao, and S.Z. Li, "Bayesian Face Recognition Based on Markov Random Field Modeling," *Proc. Third Int'l Conf. Advances in Biometrics,* pp. 42-51, 2009.

[20] B.-G. Park, K.-M. Lee, and S.-U. Lee, "Face Recognition Using Face-ARG Matching," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 12, pp. 1982-1988, Dec. 2005.

[21] R. Huang, V. Pavlovic, and D. Metaxas, "A Hybrid Face Recognition Method Using Markov Random Fields," *Proc. Int'l Conf. Pattern Recognition,* vol. 3, pp. 157-160, Aug. 2004.

[22] D. Kisku, A. Rattani, M. Tistarelli, and P. Gupta, "Graph Application on Face for Personal Authentication and Recognition," *Proc. 10th Int'l Conf. Control, Automation, Robotics and Vision,* pp. 1150 -1155, Dec. 2008.

[23] B. Glocker, N. Komodakis, N. Paragios, G. Tziritas, and N. Navab, "Inter- and Intra-Modal Deformable Registration: Continuous Deformations Meet Efficient Optimal Linear Programming," *Proc. Information Processing in Medical Imaging,* pp. 408-420, 2007.

[24] Y. Boykov, O. Veksler, and R. Zabih, "Fast Approximate Energy Minimization via Graph Cuts," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, no. 11, pp. 1222-1239, Nov. 2001.

[25] M. Pawan Kumar, P.H. Torr, and A. Zisserman, "Learning Layered Motion Segmentations of Video," *Int'l J. Computer Vision,* vol. 76, no. 3, pp. 301-319, 2008.

[26] V. Kolmogorov, "Convergent Tree-Reweighted Message Passing for Energy Minimization," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 10, pp. 1568-1583, Oct. 2006.

[27] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother, "A Comparative Study of Energy Minimization Methods for Markov Random Fields with Smoothness-Based Priors," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 30, no. 6, pp. 1068-1080, June 2008.

[28] A. Ashraf, S. Lucey, and T. Chen, "Learning Patch Correspondences for Improved Viewpoint Invariant Face Recognition," *Proc. IEEE CS Int'l Conf. Computer Vision and Pattern Recognition,* pp. 1-8, June 2008.

[29] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing,* second ed. Cambridge Univ. Press, Oct. 1992.

[30] A.R. Ahmadyfard and J. Kittler, "Region-Based Object Recognition: Pruning Multiple Representations and Hypotheses," *Proc. British Machine Vision Conf.,* vol. 2, pp. 745-754, 2000.

[31] J. Kittler and A. Lucas, "A New Method for Dynamic Time Alignment of Speech Waveforms in Pattern Recognition and Understanding," *Speech Recognition and Understanding,* P. Laface and R. D. Mori, eds., pp. 537-542, Springer-Verlag, 1991.

[32] K. Messer, J. Matas, J. Kittler, and K. Jonsson, "XM2VTSDB: The Extended M2VTS Database," *Proc. Second Int'l Conf. Audio and Video-Based Biometric Person Authentication,* pp. 72-77, 1999.

[33] X. Tan and B. Triggs, "Enhanced Local Texture Feature Sets for Face Recognition under Difficult Lighting Conditions," *Proc. Analysis and Modelling of Faces and Gestures,* pp. 168-182, 2007.

[34] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 7, pp. 971-987, July 2002.

[35] C.-H. Chan, "Multi-Scale Local Binary Pattern Histogram for Face Recognition," PhD dissertation, 2008.

[36] J. Tena, R. Smith, M. Hamouz, J. Kittler, A. Hilton, and J. Illingworth, "2D Face Pose Normalisation Using a 3D Morphable Model," *Proc. Int'l Conf. Video and Signal Based Surveillance,* pp. 1-6, Sept. 2007.

[37] R. Gross, I. Matthews, and S. Baker, "Appearance-Based Face Recognition and Light-Fields," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, no. 4, pp. 449-465, Apr. 2004.

[38] X. Zhang, Y. Gao, and M. Leung, "Recognizing Rotated Faces from Frontal and Side Views: An Approach toward Effective Use of Mugshot Databases," *IEEE Trans. Information Forensics and Security,* vol. 3, no. 4, pp. 684-697, Dec. 2008.

[39] S. Romdhani, V. Blanz, and T. Vetter, "Face Identification by Fitting a 3d Morphable Model Using Linear Shape and Texture Error Functions," *Proc. Seventh European Conf. Computer Vision-Part IV,* pp. 3-19, 2002.

[40] S. Arashloo, J. Kittler, and W. Christmas, "Facial Feature Localization Using Graph Matching with Higher Order Statistical Shape Priors and Global Optimization," *Fourth IEEE Int'l Conf. Biometrics: Theory Applications and Systems.* pp. 1-16, 2010.