Open Access

Face Verification Using Modeled Eigenspectrum

Bappaditya Mandal, Xudong Jiang* and Alex Kot

School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

Abstract: Face verification is different from face identification task. Some traditional subspace methods that work well in face identification may suffer from severe over-fitting problem when applied for the verification task. Conventional discriminative methods such as linear discriminant analysis (LDA) and its variants are highly sensitive to the training data, which hinders them from achieving high verification accuracy. This work proposes an eigenspectrum model that alleviates the over-fitting problems by replacing the unreliable small and zero eigenvalues with the model values. It also enables the discriminant evaluation in the whole space to extract the low dimensional features effectively. The proposed approach is evaluated and compared with 8 popular subspace based methods for a face verification task. Experimental results on three face databases show that the proposed method consistently outperforms others.

Keywords: Biometrics, face verification, subspace methods, feature extraction, discriminant analysis.

INTRODUCTION

In Biometrics, face recognition has two main applications, one is verification and the other is identification. Face verification is a task to determine whether a person claiming a given identity is the true claimant or an imposter. This can be done by computing the similarity between the probe sample and the samples of the claimed person in the gallery. The final decision is made based on a threshold. For face identification, however, the system has to determine the identity of a person by computing similarities between the probe and all the gallery samples in the database. The identity is determined based on the highest similarity score. Over the last decade numerous algorithms using linear as well as nonlinear techniques for face identification have been proposed and considerable good performance has been achieved [1-3]. However, these algorithms focus on face identification and very few of them are evaluated for face verification task. Although there are some attempts to distinguish face verification from face identification [4, 5] many works ignore the intrinsic difference between face verification and face identification tasks.

In general, similar training algorithms can be used for both face verification and identification. However, there are some intrinsic differences between the two cases. While an identification system could be limited to identify an input as one of its known users, a verification system must be able to reject the unknown imposters.

Consequently, while the training set could include some sample images of all test subjects for certain identification applications, no sample of test imposters should be included in the training set for a verification system. Hence, a learning algorithm may suffer more severe over-fitting problem (or poor generalization) for face verification task than for the face identification. Furthermore, decisions of a verification system depend on the operating points or thresholds while those of an identification system depend on the rank of the similarity. This leads to quite different evaluation methods for the verification system from the identification system. Thus, a method that has a high accuracy for the identification task may not necessarily achieve a high accuracy for the verification task.

Both face verification and identification are challenging as the human faces may undergo significant variations in appearance due to different facial expressions, illumination changes and different pose conditions. Subspace based methods such as the principal component analysis (PCA) [6], Bayesian maximum likelihood (BML) [7-9] and linear discriminant analysis (LDA) [10, 11] have shown promising results for the face identification problem. This work explores some outstanding challenging problems of existing subspace based methods caused by the high dimensionality of the face image and the finite number of training samples in practice and proposes a new approach to alleviate these problems for the face verification task.

PCA maximizes the variances of the extracted features and hence minimizes the reconstruction error and removes noise residing in the discarded dimensions. The best representation of data may not perform well from classification point of view because the total scatter matrix is contributed by both the within- and between-class variations. To differentiate face images of one person from those of the others, the discrimination of the features is the most important. LDA is an efficient way to extract the discriminative features as it handles the within- and between-class variations separately. However, this method needs the inverse of the within-class scatter matrix. This is problematic in many practical face recognition tasks because the dimensionality of the face image is usually very high compared to the number of available training samples and hence the within-class scatter matrix is often singular.

*Address correspondence to this author at the School of Electrical and Electronic Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798; Tel: +65 6790 5018; Fax: +65 6793 3318;
E-mail: exdjiang@ntu.edu.sg

1874-0618/08 2008 Bentham Open

Numerous methods have been proposed to solve this problem in the last decade. A popular approach called Fisherface (FLDA) [12] applies PCA first for dimensionality reduction so as to make the within-class scatter matrix nonsingular before the application of LDA. However, applying PCA for dimensionality reduction may lose discriminative information [13-15]. Direct-LDA (DLDA) method [16, 17] removes null space of the between-class scatter matrix and extracts the eigenvectors corresponding to the smallest eigenvalues of the within-class scatter matrix. It is an open question of how to scale the extracted features as the smallest eigenvalues are very sensitive to noise. The null space approach (NDA) [15, 18] assumes that the null space contains the most discriminative information. Interestingly, this appears to be contradicting the popular FLDA that only uses the principal space and discards the null space. A common problem of all these approaches is that they all lose some discriminative information, either in the principal or in the null space.

In fact, the discriminative information resides in the both subspaces. To use both the subspaces, dual-space approach (DSL) [14] extracts features separately from the principal and its complementary subspaces of the within-class scatter matrix. It scales features in the complementary subspace by the average eigenvalue of the within-class scatter matrix over this subspace. As eigenvalues in this subspace are not well estimated [14], their average may not be a good scaling factor relative to those in the principal subspace. Features extracted from the two complementary subspaces are properly fused by using summed normalized-distance [19]. Open questions of these two approaches are how to divide the space into the principal and the complementary subspaces and how to apportion a given number of features to the two subspaces. Furthermore, as the discriminative information resides in the both subspaces, it is inefficient and only suboptimal to extract features separately from the two subspaces.

The above approaches focus on the problem of singularity of the within-class scatter matrix. In fact, the instability and noise disturbance of the small eigenvalues cause great problems when the inverse of the matrix is applied such as in the Mahalanobis distance, in the BML estimation and in the whitening process of various LDA approaches. Problems of the noise disturbance were addressed in [20] and a unified framework of subspace methods (UFS) was proposed. The good recognition performance of this framework shown in [20] verifies the importance of the noise suppression. However, this approach applies three stages of subspace decompositions sequentially on the face training data and the dimensionality reduction occurs at the very first stage. As addressed in the literature [13-15], applying PCA for dimensionality reduction may lose discriminative information. Another open question of UFS is how to choose the number of principal dimensions for the first two stages of subspace decompositions before selecting the final number of features in the third stage. The experimental results in [20] show that the recognition performance is sensitive to these choices at different stages.

In this work, we propose a three subspace based eigenspectrum decomposition methodology which uses two control points to differentiate the reliable, unreliable and zero eigenvalues. An eigenspectrum modeling procedure is proposed that enables us to perform discriminant evaluation in the whole space. This eigenspace decomposition is not used to limit the discriminant evaluation in one subspace but to enable the evaluation in the whole space. The extracted feature hence contains both the reconstructive and the discriminative information of training samples. The replacement of unreliable small and zero eigenvalues by the modeled values reduces the sensitivity of discriminative methods to the number of training samples, high dimensionality of the face images and noises in the data. It provides better generalization or less over-fitting as compared to the existing methods for face verification tasks.

In addition, many subspace methods are evaluated only for face identification task. Their performances for face verification are unknown in the literature. This work experimentally evaluates eight popular subspace based approaches for the face verification task and compares them with the proposed method. Experimental results on three face databases demonstrate that the proposed method consistently outperforms others. In the following section, we present the problems of feature scaling, the subspace decomposition and eigenspectrum modeling. Following the above section, we discuss the eigenfeature scaling and extraction in the whole space using the proposed eigenmodel. Experimental results and discussions that compare our method with others are presented before drawing conclusions.

FEATURE SCALING AND SUBSPACE DECOMPOSITION

Problems in Feature Scaling and Extraction

Given a set of properly normalized h-by-w face images, we can form a training set of column vectors $\{X_{ij}\}$, where $X_{ij} \in \mathrm{R}^{n=hw}$ is called image vector, by ordering the pixel elements of image j of person i. Let the training set contain p persons and q_i sample images for person i. The number of total training samples is $l = \sum_{i=1}^{p} q_i$. Letting c_i be the prior probability of person i, the within-class scatter matrix is defined by

$$\mathbf{S}^w = \sum_{i=1}^{p} \frac{c_i}{q_i} \sum_{j=1}^{q_i} (X_{ij} - \overline{X}_i)(X_{ij} - \overline{X}_i)^T, \quad (1)$$

where $\overline{X}_i = \frac{1}{q_i} \sum_{j=1}^{q_i} X_{ij}$. The between-class scatter matrix $\mathbf{S}^b$ and the total (mixture) scatter matrix $\mathbf{S}^t$ are defined by

$$\mathbf{S}^b = \sum_{i=1}^{p} c_i (\overline{X}_i - \overline{X})(\overline{X}_i - \overline{X})^T, \quad (2)$$

$$\mathbf{S}^t = \sum_{i=1}^{p} \frac{c_i}{q_i} \sum_{j=1}^{q_i} (X_{ij} - \overline{X})(X_{ij} - \overline{X})^T, \qquad (3)$$

where $\overline{X} = \sum_{i=1}^{p} c_i \overline{X}_i$. If all persons have equal prior probability, then $c_i = 1/p$. Let $S^g, g \in \{t, w, b\}$ represent one of the above scatter matrices. If we regard the elements of the image vector or the class mean vector as features, these preliminary features will be de-correlated by solving the eigenvalue problem

$$\Lambda^g = \Phi^{gT} \mathbf{S}^g \Phi^g, \qquad (4)$$

where $\Phi^g = [\phi_1^g, ..., \phi_n^g]$ is the eigenvector matrix of $\mathbf{S}^g$, and Λ^g is the diagonal matrix of eigenvalues $\lambda_1^g, ..., \lambda_n^g$ corresponding to the eigenvectors. Suppose that the eigenvalues are sorted in descending order $\lambda_1^g \geq, ..., \geq \lambda_n^g$. The plot of eigenvalues λ_k^g against the index k is called eigenspectrum of the face training data. It plays a critical role in the subspace methods as the eigenvalues are used to scale and extract features.

If we compute all the eigenvalues $\Lambda^w = diag(\lambda_1^w, ..., \lambda_n^w)$ and eigenvectors $\Phi^w = [\phi_1^w, ..., \phi_n^w]$ of the n-by-n dimensional matrix $\mathbf{S}^w$ using (4). The projection matrix $\Phi^w = [\phi_1^w / \sigma_1^w, ..., \phi_n^w / \sigma_n^w]$ is so called whitened eigenvector matrix of $\mathbf{S}^w$ with $\| \phi_k^w \| = 1$ and $\sigma_k^w = \sqrt{\lambda_k^w}$. This implies that if any one of the eigenvalues in (4) of these matrices is zero or close to zero then the corresponding eigenvector ϕ_i^w gets an infinite or semi-infinite weighting factor. If we discard the eigenvectors corresponding to the zero eigenvalues (that appear beyond the rank of $\mathbf{S}^w$), these eigenvectors are equivalently multiplied by zero. This can be viewed as an n-dimensional pattern vector X_{ij} is first represented by an n-dimensional eigenfeature vector $Y_{ij} = \Phi^{w^T} X_{ij}$, and then multiplied by a weighting function

$$w_k^w = \begin{cases} 1/\sqrt{\lambda_k^w}, & k \leq r_w \\ 0, & r_w < k \leq n \end{cases}, \qquad (5)$$

as shown in Fig. (**1**), where r_w is the rank of $\mathbf{S}^w$.

Conventional FLDA applies PCA first for dimensional reduction (DR) and then LDA is used for discriminant analysis [12]. Several questions arise pertinent to the amount of basis vectors or principal components to be retained in this DR and how they affect the performance from discrimination point of view. Detail discussions and experimental results can be found in [21]. To facilitate analysis one has to consider three important factors: limited number of training samples, dimensionality of the data and the presence of noise in the data. For image representation PCA is optimal and gives the most compact representation. Its reconstruction performance improves when more number of principal components are used. However, it is well known that keeping more components may lead to decreased classification performance in FLDA [11] and also for BML [22]. In the PCA step of FLDA, if more components are kept corresponding to the small eigenvalues (which encode noises), LDA process has to cater for noises as well, consequently over-fitting problem occurs and classification performance decreases [23]. For face verification applications, this problem can be more severe as compared to face identification because in the former, the probe imposter subjects are distinct and unknown during the training session. Therefore, the algorithm used for verification purpose should be designed to have more generalization capability (or less over-fitting problem) as compared to face identification algorithms.

Two major problems arise which are visible in the graph (Fig. **1**). Firstly, although the eigenvalues in the region $(\lambda_k^w \mid m_1 < k \leq r_w)$, where m_1 is a control point, are within the range space (rank) of $\mathbf{S}^w$ matrix, they are very small and noise component may dominate the eigenvalues. Furthermore, the finite number of training samples results in faster decay of the eigenvalues. When their inverses are used for scaling (5), the corresponding eigenvectors get undue heavy or semi-infinite weightage in this range, as shown in Fig. (**1**). These small eigenvalues cause misleading scaling in the whitening step and thus generalize poorly when exposed to entirely new (subject) data - a scenario which is commonly encountered in face verification and other pattern recognition applications.

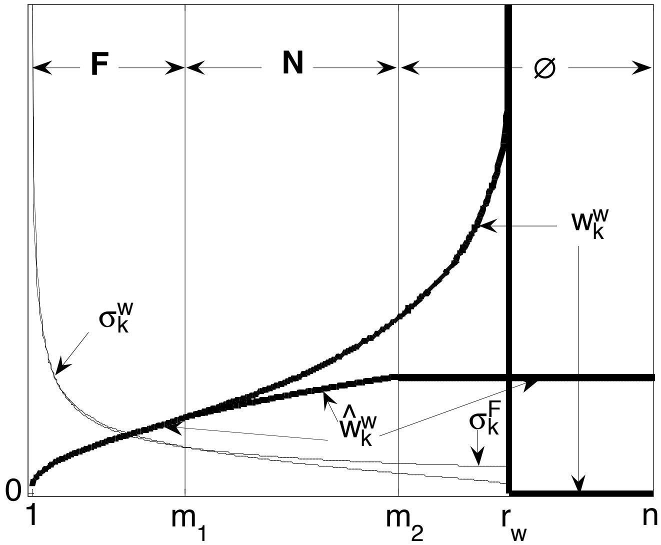

Fig. (1). Weighting functions of (5) and (14) in the face-, noise- and null-subspaces based on a typical real eigenspectrum.

Secondly, it is apparent from Fig. (**1**) that the eigenvectors $\{\phi_k^w\}_{k=r_w+1}^{n}$ in the null space of $\mathbf{S}^w$ are weighted by zero and thus this subspace is excepted from the discriminant evaluation. This is unreasonable because features in the null space have zero within-class variances based on the training data and hence should be more heavily weighted. It seems anomalous that the weighting function increases with the decrease of the eigenvalues and then suddenly has a big drop from the maximum value to zero as shown in Fig. (**1**). Furthermore, weights determined by the inverse of σ^w_k is, though optimal in terms of the ML estimation, dangerous when σ^w_k is small ($m_2 < k \le r_w$), where m_2 is a control point. The small and zero eigenvalues are training-set-specific and very sensitive to different training sets [24]. Adding new samples to the training set or using different training set may easily change some zero eigenvalues to nonzero and make some very small eigenvalues several times larger. Therefore, these eigenvalues of the within-class scatter matrix are unreliable and need to be replaced by suitably modeled eigenvalues. Before proposing the eigenspectrum model, we first need to decompose the range space into reliable, unstable noise variation dominating subspace and null spaces.

SUBSPACE DECOMPOSITION

We propose a methedology to estimate two control points m_1 and m_2 which will segregate the reliable eigenvalues from the unreliable and near zero ones. As the inverse of the eigenvalues are used in the feature scaling (5), the noise disturbances and the limited training samples have little effect on the initial portion of the eigenspectrum (Fig. **1**) but may substantially affect the feature stability in the latter portion of the range space where the eigenvalues are small or close to zero. Hence, the whole eigenspace R^n spanned by eigenvectors $\{\phi^w_k\}^n_{k=1}$ is decomposed into three subspaces: a reliable face variation dominating subspace (or simply face space) $\mathbf{F} = \{\phi^w_k\}^{m_1}_{k=1}$, an unreliable noise variation dominating subspace (or simply noise space) $\mathbf{N} = \{\phi^w_k\}^{m_2}_{k=m_1}$ and a null space $\emptyset = \{\phi^w_k\}^n_{k=m_2}$ as illustrated in Fig. (**1**).

Estimation of m_1

The rank of $\mathbf{S}^w$ is $r_w \le min(n, l-p)$. In practice, the rank of a scatter matrix usually reaches this maximum values unless some training images are linearly dependent. Even in this rare case, the rank r_w can be easily determined by finding the maximal value of k that satisfies $\lambda^w_k > \varepsilon$, where ε is a very small positive value comparing to λ^w_1. As face images have similar structure, significant face components reside intrinsically in a very low-dimensional (m_1-dimensional) subspace. For a robust training, the database size should be significantly larger than the face dimensionality m_1, although it could be, and usually in practice is, much smaller than the image dimensionality n. Thus, in many practical face verification training tasks we usually have $m_1 = r_w = n$. As the face component typically decays rapidly and stabilizes, eigenvalues in the face dominant subspace, which constitute the initial portion of the eigenspectrum, are the outliers of the whole spectrum. It is well known that median operation works well in separating outliers from a data set. To determine the start point of the noise dominant region $m_1 + 1$, we first find a point near the center of the noise region by

$$\lambda^w_{med} = \text{median}\left\{\forall\lambda^w_k \mid k \le r_w\right\}. \tag{6}$$

The distance between λ^w_{med} and the smallest nonzero eigenvalue is $d_{m_1,r_w} = \lambda^w_{med} - \lambda^w_{r_w}$. As eigenvalues tend toward equality in **N** and are highly disparate in **F**, the distance d_{m_1,r_w} that indicates the half range of the eigenvalue variation in **N** is very small comparing to the highly disparate eigenvalues in **F**. Therefore, $\lambda^w_{med} + d_{m_1,r_w}$ is proposed as the upper bound of the unreliable eigenvalues to separate the highly disparate subspace **F** from the flat subspace **N**. Although this is a reasonable choice of the upper bound of the unreliable eigenvalues, it may not be optimal in all cases considering the great variation of image size and number of training samples in different applications. More generally, the start point of the noise region $m_1 + 1$ is estimated by

$$\lambda^w_{m_1+1} = \max\left\{\forall\lambda^w_k \mid \lambda^w_k < (\lambda^w_{med} + \mu(\lambda^w_{med} - \lambda^w_{r_w}))\right\}, \tag{7}$$

where μ is a constant. The optimal value of μ may be slightly larger or smaller than 1 for different applications. To avoid exhaustive search for the best parameter value, μ is fixed to be 1 in all experiments of this paper for fair comparisons with other approaches.

Estimation of m_2

To differentiate the unreliable eigenvalues from the larger ones we employ the ratios of the successive eigenvalues of the eigenspectrum to decompose the whole eigenspace. The phenomenon that the eigenspectrum accelerates its decrease is caused by the limited number of training samples and noises present in them. To study this, we define eigenratios as

$$\gamma^w_k = \frac{\lambda^w_k}{\lambda^w_{k+1}}, \quad 1 \le k < r_w. \tag{8}$$

The plot of eigenratios γ^w_k against index k is called eigenratio-spectrum. Fig. (**2**) shows a typical eigenratio-spectrum of a real face training database. The limited number of the training samples causes the increase of the eigenratios. The corresponding eigenvalues are thus unreliable.

We examined several different face databases, the eigenratio plots shown in Fig. (**2**) is a general behavioral pattern that all the eigenratios of different databases portray. It is apparent from the graph that the eigenratios first decreases very rapidly, then stabilizes and finally increases. The increase of the eigenratios should not be the behavior of the true variances. Therefore, the start point of the unreliable region $m_2 + 1$ is estimated by

$$\gamma^w_{m_2+1} = \min\left\{\forall\lambda^w_k, 1 \le k < r_w\right\}. \tag{9}$$

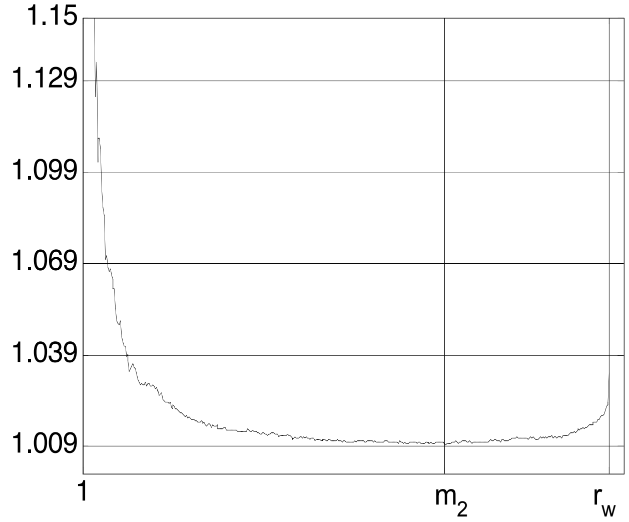

Fig. (2). Eigenratio-spectrum (8) from a real eigenspectrum.

A typical such m_2 value of a real eigenspectrum is shown in Fig. (**2**).

EIGENSPECTRUM MODELING

If we regard X_{ij} as samples of a random variable vector X, the eigenvalue λ_k^w is a variance estimate of X projected on the eigenvector ϕ_k^w estimated on the training samples. It usually deviates from the true variance of the projected random vector X due to noise and the finite number of training samples. Thus, unreliable eigenvalues need to be replaced by some model value to alleviate the over-fitting problem. As the eigenspectrum typically decays rapidly and stabilizes, we can model it by a function of the form $1/f$ that can well fit to the decaying nature of the eigenspectrum. The function form $1/f$ was used in BML approach [7] to fit the eigenspectrum in the whole range subspace $\{\lambda_k^w \mid 1 \le k \le r_w\}$ and then to extrapolate eigenvalues in the null space $\{\lambda_k^w \mid r_w < k \le n\}$ for computing an average eigenvalue. Different from BML approach [7], this work uses function form $1/f$ to fit only the reliable part of eigenspectrum $\{\lambda_k^w \mid 1 \le k \le m_1\}$ and then to extrapolate eigenvalues in the noise subspace $\{\lambda_k^w \mid m_1 < k \le m_2\}$.

We propose to model the eigenspectrum by

$$\hat{\lambda}_k^w = \frac{\alpha}{k+\beta}, \quad 1 \le k \le r_w, \tag{10}$$

where α and β are two constants. As the eigenspectrum in the face space is dominated by the face structural component, the parameters of α and β is determined by fitting the model to the real eigenspectrum in the reliable face space **F**. While not limiting ourselves from other possible fitting methods, in all experiments of this work, we simply determine α and β by letting $\hat{\lambda}_1^w = \lambda_1^w$ and $\hat{\lambda}_{m_1}^w = \lambda_{m_1}^w$, which yields

$$\alpha = \frac{\lambda_1^w \lambda_{m_1}^w (m_1 - 1)}{\lambda_1^w - \lambda_{m_1}^w}, \tag{11}$$

$$\beta = \frac{m_1 \lambda_{m_1}^w - \lambda_1^w}{\lambda_1^w - \lambda_{m_1}^w}. \tag{12}$$

Since the eigenspectrum decays very fast, we plot the square roots $\sigma_k^w = \sqrt{\lambda_k^w}$ and $\hat{\sigma}_k^w = \sqrt{\hat{\lambda}_k^w}$ for clearer illustration (we still call them eigenspectrum for simplicity). A typical real eigenspectrum σ_k^w and its model $\hat{\sigma}_k^w$ are shown in Fig. (**1**). We see that the model $\hat{\sigma}_k^w$ fits closely to the real σ_k^w in the face space **F** but has slower decay in the noise space **N**. The faster decay of the real eigenspectrum σ_k^w in **N** due to noise and the limited number of training samples is what we want to slow down (Fig. **1**).

EIGENFEATURE SCALING AND EXTRACTION

According to LDA criteria, the optimal discriminative features should have smallest within-class variations and largest between-class variations. As pointed out in [25], LDA can be carried out in two operations: first eigen-decomposition of the with-class scatter matrix, followed by the eigenvector scaling or whitening and the second eigen-decomposition of the transformed between-class variation matrix. The biggest hurdle comes in the whitening process where the inverses of the eigenvalues of the within-class scatter matrix are used to scale the eigenvectors or features. Over-fitting problems occur here due to the high dimensionality of face image and the limited number of training samples. As elaborated before the problems of feature scaling and extraction, the existing scaling function (5) gives undue heavy weightage to the corresponding eigenvectors in the range $\{k \mid m_1 < k \le r_w\}$, resulting in poor generalization to the new data. Moreover, eigenvectors corresponding to the zero eigenvalues are lost and they fail to contribute to the discriminant evaluation. Hence, the small and zero eigenvalues need to be replaced by the modeled eigenvalues. This replacement of eigenvalues alleviates over-fitting problems and also enables us to perform discriminant evaluations in the whole space.

Eigenfeature Scaling

Although there is always noise component in **F** as noise affects every element of the image vector, its variance is very small comparing to the large variance of the face structural component in **F**. In **N**, however, noise component may dominate in the variance changes and the finite number of training samples results in faster decay of the variances. Therefore, the decay of the eigenvalues should be slowed down to compensate the effect of noise and the finite number of training samples. This can be done by replacing the eigenspectrum with the proposed model (10). In the null space, we have no information about the variation of the eigenvalues and hence all features are treated in a same way. The zeros variance in the null space is only an estimate on one set of the training data. Another set of training data may easily make them nonzero, especially when larger number of training samples are used. Therefore, we should not trust the zero variance and derive an infinite or very large feature weights in this space. However, based on the available training data that result in zero variances in the null space, the feature weights in the null space should not be smaller than those in the other subspaces.

Therefore, we keep the eigenvalues in the face structure subspace **F** unchanged, replace the unreliable small eigenvalues in the noise dominating subspace **N** by the model $\hat{\lambda}_k^w = \frac{\alpha}{k+\beta}$, and replace the zero eigenvalues in null space Ø by the constant $\hat{\lambda}^w_{r_w+1}$. Thus, the modeled eigenspectrum $\widetilde{\lambda}_k^w$ is given by

$$\widetilde{\lambda}_k^w = \begin{cases} \lambda_k^w, & k < m_1 \\ \dfrac{\alpha}{k+\beta}, & m_1 \le k \le m_2. \\ \dfrac{\alpha}{r_w + 1 + \beta}, & m_2 < k \le n \end{cases} \tag{13}$$

The proposed feature weighting function is then

$$\widetilde{w}_k^w = \frac{1}{\sqrt{\widetilde{\lambda}_k^w}}, \; k = 1,2,...n. \tag{14}$$

Fig. (**1**) shows the proposed feature weighting function $\widetilde{w}_k^w$ calculated by (11), (12), (13) and (14) comparing with that w_k^w of (5). Obviously, the new weighting function $\widetilde{w}_k^w$ is identical to w_k^w in the face structural dominating subspace **F**, increases along with k at a much slower pace than w_k^w in the noise dominating subspace **N** and has maximal constant weights instead of zero of w_k^w in the null space Ø.

Using this weighting function and the eigenvectors ϕ_k^w, training data are transformed to

$$\widetilde{Y}_{ij} = \Phi_n^{wT} X_{ij}, \tag{15}$$

where

$$\Phi_n^w = [\tilde{w}_k^w \phi_k^w]_{k=1}^n = [\tilde{w}_1^w \phi_1^w, ..., \tilde{w}_n^w \phi_n^w] \tag{16}$$

is a full rank matrix that transforms an image vector to an intermediate feature vector. There is no dimension reduction in this transformation as $\widetilde{Y}_{ij}$ and X_{ij} have the same dimensionality n.

Eigenfeature Extraction

After the above feature transformation and scaling, a new between-class scatter matrix is formed by vectors $\widetilde{Y}_{ij}$ of the transformed training data as

$$\widetilde{\mathbf{S}}^b = \sum_{i=1}^{p} c_i(\overline{\widetilde{Y}}_i - \overline{Y})(\overline{\widetilde{Y}}_i - \overline{Y})^T, \tag{17}$$

where $\overline{\widetilde{Y}}_i = \frac{1}{q_i}\sum_{j=1}^{q_i}\widetilde{Y}_{ij}$ and $\overline{Y} = \sum_{i=1}^{p}\frac{c_i}{q_i}\sum_{j=1}^{q_i}\widetilde{Y}_{ij}$.

The transformed features $\widetilde{Y}_{ij}$ will be de-correlated for $\widetilde{\mathbf{S}}^b$ by solving the eigenvalue problem as (4). Suppose that the eigenvectors in the eigenvector matrix $\Phi_n^b = [\tilde{\phi}_1^b, ..., \tilde{\phi}_n^b]$ are sorted in descending order of the corresponding eigenvalues. The dimensionality reduction or feature extraction is performed here by keeping the eigenvectors with the d largest eigenvalues

$$\Phi_d^b = [\tilde{\phi}_k^b]_{k=1}^d = [\tilde{\phi}_1^b, ..., \tilde{\phi}_d^b], \tag{18}$$

where d is the number of features usually selected by a specific application. Thus, the proposed feature scaling and extraction matrix $\mathbf{U}$ is given by

$$\mathbf{U} = \Phi_n^w \Phi_d^b, \tag{19}$$

which transforms a face image vector X, $X \in \mathsf{R}^n$, into a feature vector F, $F \in \mathsf{R}^d$, by

$$F = \mathbf{U}^T X. \tag{20}$$

We witness that the three-subspace based eigenspectrum decomposition proposed in this work is only used to replace the unreliable small and zero eigenvalues by the model values. The discriminant evaluation (here the evaluation of the eigenvalues of $\widetilde{\mathbf{S}}^b$) is performed in the full space R^n. This approach extracts the discriminative features from the whole space by searching the most discriminative features in the full space. Thus, the proposed method is based on the global optimization, different from the local optimization in a subspace of approaches FLDA [12], DLDA [16, 17], NDA [15, 18] and UFS [20] and different from the two separate local optimizations in two subspaces of the dual-space approaches in [14, 19].

THE PROPOSED ALGORITHM

The proposed face verification using modeled eigenspectrum (MES) approach is summarized below:

At the training stage:

1. Given a training set of normalized face image vectors $\{X_{ij}\}$, compute $\mathbf{S}^w$ by (1) and solve the eigenvalue problem as (4).
2. Estimate m_1 value using (6) and (7).
3. Estimate m_2 value using (8) and (9).
4. Decompose the eigenspace into face-, noise-, and null-space uisng m_1 and m_2 values.
5. Transform the training samples represented by X_{ij} into $\widetilde{Y}_{ij}$ by (15) with the weighting function (14) determined by (8), (9), (11), (12) and (13).

6. Compute $\widetilde{\mathbf{S}}^b$ by (17) with $\widetilde{Y}_{ij}$ and solve the eigenvalue problem as (4).
7. Obtain the final feature scaling and extraction matrix by (16), (18) and (19) with a predefined number of features d.

At the enrollment or registration stage:

1. Extract d-D feature vector F from the enrolled n-D normalized face image vector X by (20) using the feature scaling and extraction matrix $\mathbf{U}$ obtained in the training stage (19).
2. Store the extracted feature vector and the registration ID into the gallery feature vector set.

At the verification stage:

1. Extract d-D feature vector F from the n-D normalized probe face image vector X by (20) using the feature scaling and extraction matrix $\mathbf{U}$ obtained in the training stage (19).
2. Compare or match the probe feature vector with that in the gallery feature vector set corresponding to the claimed ID.

In the experiments of this work, cosine distance measure between a probe feature vector F_P and a gallery feature vector F_G

$$dst(F_P, F_G) = \frac{{F_P}^T F_G}{\|F_P\|_2 \|F_G\|_2} \tag{21}$$

is applied in matching two feature vectors, where $\|\bullet\|_2$ is the norm 2 operator.

ANALYSIS AND COMPARISONS

Subspace methods like FLDA, DLDA, UFS and NDA approaches discard a subspace before the discriminant evaluation. Therefore, they perform a local optimization process and hence their extracted features are suboptimal or are the most discriminative only in a subspace. Although BML works in the whole space, it does not evaluate the discriminant value and hence the whole face image must be used as features in the verification process. It performs well for the identification problem but as we will see in the experiments, it suffers from severe over-fitting problem for the verification task. Although the dual-space based approaches do not throw away any subspace before the discriminant evaluation, it is inefficient or only suboptimal to evaluate the discriminant value and extract features separately in two subspaces and then to combine them. Other open questions of the dual-space based approaches include how to divide the space into the two subspaces and how to apportion the given number of features to the two subspaces.

The proposed algorithm (MES) in this work has two important and novel ingredients. First, the three subspace based eigenspectrum decomposition differentiates the reliable, unreliable and zero eigenvalues. Two reasonable decomposition points are determined by the algorithm. This eigenspace decomposition is not used to limit the discriminant evaluation in one subspace but to enable the evaluation in the whole space. Thus, the proposed method is based on the global optimization that extracts features by searching the most discriminative ones in the whole space. Second, the parameters of the eigenspectrum model are determined by the reliable portion of the real eigenspectrum and the modeled eigenvalues are then used to replace the unreliable small and zero real eigenvalues. This reduces the sensitivity of the extracted features to the dimensionality of face image, the number of training samples and noise disturbance. Consequently, the proposed approach alleviates the over-fitting problem that often occurs in the training process with limited number of high dimensional samples. As discussed before, the face verification task may suffer from more severe over-fitting problem (or poorer generalization) than the face identification task. The proposed MES approach facilitates a discriminative and stable low-dimensional feature representation of the face image, which is verified in the following experiments on the face verification task.

EXPERIMENTS

Three popular face databases: AR database, FERET database 1 and FERET database 2 are used in the experiments. Each database is partitioned into training, gallery and probe sets. In all experiments reported in this work, images are preprocessed, aligned and normalized following the CSU Face Identification Evaluation System [26]. Face verification is performed by accepting a claimant if the subject's matching score is greater than or equal to a threshold and rejecting the claimant if its matching score is lower than the threshold. Verification performance is evaluated using two measures: false acceptance rate (FAR) and the false rejection rate (FRR). FAR is the ratio of the number of accepted imposter matchings to the total number of imposter matchings. FRR is the ratio of the number of rejected genuine matchings to the total number of genuine matchings. The plot of FRR against FAR is called the receiver operating characteristics (ROC) curve. The system performances at various different operating points (thresholds) are characterized by the ROC curve. The equal error rate EER, defined by EER=FAR=FRR at a specific threshold, serves as a single number indicator of a verification system's performance.

The proposed MES method is tested and compared with 8 other popular subspace based approaches: PCA [6] with Euclidian distance (PCAE), PCA with Mahalanobis distance (PCAM), FLDA [12], DLDA [16], BML [8], NDA [18], UFS [20] and DSL [14] approaches. The proposed MES approach has only one free parameter μ in (7). To avoid exhaustive search for the best parameter value, μ is fixed to be 1 in all experiments of this paper for fair comparisons with other approaches. The parameters of UFS are applied that result in the best performance through an exhaustive search in the experiments of [20]. We conduct the experiments starting with the number of features $d = 10$, incremented by 2 each time up to $p-1$, where p is the number of training subjects. Experimental results are presented in this paper for each approach where the minimum EER is obtained.

Results on AR Database

The color images in AR database [27] are converted to gray-scale and cropped into the size of 120×170, same as the image size used in [27, 28]. The pictures of most subjects were taken in two sessions (separated by two weeks). In our experiment, 75 subjects with 14 non-occluded images per subject were selected from the AR database. The first 7 images (numbers 1-7, first session [27]) of 60 subjects are used in the training and also serve as gallery images. The second 7 images (numbers 14-20, second session [27]) of the 60 subjects serve as probe genuine images. The remaining 15 subjects with 14 images per subject are used as probe imposters. The total number of genuine matches is $7 \times 7 \times 60 = 2,940$ and the total number of imposter matches is $14 \times 7 \times 15 \times 60 = 88,200$. For this large image size, we first apply PCA to remove the null space of $\mathbf{S}^t$ and then apply the MES approach on the 419-dimensional feature vectors. Fig. (**3**) shows the ROC curve that plots the false rejection rate (%) against the false acceptance rate (%).

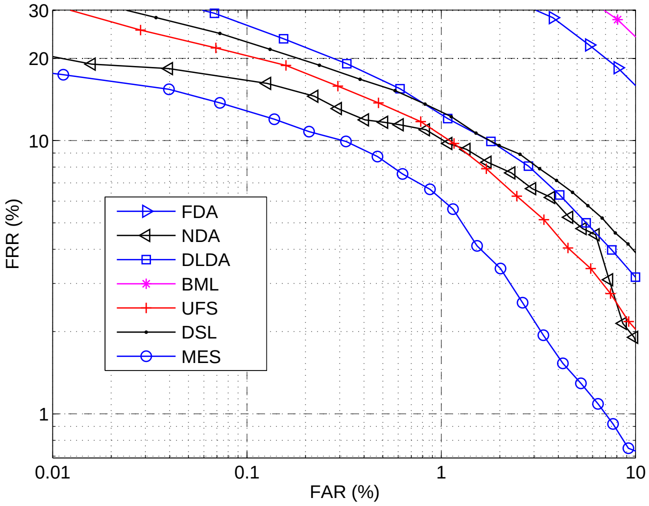

Fig. (3). False rejection rate against the false acceptance rate on the AR face database of 420 training and gallery images (60 subjects), 420 probe genuine images (60 subjects) and 210 probe imposter images (15 subjects). The total number of genuine matches is $7 \times 7 \times 60 = 2,940$ and the total number of imposter matches is $14 \times 7 \times 15 \times 60 = 88,200$.

The ROC curves of PCAE and PCAM do not appear in Fig. (**3**) because their FRRs and FARs are so high that their values are out of the range of Fig. (**3**). Their EERs are numerically recorded in Table **1**. We see that BML approach does not perform well for the face verification task although it is one of the best approaches for the face identification task. The problem of BML for face verification was also addressed in [29]. Fig. (**3**) shows that the proposed MES method consistently outperforms all other 8 approaches for all different operating points (thresholds).

Results on FERET Database 1

In FERET database, the face image variations include facial expression and other details (like glasses or no glasses), illumination, pose, and aging [30]. 2388 images comprising of 1194 subjects (two images FA/FB per subject) are selected from the FERET database. Images are cropped into the size of 33×38. Images of 497 subjects are randomly selected for training and the remaining images of 697 subjects are used for testing. For this database, the subjects used for training are different from those used for the testing. There is no overlap in subjects between the training and the testing data sets. The gallery data set contains 697 subjects with 1 image per subject. The remaining 697 images of the same subjects as in the gallery serve as both the probe genuine images (when matched with gallery images of same subjects) and the probe imposters (when matched with gallery images of different subjects). The total number of genuine matches is $697 \times 1 = 697$ and the total number of imposter matches is $697 \times 696 = 485,112$.

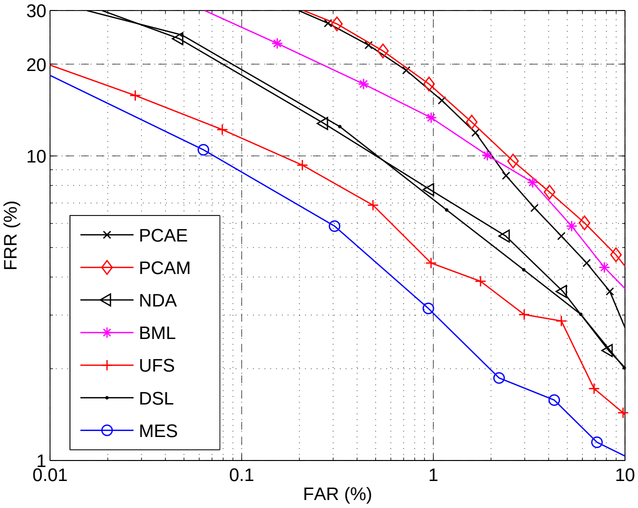

Fig. (4). The ROC curve that plots the false rejection rate (%) against the false acceptance rate (%). False rejection rate against the false acceptance rate on the FERET database 1 of 994 training images (497 subjects), 697 gallery images (697 subjects) and 697 probe images (697 subjects). The total number of genuine matches is $697 \times 1 = 697$ and the total number of imposter matches is $697 \times 696 = 485,112$.

The ROC curves of FLDA and DLDA do not appear in Fig. (**4**) because their FRRs and FARs are too high to be included in Fig. (**4**). Their EERs are numerically recorded in Table **1**. This experiment shows that FLDA and DLDA suffer from severe over-fitting problem. Although BML performs better than in the first experiment, it still underperforms the traditional PCAE for some operating points. Fig. (**4**) shows again that the proposed MES method consistently achieves the lowest FAR and FRR among the 9 tested approaches for all different operating points.

Results on FERET Database 2

This database is constructed, similar to one data set used in [31], by choosing 256 subjects randomly with at least four images per subject form FERET database. However, we use the same number of images (four) per subject for all subjects. Three images per subject of the first 200 subjects are used for training and also serve as gallery images. The remaining 200 images of the 200 subjects are used as probe genuine images. All 4 images of the remaining 56 subjects serve as probe imposter images. The size of the normalized image is 130×150, same as that in [31]. For such a large image size, we first apply PCA to remove the null space of $\mathbf{S}^t$ and then apply the proposed MES approach on the 599-dimensional feature vectors.

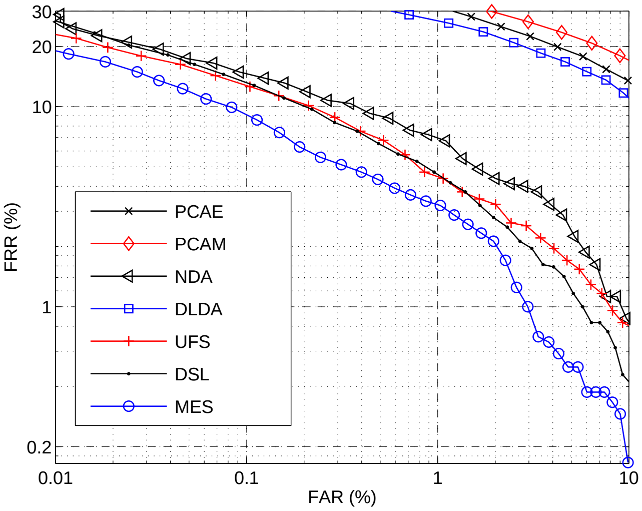

Fig. (5). False rejection rate against the false acceptance rate on the FERET database 2 of 600 training and gallery images (200 subjects), 200 probe genuine images (200 subjects) and 224 probe imposter images (56 subjects). The total number of genuine matches is $200 \times 3 = 600$ and the total number of imposter matches is $4 \times 3 \times 56 \times 200 = 134,400$.

For this database we conducted 4 runs of training and testing with distinct probe genuine image set in each run. More specifically, the i^{th} images ($i = 1,2,3,4$) of all training subjects are chosen to form probe genuine set and the remaining 3 images per subject serve as the training and gallery images. The total number of genuine matches is $200 \times 3 = 600$ and the total number of imposter matches is $4 \times 3 \times 56 \times 200 = 134,400$. Fig. (**5**) shows the ROC curves of the 4 runs of training and testing.

The ROC curves of FLDA and BML do not appear in Fig. (**5**) because their FRRs and FARs are so high that their values are out of the range of Fig. (**5**). Their EERs are numerically recorded in Table **1**. Although the second lowest ROC curve is different from the previous two experiments, Fig. (**5**) shows once more that the proposed MES method consistently delivers the most accurate face verification for all different operating points.

For an accurate record, various ERRs (in %) obtained from the above three experiments are numerically recorded in Table **1**. It clearly demonstrates the superior performance of the proposed MES approach to all other approaches tested in the experiments on three different face databases.

Table 1. Equal Error Rate (ERR %) of Various Approaches on Three Different Databases

Databases	AR	FERET1	FERET2
PCAE	34.2	5.1	13.0
PCAM	21.9	6.0	15.5
FDA	13.2	35.0	34.8
NDA	5.0	4.0	3.6
DLA	5.2	27.0	10.5
BML	19.1	7.9	23.8
UFS	4.3	3.0	2.7
DSL	5.9	3.8	2.5
MES	**2.6**	**2.0**	**2.0**

Summary of the Experimental Results

We have conducted 3 sets of experiments on 3 different face databases that evaluate 9 subspace based approaches for the face verification task. Unlike face identification experiments where some sample images of all probe subjects can be included in the training, in all verification experiments of this work, the subjects of the probe imposters are excluded in the training. Moreover, in FERET database 1, the training subjects are different from those in the gallery and probe sets. The experimental results verify the difference in terms of accuracy between the face verification and the face identification. Methods that work well for the face identification may not necessarily do the same for the face verification task. BML is a good example for this. It is thus useful to test the verification performances of various approaches that were developed and tested for identification task.

The experiments show that UFS, NDA and DSL approaches perform better than PCAE, PCAM, FLDA, DLDA and BML approaches. UFS keeps only a small principal subspace with largest eigenvalues for the discriminant evaluation. It suppresses more noise and thus has less over-fitting problem comparing to the FLDA and DLDA that perform the discriminant evaluation in the whole range space. The good performance of NDA verifies that the null space contains important discriminative information and should not be simply discarded in the feature extraction. Another property of NDA is that it does not scale the features by the eigenvalues. This is one possible reason why NDA has better generalization than FLDA, DLDA and BML. DSL extract two sets of features, one from a principal subspace and the other from its complementary subspace including the null space. Its relative good performance shows that the discriminative information resides in the both subspaces.

However, none of the three better approaches, UFS, NDA and DSL can consistently achieve the second best performance in the three experiments. One reason could be that all of them are suboptimal that extract features by the discriminant evaluation in a subspace or separately in two subspaces. The proposed MES method shows superior verification performances to all the other 8 subspace based approaches. In all three experiments on the different face databases, the proposed MES method consistently achieves the lowest error rates at all different operating points. It is important to test a verification system at different operating points because there is no optimal threshold for a verification system and different applications in practice has different requirement of FAR and FRR. The superior verification performance of the proposed method is attributed to the eigenspectrum modeling that enables a global optimization by the discriminant evaluation in the whole space and alleviates the over-fitting problem by replacing the unreliable or noise sensitive small and zero eigenvalues by the modeled values.

CONCLUSIONS

There are some intrinsic differences between the face verification and face identification. A method that performs well for the identification task may not necessarily achieve a high accuracy for the verification task. This paper addresses the face verification problem and explores several popular subspace based approaches for face verification. Experiments on three face databases compare the verification performances of eight well known approaches, PCAE, PCAM, FLDA, NDA, DLDA, BML, UFS and DSL. The verification performances of some of these approaches are indeed quite different from those in the identification evaluation reported in the literature.

This work shows problems of feature scaling and extraction from high dimensional data such as face image that are often encountered in computer vision and pattern recognition where the within-class scatter matrix degenerates due to the very small and zero eigenvalues. To alleviate these problems we decompose the eignespace into three subspaces and generate an eigenspectrum model for face data. The proposed eigenspectrum model alleviates the over-fitting problem by replacing the small and zero eigenvalues in the noise dominating and null spaces with the model values. It also enables a global optimization in the feature extraction by performing the discriminant evaluation in the whole space. Therefore, the extracted features are the most discriminative in the whole space and stable or less sensitive to the noise disturbance, the data dimensionality and the number of training samples. Extensive experiments on three face databases demonstrate that the proposed approach consistently outperforms other 8 popular subspace based approaches tested in this work.

REFERENCES

[1] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey", *ACM Comput. Surv.*, vol. 35, no. 4, pp. 399-458, 2003.

[2] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge", in IEEE Conference on Computer Vision & Pattern Recognition, San Diego, pp. 947-954, 2005.

[3] X. D. Jiang, B. Mandal, and A. Kot, "Complete Discriminant Evaluation and Feature Extraction in Kernel Space for Face Recognition", *Mach. Vis. Appl.*, to appear, DOI: 10.1007/s00138-007-0103-1, 2008.

[4] M. Sadeghi, and J. Kittler, "Decision making in the lda space: Generalised gradient direction metric", in the Sixth IEEE Int. Conf. Automatic Face and Gesture Recognition, 2004, pp. 248-253.

[5] H. Liu, C. Su, Y. Chiang, and Y. Hung, "Personalized face verification system using owner-specific cluster-dependent lda-subspace", in *Int. Conf. Pattern Recogn*., pp. 344-347, 2004.

[6] M. Kirby, and L. Sirovich, "Application of karhunen-loeve procedure for the characterization of human faces*", IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 12, no. 1, pp. 103-108, 1990.

[7] B. Moghaddam, and A. Pentland, "Probabilistic visual learning for object representation", *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 19, no. 7, pp. 696-710, 1997.

[8] B. Moghaddam, T. Jebara, and A. Pentland, "Bayesian face recognition", *Pattern Recognit.*, vol. 33, no. 11, pp. 1771-1782, 2000.

[9] B. Moghaddam, "Principal manifolds and probabilistic subspace for visual recognition", *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 24, no. 6, pp. 780-788, 2002.

[10] D. L. Swets, and J. Weng, "Using discriminant eigenfeatures for image retrieval", *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 18, no. 8, pp. 831-836, 1996.

[11] C. Liu, and H. Wechsler, "Enhanced Fisher linear discriminant models for face recognition", in *Int. Conf. Pattern Recogn*., Queensland, Australia, pp. 1368-1372, 1998.

[12] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces *vs* Fisherfaces: Recognition using class specific linear projection", *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 19, no. 7, pp. 711-720, 1997.

[13] X. D. Jiang, B. Mandal, and A. Kot, "Face Recognition Based On Discriminant Evaluation in the Whole Space", *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Hawaii, April, 2007.

[14] X. Wang, and X. Tang, "Dual-space linear discriminant analysis for face recognition", in *IEEE Conf. Compu. Vis. Pattern Recognit.,* Washington DC, USA, pp. 564-569.

[15] H. Cevikalp, M. Neamtu, M. Wilkes, and A. Barkana, "Discriminative common vectors for face recognition", *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 27, no. 1, pp. 4-13, 2005.

[16] H. Yu, and J. Yang, "A direct LDA algorithm for high-dimensional data with application to face recognition", *Pattern Recognit.*, vol. 34, no. 10, pp. 2067-2070, 2001.

[17] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Face recognition using LDA-based algorithms", *IEEE Trans. Neural Netw.,* vol. 14, no. 1, pp. 195-200, 2003.

[18] W. Liu, Y. Wang, S. Z. Li, and T. N. Tan, "Null space approach of Fisher discriminant analysis for face recognition", in ECCV workshop on Biometric Authentication, Prague, Czech Republic, pp. 32-44, 2004.

[19] J. Yang, A. F. Frangi, J. Y. Yang, D. Zhang, and Z. Jin, "KPCA plus LDA: A complete kernel Fisher discriminant framework for feature extraction and recognition", *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 27, no. 2 pp. 230-244, 2005.

[20] X. Wang, and X. Tang, "A unified framework for subspace face recognition", *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 26, no. 9, pp. 1222-1228, 2004.

[21] B. Mandal, X. D. Jiang, and A. Kot, "Dimensionality Reduction in Subspace Face Recognition", IEEE Sixth International Conference on Information, Communications and Signal Processing, Singapore, 10-13 December 2007, pp. 1-5.

[22] M. Teixeira, "The bayesian intrapersonal/extrapersonal classifier", MS Thesis: http://www.cs.colostate.edu/evalfacerec/papers/teixeiraThesis.pdf, 2003. [Accessed May 17, 2008].

[23] X. D. Jiang, B. Mandal, and A. Kot, "Eigenfeature Regularization and Extraction in Face Recognition", *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 30, no. 3, pp. 383–394, 2008.

[24] X. D. Jiang, B. Mandal, and A. Kot, "Enhanced maximum likelihood face recognition", *Electron. Lett.,* vol. 42, no. 19, pp. 1089-1090, 2006.

[25] K. Fukunaga, "*Introduction to Statistical Pattern Recognition*", Academic Press, second edition, 1991.

[26] M. Teixeira, R. Beveridge, D. Bolme, M. Teixeira, and B. Draper, "The csu face identification evaluation system users guide: Version 5.0, 2005, Technical Report": http://www.cs.colostate.edu/evalfacerec/data/normalization.html. [Accessed May 17, 2008].

[27] A. M. Martinez, "Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class", *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 24, no. 6, pp. 748-763, 2002.

[28] B. G. Park, K. M. Lee, and S. U. Lee, "Face recognition using face-arg matching," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 27, no. 12, pp. 1982-1988, 2005.

[29] Bazin, and M. Nixon, "Facial verification using probabilistic methods", in British Machine Vision Association Workshop on Biometrics, London, 2004.

[30] P. J. Phillips, H. Moon, S. Rizvi, and P. Rauss, "The feret evaluation methodology for face recognition algorithms", *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 22, no. 10, pp. 1090-1104, 2000.

[31] J. Lu, K. N. Plataniotis, A. N. Venetsanopoulos, and S. Z. Li, "Ensemble-based discriminant learning with boosting for face recognition", *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 166-178, 2006.

Received: March 25, 2008 Revised: May 19, 2008 Accepted: May 26, 2008