# Eigenfeature Regularization and Extraction in Face Recognition

Xudong Jiang, *Senior Member*, *IEEE*, Bappaditya Mandal, and Alex Kot, *Fellow*, *IEEE*

**Abstract**—This work proposes a subspace approach that regularizes and extracts eigenfeatures from the face image. Eigenspace of the within-class scatter matrix is decomposed into three subspaces: a reliable subspace spanned mainly by the facial variation, an unstable subspace due to noise and finite number of training samples, and a null subspace. Eigenfeatures are regularized differently in these three subspaces based on an eigenspectrum model to alleviate problems of instability, overfitting, or poor generalization. This also enables the discriminant evaluation performed in the whole space. Feature extraction or dimensionality reduction occurs only at the final stage after the discriminant assessment. These efforts facilitate a discriminative and a stable low-dimensional feature representation of the face image. Experiments comparing the proposed approach with some other popular subspace methods on the FERET, ORL, AR, and GT databases show that our method consistently outperforms others.

**Index Terms**—Face recognition, linear discriminant analysis, regularization, feature extraction, subspace methods.

## 1 Introduction

Face recognition has attracted many researchers in the area of pattern recognition, machine learning, and computer vision because of its immense application potential. Numerous methods have been proposed in the last two decades [1], [2]. However, there are still substantial challenging problems, which remain to be unsolved. One of the critical issues is how to extract discriminative and stable features for classification. Linear subspace analysis has been extensively studied and becomes a popular feature extraction method since the principal component analysis (PCA) [3], Bayesian maximum likelihood (BML) [4], [5], [6], and linear discriminant analysis (LDA) [7], [8] were introduced into face recognition. A theoretical analysis showed that a low-dimensional linear subspace could capture the set of images of an object produced by a variety of lighting conditions [9]. The agreeable properties of the linear subspace analysis and its promising performance achieved in the face recognition encourage researchers to extend it to higher order statistics [10], [11], nonlinear methods [12], [13], [14], Gabor features [15], [16], and locality preserving projections [17], [18]. However, the basic linear subspace analysis has still outstanding challenging problems when applied to the face recognition due to the high dimensionality of face images and the finite number of training samples in practice.

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 the 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.

Numerous methods have been proposed to solve this problem in the last decade. A popular approach called Fisherface (FLDA) [19] 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 result in the loss of discriminative information [20], [21], [22]. Direct LDA (DLDA) method [23], [24] 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 (NLDA) approaches [25], [26], [22] assume 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 both subspaces. To use both subspaces, a modified LDA approach [27] replaces the within-class scatter matrix by the total scatter matrix. Subsequent work [28] extracts features separately from the principal and null spaces of the within-class scatter matrix. However, the extracted features may not be properly scaled and undue emphasis is placed on the null space in these two approaches due to the replacement of the within-class scatter matrix by the total scatter matrix. The dual-space

- *The authors are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Nanyang Link, Singapore 639798. E-mail: {exdjiang, bapp0001, eackot}@ntu.edu.sg.*

*Manuscript received 27 June 2006; revised 6 Nov. 2006; accepted 15 May 2007; published online 31 May 2007.*
*Recommended for acceptance by T. Tan.*
*For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0475-0606.*
*Digital Object Identifier no. 10.1109/TPAMI.2007.70708.*

0162-8828/08/$25.00 © 2008 IEEE Published by the IEEE Computer Society

approach (DSL) [21] 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 [21], 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 [14]. Open questions of these 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 both subspaces, it is inefficient or 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 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 [29], and a unified framework of subspace methods (UFS) was proposed. Good recognition performance of this framework shown in [29] verifies the importance of noise suppression. However, this approach applies three stages of subspace decompositions sequentially on the face training data, and dimensionality reduction occurs at the very first stage. As addressed in the literature [20], [21], [22], applying PCA for dimensionality reduction may result in the lost of 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 at the third stage. The experimental results in [29] show that recognition performance is sensitive to these choices at different stages.

In this paper, we present a new approach for facial eigenfeature regularization and extraction. Image space spanned by the eigenvectors of the within-class scatter matrix is decomposed into three subspaces. Eigenfeatures are regularized differently in these subspaces based on an eigenspectrum model. This alleviates the problem of unreliable small and zero eigenvalues caused by noise and the limited number of training samples. It also enables discriminant evaluation to be performed in the full dimension of the image data. Feature extraction or dimensionality reduction occurs only at the final stage after the discriminant assessment. In Section 2, we model the eigenspectrum, study the effect of the unreliable small eigenvalues on the feature weighting, and decompose the eigenspace into face, noise, and null subspaces. Eigenfeature regularization and extraction are presented in Section 3. Analysis of the proposed approach and comparison with other relevant methods are provided in Section 4. Experimental results are presented in Section 5 before drawing conclusions in Section 6.

## 2 Eigenspectrum Modeling and Subspace Decomposition

Given a set of properly normalized $w$-by-$h$ face images, we can form a training set of column image vectors $\{X_{ij}\}$, where $X_{ij} \in \mathbb{R}^{n=wh}$, by lexicographic 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 sample is $l = \sum_{i=1}^{p} q_i$. For face recognition, each person is a class with prior probability of $c_i$. The within-class scatter matrix is defined by

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

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

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

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

where $\overline{X} = \sum_{i=1}^{p} c_i \overline{X}_i$. If all classes have equal prior probability, then $c_i = 1/p$.

Let $\mathbf{S}^g$, $g \in \{t, w, b\}$ represent one of the above scatter matrices. If we regard the elements of the image vector and the class mean vector as features, these preliminary features will be decorrelated by solving the eigenvalue problem

$$\mathbf{\Lambda}^g = \mathbf{\Phi}^{gT} \mathbf{S}^g \mathbf{\Phi}^g, \tag{4}$$

where $\mathbf{\Phi}^g = [\phi_1^g, \dots, \phi_n^g]$ is the eigenvector matrix of $\mathbf{S}^g$, and $\mathbf{\Lambda}^g$ is the diagonal matrix of eigenvalues $\lambda_1^g, \dots, \lambda_n^g$ corresponding to the eigenvectors.

Suppose that the eigenvalues are sorted in descending order $\lambda_1^g \geq, \dots, \geq \lambda_n^g$. The plot of eigenvalues $\lambda_k^g$ against the index $k$ is called eigenspectrum of the face training data. It plays a critical role in subspace methods as the eigenvalues are used to scale and extract features. We first model the eigenspectrum to show its problems in feature scaling and extraction.

### 2.1 Eigenspectrum Modeling

If we regard $X_{ij}$ as samples of a random variable vector $X$, the eigenvalue $\lambda_k^g$ is a variance estimate of $X$ projected on the eigenvector $\phi_k^g$ estimated from the training samples. It usually deviates from the true variance of the projected random vector $X$ due to the finite number of training samples. Thus, we model the eigenspectrum in the range subspace $\lambda_k^g$, $1 \leq k \leq r$, as the sum of the true variance component $v_k^F$ and a deviation component $\delta_k$. For simplicity, we call $v_k^F$ face component and $\delta_k$ noise component. As the face component 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 [4] to extrapolate eigenvalues in the null space for computing the average eigenvalue over a subspace. The noise component $\delta_k$ that includes the effect of the finite number of training samples can be negative if the face component $v_k^F$ is modeled by a $1/f$ function that always has positive values.

We propose to model the eigenvalues first in descending order of the face component $v_k^F$ by

$$\hat{\lambda}_k^F = v_k^F + \delta_k = \frac{\alpha}{k + \beta} + \delta_k, \quad 1 \leq k \leq r, \tag{5}$$

where $\alpha$ and $\beta$ are two constants that will be given in Section 3.1. The modeled eigenspectrum $\hat{\lambda}_k^g$ is then obtained by sorting $\hat{\lambda}_k^F$ in descending order. As the eigenspectrum

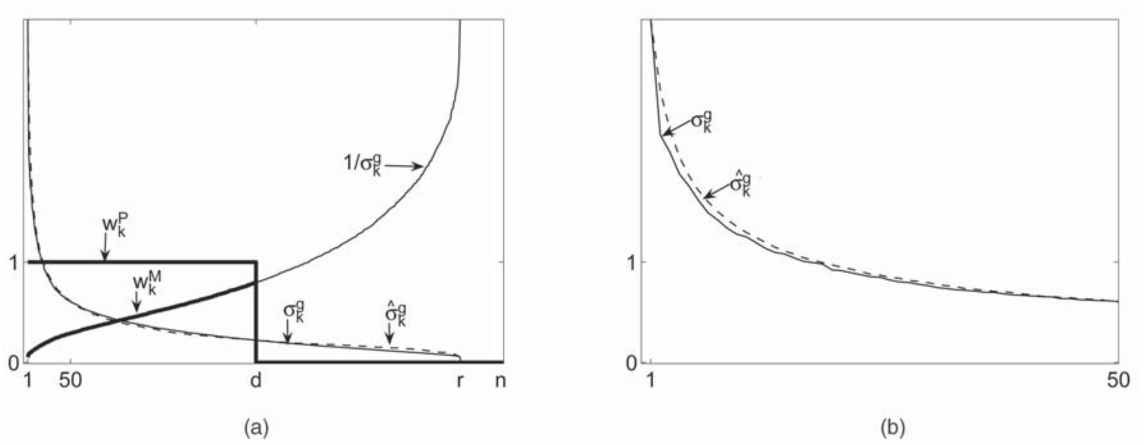

Fig. 1. A typical real eigenspectrum, (a) its model and feature weighting/extraction of PCA. The first 50 real eigenvalues and (b) the values of the model.

decays very fast, we plot the square roots $\sigma_k^g = \sqrt{\lambda_k^g}$ and $\hat{\sigma}_k^g = \sqrt{\hat{\lambda}_k^g}$ for clearer illustration (we still call them eigenspectrum for simplicity). A typical real eigenspectrum $\sigma_k^g$ and its model $\hat{\sigma}_k^g$ are shown in Fig. 1, where $\delta_k$ is a random number evenly distributed in a small range of $[-\lambda_1^g/1000, \lambda_1^g/1000]$. In Fig. 1, we see that the proposed model closely approximates to the real eigenspectrum.

## 2.2 Problems of Feature Scaling and Extraction

The PCA approach with euclidean distance selects $d$ leading eigenvectors and discards the others. This can be seen as weighting the eigenfeatures with a step function as

$$u_k^d = \begin{cases} 1, & k \leq d \\ 0, & k > d. \end{cases} \tag{6}$$

This weighting function of PCA $w_k^P = u_k^d$ is shown in Fig. 1. In a practical face recognition problem, the recognition performance usually improves with the increase of $d$. The PCA with Mahalanobis (PCAM) distance can be seen as the PCA with euclidean distance and a weighting function $w_k^M = u_k^d/\sigma_k^g$. This weighting function is shown in Fig. 1. Using the inverse of the square root of the eigenvalue to weigh the eigenfeature makes a large difference in the face recognition performance. The recognition accuracy usually increases sharply with the increase of $d$ and is better than the euclidean distance for smaller $d$. However, the recognition accuracy decreases also sharply and is much worse than the euclidean distance for larger $d$. The BML approach that weights the principal features $(1 \leq k \leq d)$ by the inverse of the square root of the eigenvalue also suffers from the performance decrease with the increase of the $d$ after $d$ reaches a small value.

Noise disturbance and poor estimates of small eigenvalues due to the finite number of training samples are the culprits. The limited number of training samples results in very small eigenvalues in some dimensions that may not well represent the true variance in these dimensions. This may result in serious problems if their inverses are used to weight the eigenfeatures. The characteristics of the eigenspectrum and the generalization deterioration caused by the small eigenvalues were well addressed in [30]. To exclude the small eigenvalues in the discriminant evaluation, probabilistic reasoning models and enhanced FLDA models were proposed and compared in [30]. The modeled eigenspectrum in Fig. 1 that approximates closely to the real one is resorted in descending order of the face component $v_k^F$ and is plotted in Fig. 2. The small noise disturbances are now visible in Fig. 2 given that the variances of the face component should always decay. These small disturbances cause large vibrations of the inverse eigenspectrum, as shown in Fig. 2. Eigenfeatures of larger index $k$ are heavily weighted by the scaling factors that are highly sensitive to noise and training data. This causes the deterioration of the recognition performance, especially on the independent testing data.

## 2.3 Subspace Decomposition

As shown in Fig. 2, small noise disturbances that have little effect on the initial portion of the eigenspectrum cause large vibrations of the inverse eigenspectrum in the region of small eigenvalues. Therefore, we propose to decompose the eigenspace $\mathbb{R}^n$ spanned by eigenvectors $\{\phi_k^g\}_{k=1}^n$ into three subspaces: a reliable face variation dominating subspace (or simply face space) $\mathbf{F} = \{\phi_k^g\}_{k=1}^m$, an unstable noise variation dominating subspace (or simply noise space) $\mathbf{N} = \{\phi_k^g\}_{k=m+1}^r$ and a null space $\emptyset = \{\phi_k^g\}_{k=r+1}^n$, as illustrated in Fig. 2. The purpose of this decomposition is to modify or regularize the unreliable eigenvalues for better generalization.

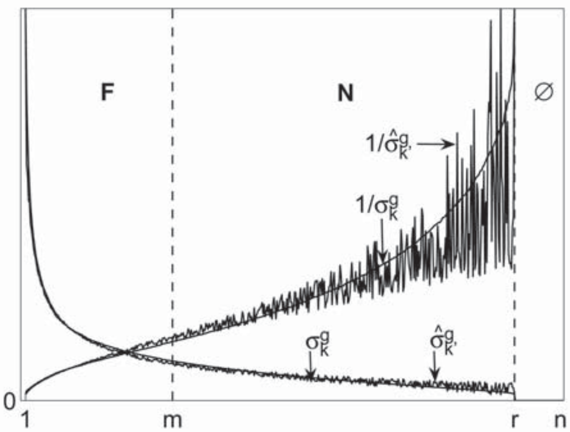

Fig. 2. A typical real eigenspectrum, its model sorted in descending order of the face component and their inverse; decomposition of the eigenspace into face, noise, and null-subspaces.

It is not difficult to determine the rank of the scatter matrix $r$. For $\mathbf{S}^t$, $r \leq min(n, l-1)$, for $\mathbf{S}^w$, $r \leq min(n, l-p)$, and for $\mathbf{S}^b$, $r \leq min(n, p-1)$. In practice, the rank of a scatter matrix usually reaches the corresponding one of these maximum values unless some training images are linearly dependent. Even in this rare case, the rank $r$ can be easily determined by finding the maximal value of $k$ that satisfies $\lambda_k^g > \varepsilon$, where $\varepsilon$ is a very small positive value comparing to $\lambda_1^g$.

As face images have similar structure, significant face components reside intrinsically in a very low-dimensional ($m$-dimensional) subspace. For a robust training, the database size should be significantly larger than the face dimensionality $m$, although it could be, and usually in practice is, much smaller than the image dimensionality $n$. Thus, in many practical face recognition training tasks, we usually have $m \ll r \ll n$. 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 $m+1$, we first find a point near the center of the noise region by

$$\lambda_{med}^g = \text{median}\{\forall \lambda_k^g | k \leq r\}. \tag{7}$$

The distance between $\lambda_{med}^g$ and the smallest nonzero eigenvalue is $d_{m,r} = \lambda_{med}^g - \lambda_r^g$. The upper bound of the unreliable eigenvalues is estimated by $\lambda_{med}^g + d_{m,r}$. 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 the number of training samples in different applications. More generally, the start point of the noise region $m+1$ is estimated by

$$\lambda_{m+1}^g = \max\{\forall \lambda_k^g | \lambda_k^g < (\lambda_{med}^g + \mu(\lambda_{med}^g - \lambda_r^g))\}, \tag{8}$$

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

## 3 Eigenfeature Regularization and Extraction

In general, it is desired to extract features that have the smallest within-class variations and the largest between-class variations. Problems occur in seeking the smallest within-class variations because the variances are estimated based on the finite number of training samples and, hence, the estimated smaller variances are unstable and tend to overfit the specific training data. It is not a surprise that the within-class variation is the biggest obstacle to achieve high recognition rate. Thus, we first work on the within-class scatter matrix $\mathbf{S}^w$.

After solving the eigenvalue problem as (4), a unit within-class scatter matrix in the subspace $\bar{Y}_{ij} \in \mathbb{R}^r$ can be obtained by representing training samples with new feature vectors $\bar{Y}_{ij}$

$$\bar{Y}_{ij} = \bar{\mathbf{\Phi}}_r^{wT} X_{ij}, \tag{9}$$

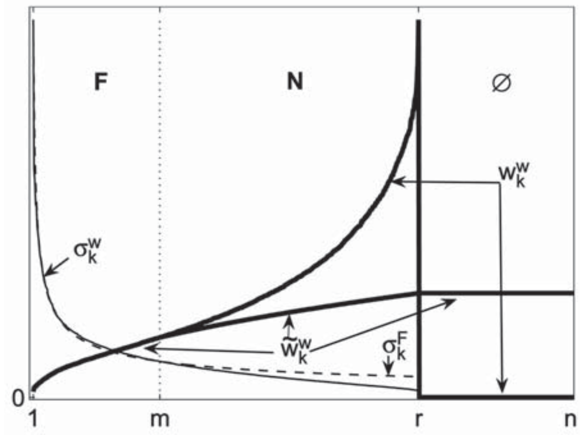

Fig. 3. Weighting functions of (10) and (14) in the face, noise, and null subspaces based on a typical real eigenspectrum.

where $\bar{\mathbf{\Phi}}_r^w = [\phi_1^w/\sigma_1^w, \ldots, \phi_r^w/\sigma_r^w]$ are composed of so called whitened eigenvectors of $\mathbf{S}^w$ with nonzero eigenvalues, and $\|\phi_k^w\| = 1$. Thus, an $n$-dimensional image vector $X_{ij}$ is first represented by an $n$-dimensional eigenfeature vector $Y_{ij} = \mathbf{\Phi}^{wT} X_{ij}$ and then multiplied by a weighting function:

$$w_k^w = \begin{cases} 1/\sqrt{\lambda_k^w}, & k \leq r \\ 0 \quad\quad , & r < k \leq n, \end{cases} \tag{10}$$

as shown in Fig. 3. The training data represented by the new feature vectors $\bar{Y}_{ij}$ will produce the same (unit) within-class variances in all directions of the reduced feature space $\mathbb{R}^r$. It appears that the problem of the within-class variation is solved.

However, there are two problems: First, face structural information in the null space is lost or, equivalently, features in the null space is weighted by a constant zero. This is unreasonable because features in the null space are of zero within-class variances based on the training data and, hence, should be much heavier 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. 3. Second, weights determined by the inverse of $\sigma_k^w$ is, though optimal in terms of the ML estimation, dangerous when $\sigma_k^w$ is small. This is shown in the modeled eigenspectrum in Fig. 2. The small and zero eigenvalues are training-set specific and very sensitive to different training sets. 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, eigenspectrum needs to be regularized.

### 3.1 Eigenspectrum Regularization

Although there is always noise component in $\mathbf{F}$ 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 $\mathbf{F}$. In $\mathbf{N}$, 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 (5).

As the eigenspectrum in the face space is dominated by the face structural component, the parameters of $\alpha$ and $\beta$ are determined by fitting the face portion of the model $v_k^F = \alpha/(k+\beta)$ to the real eigenspectrum in the face space $\mathbf{F}$. Although not limiting ourselves from other possible fitting methods, in all experiments of this work, we simply determine $\alpha$ and $\beta$ by letting $v_1^F = \lambda_1^w$ and $v_m^F = \lambda_m^w$, which yields

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

$$\beta = \frac{m\lambda_m^w - \lambda_1^w}{\lambda_1^w - \lambda_m^w}. \tag{12}$$

Fig. 3 shows the square roots of a real eigenspectrum $\sigma_k^w$ and the face portion of its model $\sigma_k^F = \sqrt{v_k^F}$. We see that the model $\sigma_k^F$ fits closely to the real $\sigma_k^w$ in the face space $\mathbf{F}$ but has slower decay in the noise space $\mathbf{N}$. The faster decay of the real eigenspectrum $\sigma_k^w$ in $\mathbf{N}$ due to noise and the effect of the limited number of training samples is what we want to slow down.

In the null space, we have no information about the variation of the eigenvalues and, hence, all features are treated in the same way. The zero 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 regularize the eigenspectrum by replacing the noise dominating $\lambda_k^w$ in $\mathbf{N}$ with the face portion of the model $v_k^F = \alpha/(k+\beta)$ and replacing the zero $\lambda_k^w$ in the null space $\emptyset$ with the constant $v_{r+1}^F$. Thus, the regularized eigenspectrum $\tilde{\lambda}_k^w$ is given by

$$\tilde{\lambda}_k^w = \begin{cases} \lambda_k^w, & k < m \\ \frac{\alpha}{k+\beta}, & m \le k \le r \\ \frac{\alpha}{r+1+\beta}, & r < k \le n. \end{cases} \tag{13}$$

The proposed feature weighting function is then

$$\tilde{w}_k^w = \frac{1}{\sqrt{\tilde{\lambda}_k^w}}, \quad k = 1, 2, \ldots n. \tag{14}$$

Fig. 3 shows the proposed feature weighting function $\tilde{w}_k^w$ calculated by (11), (12), (13), and (14) comparing with that $w_k^w$ of (10). Obviously, the new weighting function $\tilde{w}_k^w$ is identical to $w_k^w$ in the face space, increases along with $k$ at a much slower pace than $w_k^w$ in the noise space, and has maximal weights instead of zero of $w_k^w$ in the null space.

Using this weighting function and the eigenvectors $\phi_k^w$, the training data are transformed to

$$\tilde{Y}_{ij} = \tilde{\mathbf{\Phi}}_n^{wT} X_{ij}, \tag{15}$$

where

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

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

### 3.2 Discriminant Eigenfeature Extraction

After the feature regularization, a new total scatter matrix is formed by vectors $\tilde{Y}_{ij}$ of the training data as

$$\tilde{\mathbf{S}}^t = \sum_{i=1}^{p} \frac{c_i}{q_i} \sum_{j=1}^{q_i} (\tilde{Y}_{ij} - \overline{Y})(\tilde{Y}_{ij} - \overline{Y})^T, \tag{17}$$

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

There is some difference between the total scatter matrix $\tilde{\mathbf{S}}^t$ and the between-class scatter matrix $\tilde{\mathbf{S}}^b$ of the training data $\tilde{Y}_{ij}$ because the within-class scatter matrix is not fully whitened in the noise space $\mathbf{N}$, and the null space $\emptyset$ is not discarded. Some works [31], [18] show that when the training data are small, PCA can outperform LDA and that PCA is less sensitive to the different training databases. Our experiments show that $\tilde{\mathbf{S}}^t$ outperforms $\tilde{\mathbf{S}}^b$, but only very marginally and not consistently for different numbers of features. For a training set that contains images of only a few people, $\tilde{\mathbf{S}}^b$ may have a problem of extracting sufficient number of features. The maximal number of features extracted based on $\tilde{\mathbf{S}}^b$ is $p-1$ at most while that based on $\tilde{\mathbf{S}}^t$ could be much larger $(\sum_{i=1}^p q_i - 1)$. Thus, in this work, we suggest to employ the total scatter matrix $\tilde{\mathbf{S}}^t$ of the regularized training data to extract the discriminative features.

The regularized features $\tilde{Y}_{ij}$ will be decorrelated for $\tilde{\mathbf{S}}^t$ by solving the eigenvalue problem as (4). Suppose that the eigenvectors in the eigenvector matrix $\tilde{\mathbf{\Phi}}_n^t = [\tilde{\phi}_1^t, \ldots, \tilde{\phi}_n^t]$ are sorted in a descending order of the corresponding eigenvalues. The dimensionality reduction is performed here by keeping the eigenvectors with the $d$ largest eigenvalues

$$\tilde{\mathbf{\Phi}}_d^t = \left[\tilde{\phi}_k^t\right]_{k=1}^d = \left[\tilde{\phi}_1^t, \ldots, \tilde{\phi}_d^t\right], \tag{18}$$

where $d$ is the number of features usually selected by a specific application.

Thus, the proposed feature regularization and extraction matrix $\mathbf{U}$ is given by

$$\mathbf{U} = \tilde{\mathbf{\Phi}}_n^w \tilde{\mathbf{\Phi}}_d^t, \tag{19}$$

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

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

We see that the decomposition of the image space into three subspaces is used only for the regularization of the feature scaling. The discriminant evaluation (here, the evaluation of the eigenvalues of $\tilde{\mathbf{S}}^t$) is performed in the full space $\mathbb{R}^n$. Thus, feature extraction is not restricted to project an image vector into one or two of these three subspaces. More specifically, any single feature in $F$ is extracted from the whole space $\mathbb{R}^n$ since any final projection vector in $\mathbf{U}$ may have nonzero components in all the three subspaces.

### 3.3 The Proposed Algorithm

The proposed eigenfeature regularization and extraction (ERE) approach is summarized below:

At the training stage

1. Given a training set of face image vectors $\{X_{ij}\}$, compute $\mathbf{S}^w$ by (1) and solve the eigenvalue problem as (4).
2. Decompose the eigenspace into face, noise, and null spaces by determining the $m$ value using (7) and (8).
3. Transform the training samples represented by $X_{ij}$ into $\tilde{Y}_{ij}$ by (15) with the weighting function (14) determined by (11), (12), and (13).
4. Compute $\tilde{\mathbf{S}}^t$ by (17) with $\tilde{Y}_{ij}$ and solve the eigenvalue problem as (4).
5. Obtain the final feature regularization and extraction matrix by (16), (18), and (19) with a predefined number of features $d$.

At the recognition stage

1. Transform each $n$-D face image vector $X$ into $d$-D feature vector $F$ by (20) using the feature regularization and extraction matrix $\mathbf{U}$ obtained in the training stage.
2. Apply a classifier trained on the gallery set to recognize the probe feature vectors.

In the experiments of this work, a simple first nearest neighborhood classifier (1-NNK) is applied to test the proposed ERE approach. 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 to the proposed approach, where $\| \cdot \|_2$ is the norm 2 operator.

## 4 Analysis and Comparison

Two techniques are developed and integrated in the proposed ERE approach for face recognition. First, it evaluates the discriminant value in the whole space and the dimensionality reduction or feature extraction occurs only at the final stage after the discriminant assessment. Second, the proposed eigenspectrum regularization not only facilitates this discriminative information search in the whole space but also alleviates the overfitting problem.

### 4.1 Discriminant Evaluation in a Subspace, Complementary Subspaces, and the Whole Space

Most subspace approaches such as FLDA, DLDA, NLDA, and UFS discard a subspace before the discriminant evaluation. The extracted features are only suboptimal as they 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 in matching. A modified LDA approach [27] estimates the discriminant value in the full space of $\mathbf{S}^t$ by replacing $\mathbf{S}^w$ with $\mathbf{S}^t$. It is easy to see that, after whitening by $\mathbf{S}^t$, the between-class variance is always one in all dimensions of the null space of $\mathbf{S}^w$ but is always smaller than one in all dimensions of the range space of $\mathbf{S}^w$. Obviously, the null space of $\mathbf{S}^w$ is unduly overemphasized, and the extracted features may not be properly scaled due to the replacement of $\mathbf{S}^w$ by $\mathbf{S}^t$ in the discriminant assessment.

Some approaches [28], [21], [14] extract features separately from the two complementary subspaces of $\mathbf{S}^w$. One of them [28] first extracts all the $p-1$ features from the null space of $\mathbf{S}^w$. It then extracts some features from the range space of $\mathbf{S}^w$, where $\mathbf{S}^t$ is used as a replacement of $\mathbf{S}^w$ in the discriminant evaluation. Obviously, this method, similar to the approach in [27], overemphasizes the null space and may not properly scale the features. The DSL approach [21] scales features in the complementary subspace by the average eigenvalue of $\mathbf{S}^w$ over this subspace. As the eigenvalues in this subspace are not well estimated [21], their average may not be a proper scaling factor relative to those in the principal subspace. To circumvent the feature-scaling problem in the training phase, a summed normalized distance is proposed in [14] that properly combines the two groups of features in the recognition phase.

Although these double-space 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. Other open questions include how to divide the space into the two subspaces properly and how to apportion the given number of features to the two subspaces reasonably. In contrast, the proposed ERE approach searches the most discriminative information in the whole space, and the feature scaling problem caused by the small and zero eigenvalues is alleviated by the eigenspectrum regularization.

### 4.2 Eigenspectrum Regularization

It is well known that the estimated covariance matrix or its eigenvalues need to be regularized, especially for small number of training samples. Problems of the biased estimation of eigenvalues that cause overfitting and, hence, poor generalization were well addressed in [32].

A method called RD-QDA [33] was proposed to regularize $\mathbf{S}^w$ and applied in the face recognition. Another regularization approach called R-LDA [34] was derived by replacing $\mathbf{S}^w$ with $\mathbf{S}^w + \eta\mathbf{S}^b$, where $\eta$ is a regularization parameter. However, both approaches are based on the DLDA framework, and the appropriate values of the regularization parameters are different for different databases and database sizes. These critical values are determined by exhaustive search with the help of a validation data set in the experiment. Furthermore, both approaches regularize the pooled within-class scatter matrix equivalently by adding a constant to all eigenvalues. Although the largest sample-based eigenvalues are biased high and the smallest ones are biased low, as pointed out in [32], the bias is most pronounced when the population eigenvalues tend toward equality, and it is correspondingly less severe when their values are highly disparate. For the application of face recognition, it is well known that the eigenspectrum first decays very rapidly and then stabilizes. Hence, adding a constant to the eigenspectrum may bias back the rapidly changing eigenvalues in $\mathbf{F}$ too much that introduces additional error source and bias back the flat eigenvalues in $\mathbf{N}$ too little at the same time.

Therefore, the proposed approach decomposes the image space into a highly disparate face dominant subspace $\mathbf{F}$, a flat subspace $\mathbf{N}$ and a null space $\emptyset$, and regularizes eigenvalues differently in these three subspaces. The median operation well separates the highly disparate eigenvalues from the stabilized ones. As the eigenvalues tend toward equality in $\mathbf{N}$ and are highly disparate in $\mathbf{F}$, the distance $d_{m,r} = \lambda_{med} - \lambda_r$ that indicates the half range of the eigenvalue variation in $\mathbf{N}$ is very small compared to the highly disparate eigenvalues in $\mathbf{F}$. Therefore, $\lambda_{med} + d_{m,r}$ is proposed to separate the highly disparate subspace $\mathbf{F}$ from the flat subspace $\mathbf{N}$.

Although the finite number of training samples may also bias the eigenvalues in $\mathbf{F}$, we ignore this very small bias because the population eigenvalues are highly disparate in this subspace. In $\mathbf{N}$, however, the population eigenvalues tend toward equality and, hence, the finite number of training samples results in a much faster decay of the eigenvalues. Therefore, we slow down the decay of the eigenvalues by replacing the real eigenspectrum in $\mathbf{N}$ with the proposed model (5) that approximates to the real eigenspectrum in $\mathbf{F}$.

The zero eigenvalues in the null space are only estimates from one set of the training data. Adding new training samples will make them nonzero. However, based on the available training data that result in zero eigenvalues, the regularized eigenvalues in the null space should be smaller than those in the other subspaces. As we have no information about the variation of the eigenvalues in the null space, they are regularized as a positive constant smaller than the smallest regularized eigenvalue in the other subspaces.

BML can be seen as another eigenspectrum regularization approach that also keeps eigenvalues in the principal subspace unchanged and only modifies eigenvalues in the complementary subspace. DSL adopts the same regularization method. Both approaches replace eigenvalues in the complementary subspace with the constant calculated by the average eigenvalue over this subspace. Obviously, this will amplify some eigenvalues and attenuate the others in the complementary subspace. The unduly reduced eigenvalues in the complementary subspace may introduce additional overfitting problem [35] because the differences between these eigenvalues and those in the principal subspace are enlarged.

### 4.3 Computational Complexity

The computational complexity of the proposed ERE method for training is greater than other subspace methods as it evaluates the discriminant value in the full space of $\mathbf{S}^w$. Although the null spaces of $\mathbf{S}^w$ and $\mathbf{S}^b$ both contain discriminative information and, hence, are preserved in the proposed ERE approach, there is no statistical basis that the intersection of the null spaces of $\mathbf{S}^w$ and $\mathbf{S}^b$, that is, the null space of $\mathbf{S}^t$ contributes to the discriminative ability [28]. Therefore, we can apply PCA on $\mathbf{S}^t$ to remove the null space of $\mathbf{S}^t$ first and then apply the ERE approach on the $(l-1)$-dimensional subspace. This makes the ERE approach tractable for very large image sizes $(n \gg l)$. In practical applications, training is usually an offline process and recognition is usually an online process. Thus, the recognition time is usually much more critical than the training time. Although the recognition time of the ERE approach is the same as other approaches for the same number of features, it can be faster than other approaches for the same recognition rate because the proposed ERE approach, as we will see in the experiments, achieves a given recognition rate with fewer features than other approaches.

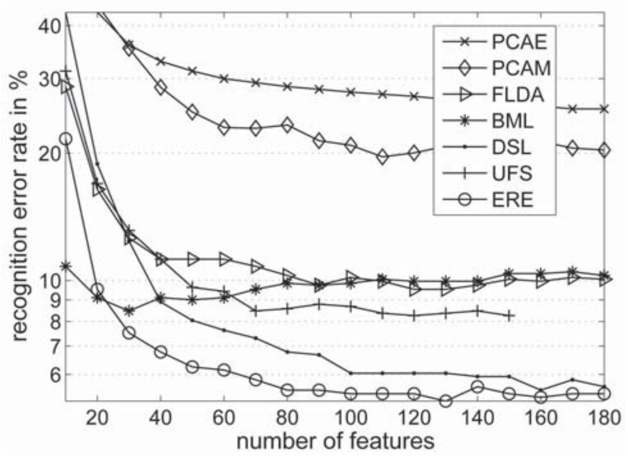

Fig. 4. Recognition error rate against the number of features used in the matching on the FERET database of 500 training images (250 people) and 1,888 testing images (944 people).

## 5 Experiments

In all experiments reported in this work, images are preprocessed following the CSU Face Identification Evaluation System [36]. Five databases ORL, AR, GT, and two from FERET are used for testing. Each database is partitioned into training and testing sets. For FERET databases, there is no overlap in person between the training and testing sets. As ORL, AR, and GT databases have only a small number of persons, both training and testing sets contain all persons. However, there is no overlap in the sample image between the training and testing sets. The recognition error rate given in this work is the percentage of the incorrect top 1 match on the testing set. The proposed ERE method is tested and graphically compared with the PCA with euclidian distance (PCAE), PCAM distance, FLDA, BML, DSL, and UFS approaches in four figures. Furthermore, three experimental results are numerically recoded in two tables, which include more results for comparison, such as the results of ERE using $\tilde{\mathbf{S}}^b$ (ERE_$\tilde{\mathbf{S}}^b$), NLDA, and SRLDA (a simple regularized LDA that replaces all zero eigenvalues of $\mathbf{S}^w$ with the minimum nonzero eigenvalue). The parameters of UFS applied are those which result in the best performance through an exhaustive search in the experiments of [29]. We implement FLDA by using PCA to reduce the dimensionality to $\eta(l-p)$, $0 < \eta \leq 1$ as the rank of $\mathbf{S}^w$ is $l-p$ at most. We present the best result of FLDA with $\eta$ varying from 0.7 to 1.

### 5.1 Results on FERET Database 1

There are 2,388 images comprising of 1,194 persons (two images FA/FB per person) selected from the FERET database [37]. Images are cropped into the size of $33 \times 38$. In the first experiment, images of 250 people are randomly selected for training, and the remaining images of 944 people are used for testing. Fig. 4 shows the recognition error rate on the testing set against the number of features $d$ used in the matching. Note that for the BML approach, the number of features used in the matching is the image dimensionality $n$ rather than $d$.

Fig. 4 shows that BML outperforms FLDA for small $d$ because BML has a large average eigenvalue $\rho$ in the

TABLE 1
Recognition Error Rate of Different Approaches for Different Number of Features

| Database | FERET (994 / 1394 training / testing images) | | | | | | | ORL (leave-one-out training-testing) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # Feature | 20 | 30 | 40 | 50 | 80 | 110 | 140 | 10 | 12 | 14 | 16 | 22 | 30 | 38 |
| PCAE | 35.87 | 31.13 | 28.26 | 27.26 | 24.82 | 23.39 | 22.24 | 11.00 | 9.00 | 9.00 | 8.25 | 7.00 | 7.00 | 7.25 |
| PCAM | 42.18 | 30.85 | 23.24 | 20.52 | 16.79 | 16.79 | 17.36 | 12.00 | 8.75 | 8.50 | 8.25 | 7.00 | 5.00 | 5.00 |
| FLDA | 14.35 | 10.76 | 8.75 | 9.33 | 8.46 | 8.61 | 8.18 | 6.25 | 5.25 | 3.75 | 3.00 | 2.50 | 2.50 | 2.00 |
| BML | 7.32 | 7.03 | 7.75 | 7.60 | 8.46 | 8.75 | 8.90 | 3.50 | 3.50 | 3.00 | 2.25 | 2.75 | 2.75 | 2.50 |
| DSL | 10.90 | 6.74 | 6.03 | 4.73 | 3.87 | 4.16 | 3.87 | 14.00 | 9.75 | 5.75 | 5.00 | 3.00 | 3.00 | 2.00 |
| UFS | 6.60 | 5.17 | 4.02 | 3.59 | 4.02 | 4.30 | 4.45 | 4.50 | 3.50 | 3.25 | 3.50 | 2.00 | 1.75 | 1.25 |
| NLDA | 10.62 | 8.46 | 7.32 | 7.32 | 6.31 | 6.89 | 7.17 | 6.00 | 4.25 | 3.00 | 3.00 | 2.50 | 2.25 | 2.00 |
| SRLDA | 8.46 | 7.03 | 5.88 | 5.74 | 4.88 | 5.31 | 5.88 | 4.50 | 3.25 | 3.25 | 2.25 | 2.00 | 2.00 | 1.25 |
| ERE_$\tilde{\mathbf{S}}^b$ | 5.31 | 4.45 | 3.01 | 2.87 | 3.30 | 3.44 | 3.30 | 2.50 | 2.50 | 2.00 | 1.25 | 1.50 | 1.00 | 0.75 |
| ERE_$\tilde{\mathbf{S}}^t$ | 5.17 | 4.30 | 3.01 | 2.87 | 3.30 | 3.44 | 3.16 | 2.75 | 2.25 | 2.00 | 1.25 | 1.50 | 1.25 | 1.00 |

complementary subspace, whereas FLDA suffers overfitting problem due to the small eigenvalues of $\mathbf{S}^w$. However, for large $d$, the small $\rho$ results in more overfitting of BML than FLDA because BML uses the inverse of $\rho$ to scale features of dimensions from $d+1$ to $n$, whereas FLDA discards the dimensions from $\eta(l-p)+1$ to $n$. UFS outperforms BML and FLDA as it solves the overfitting problem well by reducing the dimensionality to 150. DSL outperforms BML, FLDA, and UFS when a larger number of features are applied. This shows that the null space indeed contains useful discriminative information. However, extracting discriminative features separately from the two subspaces is not efficient for small number of features ($d < 40$ in Fig. 4). The proposed ERE approach consistently outperforms all other approaches for every number of features tested in the experiments, and the accuracy gain is significant for smaller number of features.

In the second experiment, more training samples (497 people) are randomly selected, and the remaining images of 697 people are used for testing. The recognition error rates are recorded in Table 1. All approaches show lower recognition error rates than those in Fig. 4 due to the larger number of training samples. The relative performances among BML, FLDA, and DSL are similar to those in Fig. 4. However, UFS achieves higher accuracy gain than BML, FLDA, and DSL. The ERE approach again outperforms all other approaches consistently in all columns in Table 1.

## 5.2 Results on FERET Database 2

This database is constructed, similar to one data set used in [38], by choosing 256 subjects with at least four images per subject. However, we use the same number of images (four) per subject for all subjects. Five hundred twelve images of the first 128 subjects are used for training, and the remaining 512 images serve as testing images. The size of the normalized image is $130 \times 150$, same as that in [38]. For such a large image size, we first apply PCA to remove the null space of $\mathbf{S}^t$ and then apply the ERE approach on the 511-dimensional feature vectors. The training time of the ERE approach by a Matlab program is about 1.3 times of that of FLDA approach. The $i$th images of all testing subjects are chosen to form a gallery set, and the remaining three images per subject serve as the probe images to be identified from the gallery set. Fig. 5 shows the average recognition error rates over the four probe sets, each of which has a distinct gallery set ($i = 1, 2, 3, 4$).

Comparing to that in Fig. 4, the recognition error rates of all methods in Fig. 5 increase due to larger variation of the testing images. They are also higher than those in [38] on a similar database because of the different training and testing procedures. In our experiment, there is no overlap in person between the training and testing sets and only one image per

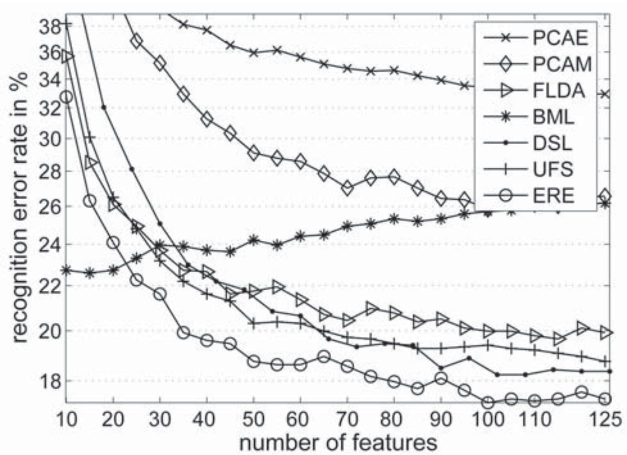

Fig. 5. Recognition error rate against the number of features used in the matching on the FERET database of 512 training images (128 people) and 512 testing images (128 people).

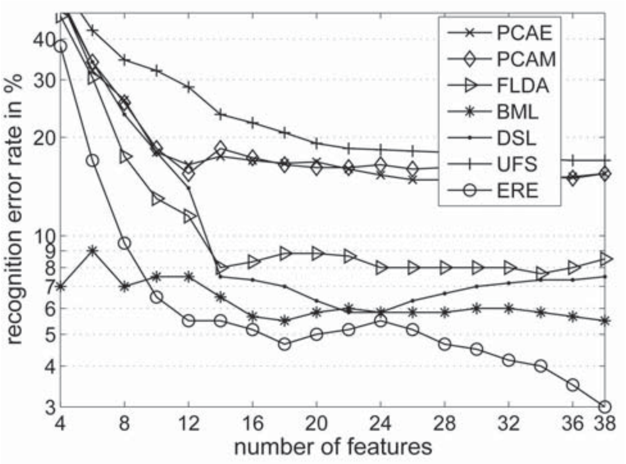

Fig. 6. Recognition error rate against the number of features used in the matching on the ORL database of 200 training images (40 people) and 200 testing images (40 people).

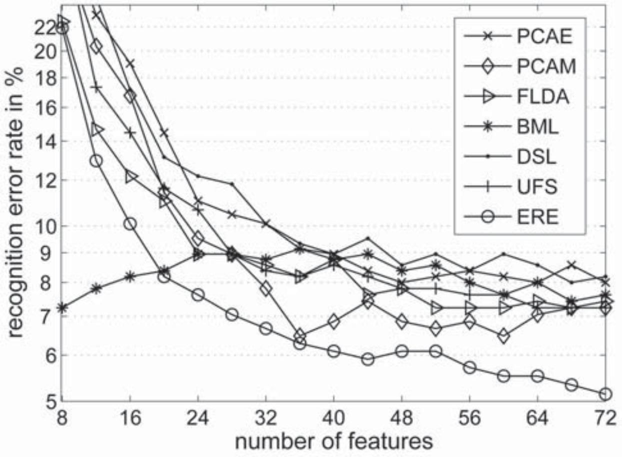

Fig. 7. Recognition error rate against the number of features used in the matching on the AR database of 525 training images (75 people) and 525 testing images (75 people).

person in the gallery set. As the large image size produces a large null space, the variation of the testing images in this large null space results in a higher recognition error of BML that uses the whole null space than FLDA that discards the null space. Similar to the first two experiments, the proposed ERE approach achieves the lowest recognition error rate consistently at all points in Fig. 5.

### 5.3 Results on ORL Database

Images of the ORL database [39] are cropped into the size of $50 \times 57$. The ORL database contains 400 images of 40 people (10 images per person). We first test various approaches using the first five samples per person for training and the remaining five samples per person for testing. Hence, there are 200 images in the training set and 200 images in the testing set. Fig. 6 shows the recognition error rate on the testing set against the number of features.

As the training set has only 200 images, it does not well represent the variations of the testing images. Therefore, the small principal space does not capture the discriminative information well. This results in poor performance of FLDA. UFS discards more dimensions and, hence, performs worse than FLDA. The DSL that extracts features in two complementary subspaces is better than FLDA. The BML that works in the whole space is better than DSL. For this small database, the proposed ERE approach also consistently outperforms other approaches.

As the ORL database is small, we conduct another experiment with leave-one-out training and testing strategy. In each of the 400 runs of training and testing, one sample is picked out for testing, and the remaining 399 samples are included in the training set. The testing results are numerically recorded in Table 1. As all images but one are used in the training, a small principal subspace on the training data can well represent the testing images. This results in better performances of the FLDA and the UFS, which only use the principal subspace, than the DSL and BML, which use both the principal and its complementary subspaces. However, the proposed ERE approach that works on the entire space outperforms FLDA and UFS. It shows that, for this training task, the complementary subspace is still useful but not well handled by the DSL and BML algorithms.

### 5.4 Results on AR Database

The color images in AR database [40] are converted to gray scale and cropped into the size of $120 \times 170$, same as the image size used in [40], [41]. There were 50 subjects with 12 images of frontal illumination per subject used in [40], and the same amount of subjects with 14 nonoccluded images per subject were used in [22]. In our experiment, 75 subjects with 14 nonoccluded images per subject are selected from the AR database. The first seven images of all subjects are used in the training, and the remaining seven images serve as testing images. For this large image size, we first apply PCA to remove the null space of $\mathbf{S}^t$ and then apply the ERE approach on the 524-dimensional feature vectors. The training time of the ERE approach by a Matlab program is about 1.4 times of that of the FLDA approach. Fig. 7 shows the recognition error rate.

As the images of the AR database were taken under tightly controlled conditions of illumination and viewpoint [40], the training set seems to represent the test set very well. FLDA, UFS, and PCAM that only use the principal subspace slightly outperform BML and DSL that use both the principal and its complementary subspaces. Furthermore, PCAM that uses leading eigenvectors of $\mathbf{S}^t$ surprisingly outperforms FLDA, DSL, and UFS. It shows that the problems of small eigenvalues of $\mathbf{S}^w$ dominate other factors for this tightly controlled database. Although the proposed ERE approach evaluates the discriminant value in the whole space of $\mathbf{S}^w$, it consistently outperforms all other approaches at all points in Fig. 7.

### 5.5 Results on Georgia Tech (GT) Database

The Georgia Tech (GT) Face Database [42] consists 750 color images of 50 subjects (15 images per subject). These images have large variations in both pose and expression and some illumination changes. Images are converted to gray scale and cropped into the size of $92 \times 112$. The first eight images of all subjects are used in the training and the remaining seven images serve as testing images. The testing results are numerically recorded in Table 2.

For this database that has large variations in both pose and expression and some illumination changes, Table 2 shows that the FLDA, BML, NLDA, SRLDA, DSL, and UFS methods have similar performances, which are significantly better than PCAE and PCAE. The proposed ERE approaches,

TABLE 2
Recognition Error Rate of Different Approaches for Different Number of Features on the GT Database

| Database | 400 training images and 350 testing images of Georgia Tech Database | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| # Feature | 8 | 12 | 16 | 20 | 24 | 28 | 32 | 36 | 40 | 44 | 48 |
| PCAE | 53.71 | 44.57 | 37.43 | 34.57 | 30.00 | 29.14 | 27.71 | 26.00 | 26.86 | 26.86 | 26.57 |
| PCAM | 48.57 | 38.57 | 31.43 | 26.57 | 26.86 | 22.29 | 20.57 | 20.00 | 19.71 | 21.43 | 19.43 |
| FLDA | 20.29 | 16.86 | 15.43 | 14.00 | 12.57 | 10.29 | 10.29 | 9.43 | 9.43 | 9.57 | 9.29 |
| BML | 13.71 | 14.57 | 12.57 | 14.29 | 13.14 | 13.43 | 13.14 | 12.57 | 12.86 | 12.86 | 14.00 |
| DSL | 29.43 | 22.57 | 16.00 | 14.00 | 14.00 | 11.71 | 10.57 | 11.71 | 10.86 | 10.00 | 9.43 |
| UFS | 21.71 | 15.71 | 14.29 | 12.00 | 11.71 | 10.86 | 10.00 | 9.43 | 9.71 | 9.14 | 9.14 |
| NLDA | 20.57 | 13.14 | 12.57 | 12.86 | 11.43 | 11.14 | 11.43 | 11.71 | 12.57 | 12.86 | 11.71 |
| SRLDA | 19.14 | 11.71 | 11.43 | 11.71 | 11.71 | 10.29 | 11.14 | 10.57 | 11.43 | 11.14 | 9.71 |
| ERE_$\tilde{\mathbf{S}}^b$ | 17.14 | 10.57 | 8.57 | 7.14 | 7.43 | 8.26 | 8.29 | 7.43 | 8.00 | 7.71 | 7.71 |
| ERE_$\tilde{\mathbf{S}}^t$ | 17.14 | 10.57 | 8.29 | 6.86 | 7.43 | 8.57 | 8.57 | 7.71 | 8.29 | 7.71 | 7.43 |

ERE_$\tilde{\mathbf{S}}^b$ and ERE_$\tilde{\mathbf{S}}^t$ both outperform all other approaches consistently in all columns in Table 2.

### 5.6 Significance of the Proposed Approach

We have performed seven sets of experiments with five different databases. The proposed ERE approach shows superior performance in the following five aspects: First, ERE consistently outperforms all other approaches in all the seven experiments, whereas no other approach can perform the second best consistently in all experiments. Second, ERE outperforms all other approaches consistently for every number of features tested in the experiments (note that BML uses $n$ rather than $d$ features in the matching process). In contrast to that, no other approach can perform the second best at all points even in a single experiment. Third, the ERE achieves the best performance in all experiments without tuning its parameter. ERE has only one free parameter $\mu$ in (8). Choosing some proper values of $\mu$ in different experiments will further enhance its recognition performances. Fourth, although ERE outperforms certain other approaches only marginally for a certain number of features in certain experiments, significant better performances of the ERE approach comparing to these approaches can always be found in at least three other experiments. Fifth, ERE significantly outperforms all other approaches for small number of features. This demonstrates that the proposed ERE approach extracts more discriminative features than others.

## 6 Conclusion

This paper addresses problems of extracting discriminant eigenfeatures from the face image based on a set of training samples. Noise disturbance and finite number of training samples in practice may cause eigenvalues of a scatter matrix deviating from the true variances of the images projected on the corresponding eigenvectors. This deviation may result in recognition performance deterioration for various subspace approaches that scale the eigenfeature by the inverse of the square root of the eigenvalue. Especially for small eigenvalues, their inverses are highly sensitive to the noise disturbance and the effect of the finite number of training samples. A different training set may easily change these values substantially. Therefore, for a good generalization, these small and zero eigenvalues should be regularized. Another problem addressed in this paper is the null space of the scatter matrix. Both the principal and the null spaces of the within-class scatter matrix contain discriminative information. Neither of them should be simply discarded in the feature extraction process. An optimal feature vector may reside in dimensions that lies in both subspaces, in other words, it may have nonzero components in both subspaces. It is inefficient or only suboptimal to construct a feature vector by extracting features separately from the principal and the null spaces. The discriminant evaluation in the whole space leads to a more efficient feature representation of a face image.

In this work, eigenspace spanned by the eigenvectors of the within-class scatter matrix is decomposed into a reliable subspace, an unstable subspace and a null subspace. Eigenfeatures are regularized differently in these three subspaces based on an eigenspectrum model to alleviate problems of instability, overfitting, or poor generalization. The discriminant evaluation is performed in the whole space and the feature extraction or dimensionality reduction is done only at the final stage after the discriminant assessment. This facilitates a discriminative and stable low-dimensional feature representation of the face image. Extensive experiments on the FERET, ORL, AR, and GT databases with different numbers of training samples demonstrate that the proposed approach consistently outperforms Eigenface, FLDA, BML, null space LDA, dual-space LDA, and unified subspace framework. Especially, it achieves more accuracy gains for a smaller number of features and for a smaller size of training set. This verifies that the proposed approach is more efficient in the feature extraction and more stable, less overfitting the training data or better generalization.

## References

[1] W. Zhao, R. Chellappa, P.J. Phillips, and A. Rosenfeld, "Face Recognition: A Literature Survey," *ACM Computing Surveys,* vol. 35, no. 4, pp. 399-458, Dec. 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," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 947-954, June 2005.

[3] M. Kirby and L. Sirovich, "Application of Karhunen-Loeve Procedure for the Characterization of Human Faces," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 12, no. 1, pp. 103-108, Jan. 1990.

[4] B. Moghaddam and A. Pentland, "Probabilistic Visual Learning for Object Representation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 696-710, July 1997.

[5] B. Moghaddam, T. Jebara, and A. Pentland, "Bayesian Face Recognition," *Pattern Recognition,* vol. 33, no. 11, pp. 1771-1782, Nov. 2000.

[6] B. Moghaddam, "Principal Manifolds and Probabilistic Subspace for Visual Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 6, pp. 780-788, June 2002.

[7] D.L. Swets and J. Weng, "Using Discriminant Eigenfeatures for Image Retrieval," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, no. 8, pp. 831-836, Aug. 1996.

[8] C. Liu and H. Wechsler, "Enhanced Fisher Linear Discriminant Models for Face Recognition," *Proc. Int'l Conf. Pattern Recognition,* vol. 2, pp. 1368-1372, Aug. 1998.

[9] R. Basri and D. Jacobs, "Lambertian Reflectance and Linear Subspaces," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 2, pp. 218-233, Feb. 2003.

[10] T.K. Kim, H. Kim, W. Hwang, and J. Kittler, "Independent Component Analysis in a Local Facial Residue Space for Face Recognition," *Pattern Recognition,* vol. 37, pp. 1873-1885, Sept. 2004.

[11] J. Kim, J. Choi, J. Yi, and M. Turk, "Effective Representation Using ICA for Face Recognition Robust to Local Distortion and Partial Occlusion," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 12, pp. 1977-1981, Dec. 2005.

[12] Q. Liu, H. Lu, and S. Ma, "Improving Kernel Fisher Discriminant Analysis for Face Recognition," *IEEE Trans. Circuits and Systems for Video Technology,* vol. 14, no. 1, pp. 42-49, Jan. 2004.

[13] W.S. Chen, P.C. Yuen, J. Huang, and D.Q. Dai, "Kernel Machine-Based One-Parameter Regularized Fisher Discriminant Method for Face Recognition," *IEEE Trans. Systems, Man, and Cybernetics Part B,* vol. 35, no. 4, pp. 659-669, Aug. 2005.

[14] 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 Analysis and Machine Intelligence,* vol. 27, no. 2, pp. 230-244, Feb. 2005.

[15] C. Liu and H. Wechsler, "Gabor Feature Based Classification Using the Enhanced Fisher Linear Discriminant Analysis for Face Recognition," *IEEE Trans. Image Processing,* vol. 11, no. 4, pp. 467-476, Apr. 2002.

[16] C. Liu, "Gabor-Based Kernel PCA with Fractional Power Polynomial Models for Face Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, no. 5, pp. 572-581, May 2004.

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

[18] X. He, S. Yan, Y. Hu, P. Niyogi, and H.J. Zhang, "Face Recognition Using Laplacianfaces," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 3, pp. 328-340, Mar. 2005.

[19] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 711-720, July 1997.

[20] D.Q. Dai and P.C. Yuen, "Regularized Discriminant Analysis and Its Application to Face Recognition," *Pattern Recognition,* vol. 36, no. 3, pp. 845-847, Mar. 2003.

[21] X. Wang and X. Tang, "Dual-Space Linear Discriminant Analysis for Face Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 564-569, June 2004.

[22] H. Cevikalp, M. Neamtu, M. Wilkes, and A. Barkana, "Discriminative Common Vectors for Face Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 1, pp. 4-13, Jan. 2005.

[23] H. Yu and J. Yang, "A Direct LDA Algorithm for High-Dimensional Data with Application to Face Recognition," *Pattern Recognition,* vol. 34, no. 10, pp. 2067-2070, Oct. 2001.

[24] J. Lu, K.N. Plataniotis, and A.N. Venetsanopoulos, "Face Recognition Using LDA-Based Algorithms," *IEEE Trans. Neural Networks,* vol. 14, no. 1, pp. 195-200, Jan. 2003.

[25] W. Liu, Y. Wang, S.Z. Li, and T.N. Tan, "Null Space Approach of Fisher Discriminant Analysis for Face Recognition," *Proc. ECCV Workshop Biometric Authentication,* pp. 32-44, May 2004.

[26] L.F. Chen, H.Y.M. Liao, M.T. Ko, J.C. Lin, and G.J. Yu, "A New LDA-Based Face Recognition System Which Can Solve the Small Sample Size Problem," *Pattern Recognition,* vol. 33, no. 10, pp. 1713-1726, Oct. 2000.

[27] K. Liu, Y.Q. Cheng, J.Y. Yang, and X. Liu, "An Efficient Algorithm for Foley-Sammon Optical Set of Discriminant Vectors by Algebraic Method," *Int'l J. Pattern Recognition Artificial Intelligence,* vol. 6, pp. 817-829, 1992.

[28] J. Yang and J.Y. Yang, "Why Can LDA Be Performed in PCA Transformed Space?" *Pattern Recognition,* vol. 36, pp. 563-566, 2003.

[29] X. Wang and X. Tang, "A Unified Framework for Subspace Face Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, no. 9, pp. 1222-1228, Sept. 2004.

[30] C. Liu and H. Wechsler, "Robust Coding Schemes for Indexing and Retrieval from Large Face Database," *IEEE Trans. Image Processing,* vol. 9, no. 1, pp. 132-137, Jan. 2000.

[31] A.M. Martinez and A.C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, no. 2, pp. 228-233, Feb. 2001.

[32] J.H. Friedman, "Regularized Discriminant Analysis," *J. Am. Statistical Assoc.,* vol. 84, no. 405, pp. 165-175, Mar. 1989.

[33] J. Lu, K.N. Plataniotis, and A.N. Venetsanopoulos, "Regularized Discriminant Analysis for the Small Sample Size Problem in Face Recognition," *Pattern Recognition Letters,* vol. 24, pp. 3079-3087, 2003.

[34] J. Lu, K.N. Plataniotis, and A.N. Venetsanopoulos, "Regularization Studies of Linear Discriminant Analysis in Small Sample Size Scenarios with Application to Face Recognition," *Pattern Recognition Letters,* vol. 26, pp. 181-191, 2005.

[35] X.D. Jiang, B. Mandal, and A. Kot, "Enhanced Maximum Likelihood Face Recognition," *Electronics Letters,* vol. 42, no. 19, pp. 1089-1090, Sept. 2006.

[36] R. Beveridge, D. Bolme, M. Teixeira, and B. Draper, "The CSU Face Identification Evaluation System User's Guide: Version 5.0," technical report, http://www.cs.colostate.edu/evalfacerec/data/normalization.html, 2003.

[37] P.J. Phillips, H. Moon, S. Rizvi, and P. Rauss, "The FERET Evaluation Methodology for Face Recognition Algorithms," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 10, pp. 1090-1104, Oct. 2000.

[38] J. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, and S.Z. Li, "Ensemble-Based Discriminant Learning with Boosting for Face Recognition," *IEEE Trans. Neural Networks,* vol. 17, no. 1, pp. 166-178, Jan. 2006.

[39] F. Samaria and A. Harter, "Parameterization of a Stochastic Model for Human Face Identification," *Proc. Second IEEE Workshop Applications of Computer Vision,* pp. 138-142, Dec. 1994.

[40] A.M. Martinez, "Recognizing Imprecisely Localized, Partially Occluded, and Expression Variant Faces from a Single Sample per Class," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 6, pp. 748-763, June 2002.

[41] 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.

[42] *Georgia Tech Face Database,* http://www.anefian.com/face_reco.htm, 2007.

**Xudong Jiang** received the BEng and MEng degrees from the University of Electronic Science and Technology of China (UESTC) in 1983 and 1986, respectively, and the PhD degree from the University of German Federal Armed Forces, Hamburg, Germany, in 1997, all in electrical and electronic engineering. From 1986 to 1993, he was a lecturer at UESTC, where he received two Science and Technology Awards from the Ministry for Electronic Industry of China. From 1993 to 1997, he was with the University of German Federal Armed Forces, as a scientific assistant. From 1998 to 2002, he was with Nanyang Technological University (NTU), Singapore, as a senior research fellow, where he developed a fingerprint verification algorithm that achieved the most efficient and the second most accurate fingerprint verification at the International Fingerprint Verification Competition (FVC '00). From 2002 to 2004, he was a lead scientist and the head of the Biometrics Laboratory at the Institute for Infocomm Research, Singapore. He joined NTU as a faculty member in 2004. Currently, he serves as the director of the Centre for Information Security, the School of Electrical and Electronic Engineering, NTU, Singapore. His research interest includes pattern recognition, signal and image processing, computer vision, and biometrics. He is a senior member of the IEEE.

**Bappaditya Mandal** received the BTech degree in electrical engineering from the Indian Institute of Technology, Roorkee, India, in 2003. He has more than one year (2003-2004) of work experience in software engineering. Since 2004, he has been a PhD candidate in the Department of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include biometrics, pattern recognition, computer vision, and machine learning. He is working on feature extraction using subspace methods for face identification and verification. In 2001, he received the summer undergraduate research award (SURA) from the Indian Institute of Technology.

**Alex Kot** received the BSEE and MBA degrees from the University of Rochester, New York, and the PhD degree from the University of Rhode Island, Kingston. He was with AT&T briefly. Since 1991, he has been with Nanyang Technological University, Singapore. He headed the Division of Information Engineering at the School of Electrical and Electronic Engineering for eight years and is currently a professor and the vice dean (Research) of the School of Electrical and Electronic Engineering. He has published extensively in the areas of signal processing for communication, biometrics, data hiding, and authentication. He has served as an associate editor for *IEEE Transactions on Signal Processing*, *IEEE Transactions on Circuits and Systems II*, and *IEEE Transactions on Circuits and Systems for Video Technology*. He has served as a guest editor for IEEE and EURASIP journals. Currently, he is an associate editor for the *IEEE Transactions on Circuits and Systems I* and *EURASIP Journal on Applied Signal Processing*. He has served on numerous conference committees and technical committees, including cochairing the IEEE International Conference on Image Processing (ICIP) in 2004 and an IEEE distinguished lecturer (2005-2006). He is a fellow of the IEEE.