Pattern Recognition Letters 31 (2010) 717–724

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Prediction of eigenvalues and regularization of eigenfeatures for human face verification ☆

Bappaditya Mandal [a,*], Xudong Jiang [b], How-Lung Eng [a], Alex Kot [b]

[a] 1 Fusionopolis Way, #21-01 Connexis (South Tower), Institute for Infocomm Research, Singapore 138632, Singapore
[b] Electrical and Electronic Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, Singapore

ARTICLE INFO

Article history:
Available online 20 October 2009

Keywords:
PCA
LDA
Feature extraction
Dimensionality reduction
Face verification

ABSTRACT

We present a prediction and regularization strategy for alleviating the conventional problems of LDA and its variants. A procedure is proposed for predicting eigenvalues using few reliable eigenvalues from the range space. Entire eigenspectrum is divided using two control points, however, the effective low-dimensional discriminative vectors are extracted from the whole eigenspace. The estimated eigenvalues are used for regularization of eigenfeatures in the eigenspace. These prediction and regularization enable to perform discriminant evaluation in the full eigenspace. The proposed method is evaluated and compared with eight popular subspace based methods for face verification task. Experimental results on popular face databases show that our method consistently outperforms others.

1. Introduction

Human beings are experts in verifying subject's identity just by analyzing face images (photographs). This ability is very appealing and has become an active research area in many real-life machine vision applications. Face verification (FV) is an important tool for authentication of an individual and has significant role in many security and e-commerce applications (Zhao et al., 2003).

Face verification and identification are the main applications of face recognition (FR). A face verification system has to discriminate between two kinds of events: either the person claiming a given identity is the true claimant or the person is an impostor. In recent years, many subspace based approaches like PCA and LDA are being applied to FR problem (Zhao et al., 2003). The results are not satisfactory because PCA does not encode the class information and LDA suffers from instability of eigenvalue decomposition due to the small number of training samples and high dimensionality of face images. Moreover, in Fisherfaces (FLDA) (Belhumeur et al., 1997), the singularity of the scatter matrices are not guaranteed (Zhuang and Dai, 2007).

In recent times, many researchers have noticed these problems and tried to solve them using different methods. Bayesian maximum likelihood (BML) is proposed in (Moghaddam et al., 2000; Moghaddam and Pentland, 1997). It uses a probabilistic similarity measure based on the Bayesian belief that the image intensity differences are characteristic of typical variations in appearance of an individual. Their similarity measure is expressed in terms of probability using the two class of facial image variations: intrapersonal variations and extrapersonal variations. Although this method performs good for FR task, one need to store the original face image of an individual in the database, which are in general, of very large dimensionality. Moreover, the computation of their distance measure has very high time complexity as it involves both distance-in-feature space and distance-from-feature space (Moghaddam et al., 1998, 2000; Jiang et al., 2006).

To deliver promising FR results, recently a myriad of algorithms based on the applications of PCA and FLDA are proposed in the existing literature (Zhao et al., 2003; Shakhnarovich and Moghaddam, 2005; Stan et al., 2004). Direct LDA (DLDA) (Yu et al., 2001) approach removes null space of the between-class scatter matrix and extracts the eigenvectors corresponding to the smallest eigenvalues of the within-class scatter matrix. However, an argument against the DLDA algorithm is presented in (Gao and Davis, 2006), where they have shown that DLDA is actually a special case of LDA by directly taking the linear space of class means as the LDA solution. The pooled covariance estimate is completely ignored. They also demonstrate that DLDA is not equivalent to traditional LDA in dealing with the small sample size problem and may impose performance limitations in general application (Gao and Davis, 2006).

Null space LDA (NDA) approach is proposed in (Liu et al., 2004; Huang et al., 2002). They have shown that the null space of the to-

☆ This paper is based on the best biometrics student paper award, received at the 19th International Conference on Pattern Recognition (ICPR), Tampa, FL, December 10, 2008.

* Corresponding author. Tel.: +65 6408 2071; fax: +65 6776 1378.
E-mail addresses: bmandal@i2r.a-star.edu.sg, bapp0001@ntu.edu.sg (B. Mandal), exdjiang@ntu.edu.sg (X. Jiang), hleng@i2r.a-star.edu.sg (H.-L. Eng), eackot@ntu.edu.sg (A. Kot).

0167-8655/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2009.10.006

tal scatter matrix is the common null space of both within-class and between-class matrices. The algorithm firstly removes the null space of the total scatter matrix and projects the samples onto the null space of within-class scatter matrix. It then removes the null space of the between-class scatter matrix in the subspace to obtain the optimal discriminant vectors. The basic notion of this algorithm is that the null space of the within-class scatter matrix is particularly useful in discriminating ability. Interestingly, this appears to be contradicting the popular FLDA that uses only the principal space and discards the null space. A common problem to all these approaches is that they all lose some discriminative information, either in the principal or in the null space.

To take advantages from both the subspaces, dual space LDA (DSL) is proposed in (Wang and Tang, 2004a). Using the probabilistic visual model (Moghaddam and Pentland, 1997), the eigenvalue spectrum in the null space of the within-class scatter matrix is estimated. It performs discriminant analysis in both the subspaces and the discriminative features are then combined in recognition phase. The features in the complementary subspace are scaled by the average eigenvalue of the within-class scatter matrix over this subspace. As eigenvalues in this subspace are not well estimated (Wang and Tang, 2004a), 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 (Yang et al., 2005). 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.

Another popular approach called unified framework of subspaces (UFS) (Wang et al., 2004b), addresses the problems of instability and noise disturbances in LDA based methods. Using this framework they demonstrate the importance of noise suppression. This approach applies three stages of subspace decompositions sequentially on the face training data and the dimensionality reduction occurs at the very first stage. However, as addressed in the literature (Jiang et al., 2007; Cevikalp et al., 2005; Wang and Tang, 2004a), 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 (Wang et al., 2004b) show that the recognition performance is sensitive to these choices at different stages.

In this paper, we revisit the short comings of FLDA approach for FV task and related ideas proposed in (Mandal et al., 2008). FLDA has instability problem due to the limited number of training samples and high dimensionality of face images. Moreover, it loses important discrimination information in the range and/or null space. To alleviate these problems, we propose to partition the entire eigenspace into reliable, unreliable and null regions using two control points. A procedure for eigenvalue prediction is proposed. The forecasted eigenvalues are used for regularization of eigenfeatures in the eigenspace. These prediction and regularization enable to perform discriminant evaluation in the full eigenspace and extract effective low-dimensional discriminative features from face images. We evaluate and compare our approach with eight other popular subspace based methods for the FV task.

In the following section, we present the partitioning of subspaces and eigenspectrum modeling. In Section 3, we discuss the eigenfeature scaling and extraction procedures. Experimental results and discussions are presented in Section 4. Finally conclusions are drawn in Section 5.

2. Partitioning of subspaces and eigenspectrum modeling

Given a set of properly normalized h-by-w face images, we can form a training set of column vectors $\{X_{ij}\}$, where $X_{ij} \in \mathbb{R}^{n=hw}$ is an image column vector, 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 total number of training samples 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

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

where $\overline{X}_i = \frac{1}{q_i} \sum_{j=1}^{q_i} X_{ij}$. The between-class scatter matrix is defined by

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

where $\overline{X} = \sum_{i=1}^{p} c_i \overline{X}_i$. If all classes have equal prior probability, then $c_i = 1/p$. The total class scatter matrix is defined by

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

In the well-known Fisher objective criteria (Duda et al., 2001), if $\boldsymbol{S}^w$ is nonsingular, the optimal projection vectors $\boldsymbol{\Phi}$ is chosen as the matrix with orthonormal columns which maximizes the ratio of the determinant of the between-class matrix of the projected samples to the determinant of the within-class scatter of the projected samples. The columns of the solution matrix are eigenvectors of matrix corresponding to its greatest eigenvalues:

$$(\boldsymbol{S}^{w^{-1}} \boldsymbol{S}^b) \boldsymbol{\Phi}^{\text{opt}} = \boldsymbol{\Phi}^{opt} \boldsymbol{\Lambda}, \tag{4}$$

which also implies:

$$\boldsymbol{\Phi}^{\text{opt}} = \arg\max_{\boldsymbol{\Phi}} \frac{|\boldsymbol{\Phi}^T \boldsymbol{S}^b \boldsymbol{\Phi}|}{|\boldsymbol{\Phi}^T \boldsymbol{S}^w \boldsymbol{\Phi}|}, \tag{5}$$

where $\boldsymbol{\Lambda}$ is the diagonal eigenvalue matrix and $\boldsymbol{\Phi}$ is the eigenvector matrix.

Let $\boldsymbol{S}^g, g \in \{t, w, b\}$ 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 de-correlated by solving the eigenvalue problem

$$\boldsymbol{\Lambda}^g = \boldsymbol{\Phi}^{gT} \boldsymbol{S}^g \boldsymbol{\Phi}^g, \tag{6}$$

where $\boldsymbol{\Phi}^g = [\phi_1^g, \ldots, \phi_n^g]$ is the eigenvector matrix of $\boldsymbol{S}^g$, and $\boldsymbol{\Lambda}^g$ is the diagonal matrix of eigenvalues $\lambda_1^g, \ldots, \lambda_n^g$ corresponding to the eigenvectors. Suppose that the eigenvalues are sorted in descending order $\lambda_1^g \geqslant, \ldots, \geqslant \lambda_n^g$. 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 subspace methods as the eigenvalues are used to scale and extract features. In the following section we discuss the problems of feature scaling in detail.

2.1. Problems in feature scaling

The result of the discriminant evaluation (Fisher criteria) in (5) cannot be directly adopted to the FR area as the $\boldsymbol{S}^w$ is often singular because of the limited number of training samples, noises and high dimensionality of the face images. Moreover, the large number of small eigenvalues that arises in the range space after eigen-decomposition of the $\boldsymbol{S}^w$ matrix give undue weightage in the feature scaling process.

To demonstrate these problems, we first perform eigen-decomposition of the $\boldsymbol{S}^w$ matrix computed from the original face samples images. Let $\boldsymbol{\Phi}^w = [\phi_1^w, \ldots, \phi_n^w]$ be the eigenvector matrix of $\boldsymbol{S}^w$, and

Λ^w be the diagonal matrix of eigenvalues $\lambda_1^w, \ldots, \lambda_n^w$ corresponding to the eigenvectors. We assume that the eigenvalues are sorted in descending order $\lambda_1^w \geqslant, \ldots, \geqslant \lambda_n^w$. The plot of eigenvalues λ_k^w against the index k is called eigenspectrum. A typical plot of $\sigma_k^w = \sqrt{\lambda_k^w}$ is shown in Fig. 1 (for simplicity, we still call it eigenspectrum). It plays a critical role in the subspace methods as the eigenvalues are used to scale and extract features. These eigenvectors then undergo a whitening process (Fukunaga, 1991). The purposes of this transformation are to change the scales of the eigenvectors in proportion to $\frac{1}{\sqrt{\lambda_k^w}}$ and also make the within-class covariance matrix invariant to any further orthonormal transformation. This property will be used for simultaneous diagonalization of $\boldsymbol{S}^w$ and $\boldsymbol{S}^b$ matrices for evaluating the Fisher criteria (5). A two-dimensional example is shown in (Fukunaga, 1991).

However, before performing this whitening step, we first analyze the reliability of the projection (eigen)vectors corresponding to all the eigenvalues of $\boldsymbol{S}^w$ matrix. After computing the $\boldsymbol{\Phi}^w = [\phi_1^w, \ldots, \phi_n^w]$ matrix from the training data, we project face images from a test dataset onto these projection vectors (using $Y_{ij} = \boldsymbol{\Phi}^{wT} X_{ij}$), and then compute the within-class variance across all the projected test data for all dimensions (indices). Let $\upsilon_k^w = \sqrt{\chi_k^w}$, where χ_k^w represents within-class variance arising from the projected test data, is shown in Fig. 1. It is evident from Fig. 1 that there is a large deviation of the small eigenvalues from the variances of novel images projected on the eigenvectors. Other datasets of training and testing face images produce results similar to Fig. 1. An estimate of the training and test datasets are given in Section 4 – experimental results and discussion.

This problem is well addressed in (Jiang et al., 2008) and recently in (Jiang, 2009). Although the largest sample-based eigenvalues are biased high and the smallest ones are biased low, as pointed out in (Friedman, 1989), the bias is most pronounced when the population eigenvalues tend toward equality, and it is correspondingly less severe when their values are highly disparate. In FR application, eigenspectrum often first decays very rapidly and then stabilizes. Therefore, the smallest eigenvalues are biased much more than the largest ones. This is evidenced by Fig. 1.

The whitened eigenvector matrix $\overline{\boldsymbol{\Phi}}^w = [\phi_1^w/\sigma_1^w, \ldots, \phi_n^w/\sigma_n^w]$, $\sigma_k^w = \sqrt{\lambda_k^w}$ as shown in Fig. 2, is used to project the image vector X_{ij} before constructing the between-class scatter matrix for the second eigen-decomposition. This is equivalent to image vector X_{ij} is first transformed by eigenvector, $Y_{ij} = \boldsymbol{\Phi}^{wT} X_{ij}$, and then multiplied

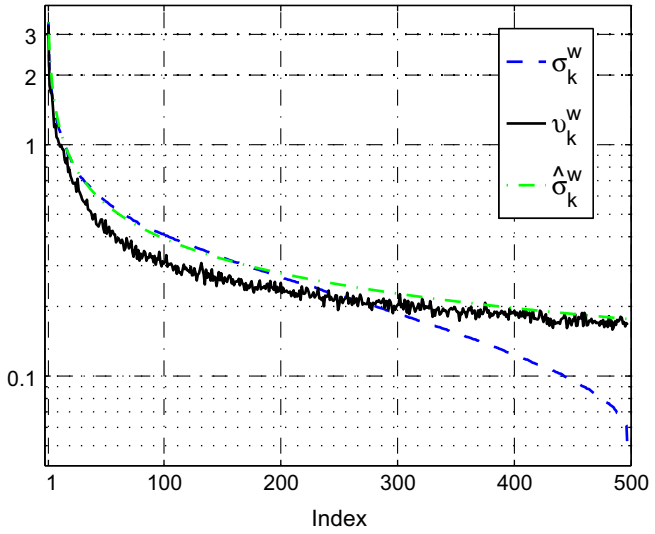

Fig. 1. Eigen-spectra of training (σ_k^w), modeled ($\hat{\sigma}_k^w$) and testing (υ_k^w) datasets.

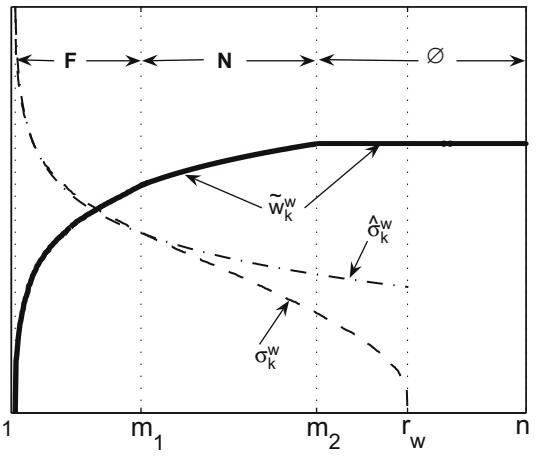

Fig. 2. Real eigenspectrum and weighting function (16).

by a weighting function $w_k^w = 1/\sqrt{\lambda_k^w}$ (whitening process). Discarding dimensions that have zero eigenvalues is equivalent to set $w_k^w = 0$ for these dimensions (as done in Fisherfaces (Belhumeur et al., 1997)). The weighting function is thus

$$w_k^w = \begin{cases} 1/\sqrt{\lambda_k^w}, & k \leqslant r_w \\ 0, & r_w < k \leqslant n \end{cases}, \tag{7}$$

where r_w is the rank of $\boldsymbol{S}^w$. Fig. 2 shows a typical real eigenspectrum.

There are two problems associated with the scaling function in (7). Firstly, the eigenvectors corresponding to the zero eigenvalues are lost or discarded as the features in the null space are weighted by a constant zero. This leads to the lose of important discriminative information that lies in the null space (Liu et al., 2004; Huang et al., 2002; Xu et al., 2008). Secondly, using the inverse of the square root of the eigenvalue (7) to weight the eigenfeature amplifies noise and tends to over-fit the training samples. The small and zero eigenvalues are training-set-specific 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. Hence, they are unreliable. In the following subsections, we first discuss procedures for estimating two index (control) points, named as m_1 and m_2, using which eigenspectrum is partitioned into three subspaces and then present a methodology for predicting the eigenvalues for replacing the unreliable ones.

2.2. Estimation of first control point m_1 for subspace partitioning

We propose to decompose the whole eigenspace spanned by eigenvectors $\mathbb{R}^n$ into three subspaces: a reliable face variation dominating subspace (or simply face space) $\boldsymbol{F} = \{\phi_k^w\}_{k=1}^{m_1}$, an unreliable noise variation dominating subspace (or simply noise space) $\boldsymbol{N} = \{\phi_k^w\}_{k=m_1+1}^{m_2}$ and a null space $\boldsymbol{\emptyset} = \{\phi_k^w\}_{k=m_2+1}^{n}$ as illustrated in Fig. 2. The purpose of this decomposition is to modify the unreliable eigenvalues for better generalization. The rank of $\boldsymbol{S}^w$ is $r_w \leqslant min(n, l-p)$. As face images have similar structure, significant face components reside intrinsically in a very low-dimensional (m_1-dimensional) subspace. 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

$$\lambda_{\text{med}}^w = \text{median}\{\forall \lambda_k^w | k \leqslant r_w\}. \tag{8}$$

The distance between λ^w_{med} and the smallest nonzero eigenvalue is $d_{m_1,r_w} = \lambda^w_{\text{med}} - \lambda^w_{r_w}$. The upper bound of the unreliable eigenvalues is estimated by $\lambda^w_{\text{med}} + d_{m_1,r_w}$. More generally, the start point of the noise region $m_1 + 1$ is estimated by

$$\lambda^w_{m_1+1} = \max\{\forall\lambda^w_k | \lambda^w_k < (\lambda^w_{\text{med}} + \mu(\lambda^w_{\text{med}} - \lambda^w_{r_w}))\}, \tag{9}$$

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.

2.3. Estimation of second control point m_2 for subspace partitioning

The phenomenon that the eigenspectrum accelerates its decrease is caused by the limited number of training samples and noises present in them (Jiang et al., 2009). To study this phenomenon, we define eigenratios as

$$\gamma^w_k = \frac{\lambda^w_{k+1}}{\lambda^w_k}, \quad 1 \leqslant k < r_w. \tag{10}$$

The plot of eigenratios γ^w_k against index k is called eigenratio-spectrum. Fig. 3 shows a typical eigenratio-spectrum of a real face training database. The eigenratios are, in general, very random in nature. To obtain a summarization of their behavior, we smoothen the eigenratios by using an average over a moving window. The original eigenratios and their smoothen values are shown in Fig. 3. We examined several different face databases, the eigenratio plots shown in Fig. 3 is a general behavioral pattern that all the eigenratios of different databases portray.

From the graph it is evident that the eigenratios first increases very rapidly, then stabilizes and finally decreases. The limited number of the training samples causes the decrease of the eigenratios. The corresponding eigenvalues are thus unreliable. Therefore, the start point of the unreliable region $m_2 + 1$ is estimated by

$$\gamma^w_{m_2+1} = \max\{\forall\gamma^w_k, \quad 1 \leqslant k < r_w\}. \tag{11}$$

A typical such m_2 value of a real eigenspectrum is shown in Fig. 3.

2.4. Prediction of eigenvalues

This work uses function form $1/f$ to fit only the reliable part of eigenspectrum $\{\lambda^w_k | 1 \leqslant k \leqslant m_1\}$ and then to extrapolate eigenvalues in the noise subspace $\{\lambda^w_k | m_1 < k \leqslant r_w\}$ (Jiang et al., 2008). Prediction of the eigenspectrum is performed by

$$\hat{\lambda}^w_k = \frac{\alpha}{k+\beta}, \quad 1 \leqslant k \leqslant r_w, \tag{12}$$

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 are determined by fitting the model to the real eigenspectrum in the reliable face space $\boldsymbol{F}$. We determine α and β by letting $\hat{\lambda}^w_1 = \lambda^w_1$ and $\hat{\lambda}^w_{m_1} = \lambda^w_{m_1}$, which yields

$$\alpha = \frac{\lambda^w_1 \lambda^w_{m_1}(m_1 - 1)}{\lambda^w_1 - \lambda^w_{m_1}}, \tag{13}$$

$$\beta = \frac{m_1 \lambda^w_{m_1} - \lambda^w_1}{\lambda^w_1 - \lambda^w_{m_1}}. \tag{14}$$

Since the eigenspectrum decays very fast, we plot the square roots $\sigma^w_k = \sqrt{\lambda^w_k}$ and $\hat{\sigma}^w_k = \sqrt{\hat{\lambda}^w_k}$ for clearer illustration (we still call them eigenspectrum for simplicity). A typical real eigenspectrum σ^w_k and its prediction $\hat{\sigma}^w_k$ are shown in Figs. 1 and 2.

From Fig. 1, firstly, we see that the predicted eigenvalues matches closely with that of the real eigenvalues. Secondly, for small eigenvalues it matches more closely with the variances obtained from the projected testing dataset as compared to that of the real eigenvalues. This shows that our prediction of the eigenvalues using few principal eigenvalues of $\boldsymbol{S}^w$ matrix provides more generalization to the unseen (test) data. In Fig. 2, we see that the predicted values $\hat{\sigma}^w_k$ fits closely to the real σ^w_k in the face space $\boldsymbol{F}$ but has slower decay in the noise space $\boldsymbol{N}$. The faster decay of the real eigenspectrum σ^w_k in $\boldsymbol{N}$ due to noise and the limited number of training samples is what we want to slow down (as shown in Figs. 1 and 2).

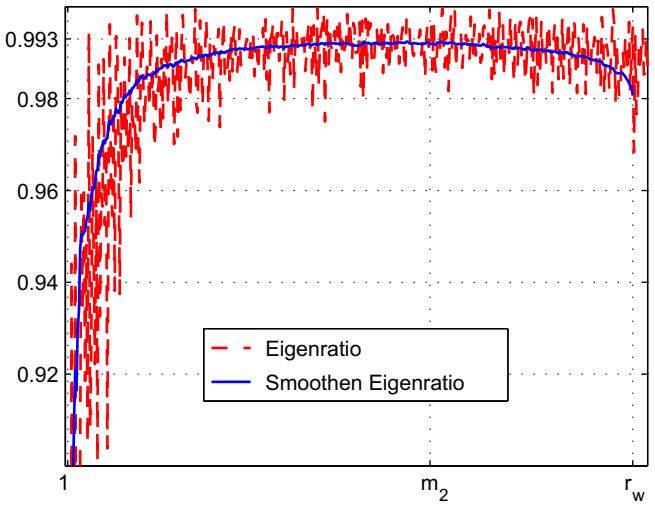

Fig. 3. Plot of eigenratios (10) and smoothened eigenratio-spectrum from a real eigenspectrum.

3. Eigenfeature scaling and extraction

The partitioning of the eigenspectrum has helped in identifying the face, noise and null regions. Eigenvalues are then forecasted using the few reliable eigenvalues from the range space. From Fig. 2 it is evident that noise component is small as compared to face components in $\boldsymbol{F}$ but it is dominating in region $\boldsymbol{N}$. Thus, the predicted eigenspectrum $\tilde{\lambda}^w_k$ is given by

$$\tilde{\lambda}^w_k = \begin{cases} \lambda^w_k, & k < m_1 \\ \frac{\alpha}{k+\beta}, & m_1 \leqslant k \leqslant m_2 \\ \frac{\alpha}{r_w+1+\beta}, & m_2 < k \leqslant n \end{cases}. \tag{15}$$

The proposed feature weighting function is then

$$\tilde{w}^w_k = \frac{1}{\sqrt{\tilde{\lambda}^w_k}}, \quad k = 1, 2, \ldots n. \tag{16}$$

Fig. 2 shows the proposed feature weighting function $\tilde{w}^w_k$ calculated by (9), (11), (13)–(16). Using this weighting function and the eigenvectors ϕ^w_k, training data are transformed to

$$\widetilde{Y}_{ij} = \widetilde{\boldsymbol{\Phi}}^{wT}_n X_{ij}, \tag{17}$$

where

$$\widetilde{\boldsymbol{\Phi}}^w_n = [\tilde{w}^w_k \phi^w_k]^n_{k=1} = [\tilde{w}^w_1 \phi^w_1, \ldots, \tilde{w}^w_n \phi^w_n], \tag{18}$$

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

A new between-class scatter matrix is formed by vectors $\widetilde{Y}_{ij}$ of the transformed training data as

$$\widetilde{\boldsymbol{S}}^b = \sum_{i=1}^{p} c_i (\overline{\widetilde{Y}}_i - \overline{Y})(\overline{\widetilde{Y}}_i - \overline{Y})^T, \tag{19}$$

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 $\widetilde{\boldsymbol{S}}^b$ by solving the eigenvalue problem (6). Suppose that the eigenvectors in the eigenvector matrix $\widetilde{\boldsymbol{\Phi}}_n^b = [\widetilde{\phi_1^b}, \ldots, \widetilde{\phi_n^b}]$ 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,

$$\widetilde{\boldsymbol{\Phi}}_d^b = [\widetilde{\phi_k^b}]_{k=1}^d = [\widetilde{\phi_1^b}, \ldots, \widetilde{\phi_d^b}], \tag{20}$$

where d is the number of features usually selected by a specific application. Thus, the proposed feature scaling and extraction matrix is given by

$$\boldsymbol{U} = \widetilde{\boldsymbol{\Phi}}_n^w \widetilde{\boldsymbol{\Phi}}_d^b, \tag{21}$$

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

$$F = \boldsymbol{U}^T X. \tag{22}$$

Below we summarize the proposed algorithm.

3.1. Proposed algorithm

The proposed approach of extracting discriminative vectors by applying predicted eigenvalues (DVPE) is summarized below:

At the training stage:

1. Given a training set of normalized face image vectors $\{X_{ij}\}$, estimate $\boldsymbol{S}^w$ by (1) and compute all its eigenvectors and eigenvalues using (6).
2. Estimate m_1 value using (8) and (9).
3. Estimate m_2 value using (10) and (11).
4. Decompose the eigenspace into face-, noise-, and null-space using m_1 and m_2 values.
5. Transform the training samples represented by X_{ij} into $\widetilde{Y}_{ij}$ using (17) with the weighting function (16) determined by (9), (11), (13)–(15).
6. Compute $\widetilde{\boldsymbol{S}}^b$ by (19) with $\widetilde{Y}_{ij}$ and solve the eigenvalue problem using (6).
7. Obtain the final feature scaling and extraction matrix by (18), (20) and (21) with a predefined number of features d.

At the enrollment or registration stage:

1. Extract d–dimensional feature vector F from the enrolled n–dimensional normalized face image vector X by (22) using the feature scaling and extraction matrix $\boldsymbol{U}$ obtained in the training stage (21).
2. Store the extracted feature vector and the registration ID into the gallery feature vector set.

At the verification stage:

1. Extract d–dimensional feature vector F from the n–dimensional normalized probe face image vector X by (22) using the feature scaling and extraction matrix $\boldsymbol{U}$ obtained in the training stage (21).
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, first nearest neighborhood classifier (1-NNK) is applied to test the proposed DVPE approach. Cosine distance measure between a probe feature vector F_P and a gallery feature vector F_G

$$dst(F_P, F_G) = -\frac{F_P^T F_G}{\|F_P\|_2 \|F_G\|_2} \tag{23}$$

is applied to the proposed approach, where $\|\cdot\|_2$ is the norm 2 operator.

4. Experimental results and discussions

AR, FERET database 1 and FERET database 2 are used in our experiments. In all the experiments reported in this work, images are preprocessed, aligned and normalized following the CSU Face Identification Evaluation System (Beveridge et al., 2003), which also employs FERET database. 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: correct verification rate (CVR) and false acceptance rate (FAR). FAR is the ratio of the number of accepted imposter matches to the total number of imposter matches. CVR is the rate at which legitimate end-users (subjects) are correctly verified. The plot of CVR 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 proposed DVPE method is tested and compared with eight other popular subspace based approaches: PCA with Euclidian distance (PCAE), PCA with Mahalanobis distance (PCAM), FLDA, DLDA (Yu et al., 2001), BML (Moghaddam et al., 2000), NDA (Liu et al., 2004), UFS (Wang et al., 2004b) and DSL (Wang and Tang, 2004a) approaches. 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 equal error rate is obtained.

4.1. Results on AR database

In AR database (Martinez, 2002), color images are converted to gray-scale and cropped into the size of 120×170. Seventy-five subjects with 14 non-occluded images per subject are selected from the AR database. The first 7 images of 60 subjects are used in the training and also serve as gallery images. The second 7 images of the 60 subjects serve as probe genuine images. The remaining 15 subjects with 14 images per subject are used as probe imposters. For this large image size, we first apply PCA to remove the null space of $\boldsymbol{S}^t$ and then apply the DVPE approach on the 419-dimensional feature vectors. Fig. 4 shows the ROC curve that plots the correct verification rate (CVR) against the false acceptance rate (FAR).

The ROC curves of PCAE do not appear in Fig. 4 because their CVRs and FARs are so low that their values are out of the range of Fig. 4. 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 (Moghaddam et al., 2000). Fig. 4 shows that the proposed DVPE method consistently outperforms all other eight approaches for all different operating points (thresholds).

4.2. Results on FERET database 1

In FERET database (Phillips et al., 2000), 2388 images comprising of 1194 subjects (two images FA/FB per subject) are selected. 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

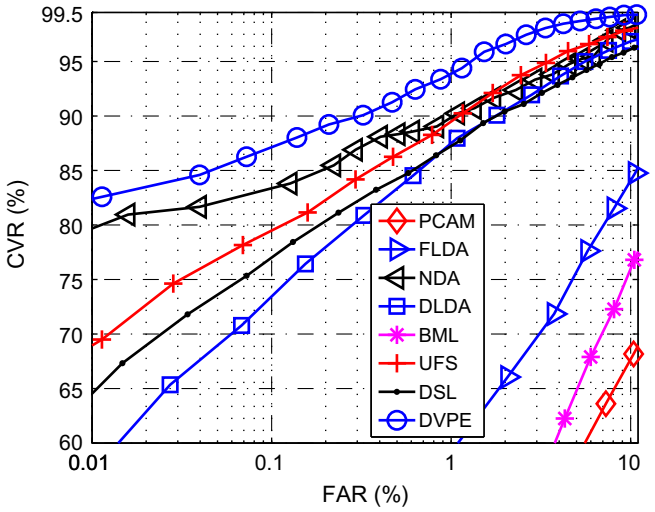

Fig. 4. Correct verification rate against the false acceptance rate on the AR face database. The total number of genuine matches is $7 \times 7 \times 60 = 2940$ and the total number of imposter matches is $14 \times 7 \times 15 \times 60 = 88,200$.

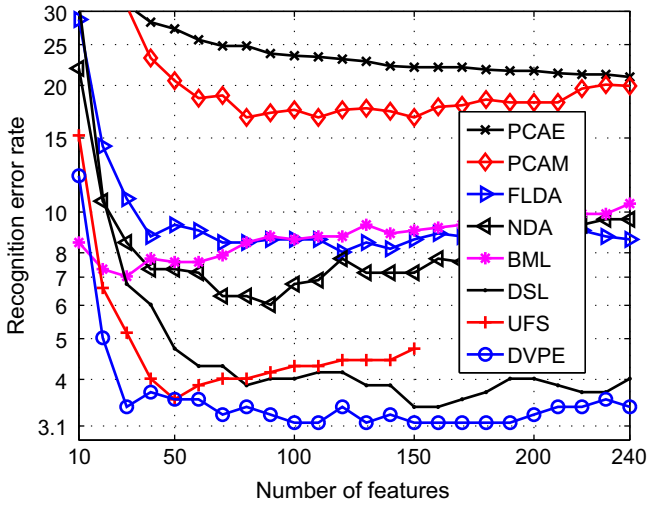

Fig. 6. Recognition error rate against the number of features used in the matching on the FERET database 1 comprising of 994 training images (497 subjects) and 1394 testing images (697 subjects).

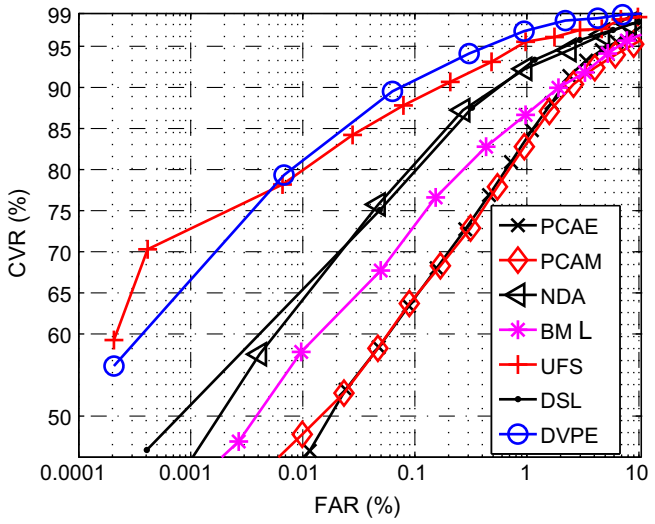

Fig. 5. Correct verification rate against the false acceptance rate on the FERET database 1. The total number of genuine matches is $697 \times 1 = 697$ and the total number of imposter matches is $697 \times 696 = 485,112$.

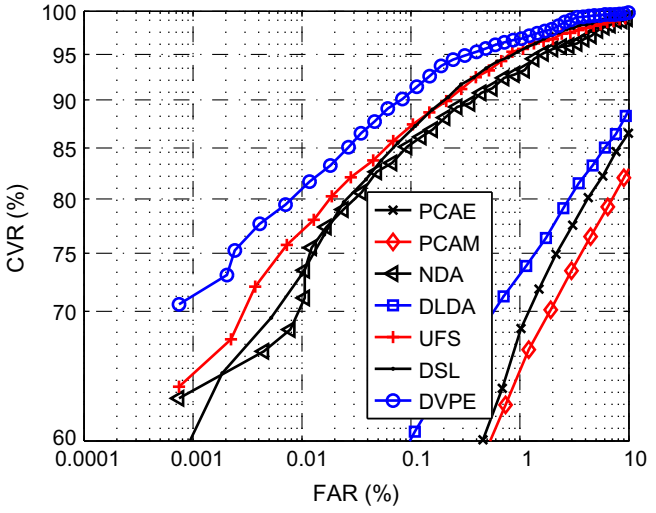

Fig. 7. Correct verification rate against the false acceptance rate on the FERET database 2 of 600 training and gallery images (200 subjects), 200 probe genuine images (200 subjects) and 224 probe imposter images (56 subjects). The total number of genuine matches is $200 \times 3 = 600$ and the total number of imposter matches is $4 \times 3 \times 56 \times 200 = 134,400$.

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). Fig. 5 shows the ROC curve that plots the CVR (%) against the FAR (%).

The ROC curves of FLDA and DLDA do not appear in Fig. 5 because their CVRs and FARs are too low to be included in Fig. 5. 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. 5 shows again that the proposed DVPE method achieves higher CVR for their corresponding FAR among other tested approaches for many different operating points.

Since this database has largest number of testing (or probe) subjects, we also evaluate our method for face identification process. The recognition error rate given in this work is the percentage of the incorrect top 1 match on the testing set. Fig. 6 shows the recognition error rate on the testing set against the number of features d used in the matching. Similar to the verification process, recognition error rates of DLDA are too high to be included in Fig. 6. DSL performs better for larger number of features because it evaluates the discrimination information separately from the two subspaces and then combined in the recognition phase. For lower number of features, UFS outperforms DSL because UFS suppresses noises by keeping very few of features in the initial stage. However, our proposed approach DVPE consistently outperforms all other approaches for all number of features.

4.3. Results on FERET database 2

In this experiment, we construct a database similar to Lu et al. (2006), by choosing 256 subjects randomly with at least four images per subjects from FERET database. However, we use the same number of images (four) per subject for all subjects. Three images per subjects of the first 200 subjects are used for training and also serve as gallery images. The remaining 200 images of

Table 1
Equal error rate (EER %) of various approaches on three different databases.

Database	AR	FERET1	FERET2
PCAE	34.2	5.1	13.0
PCAM	21.9	6.0	15.5
FLDA	13.2	35.0	34.8
NDA	5.0	4.0	3.6
DLDA	5.2	27.0	10.5
BML	19.1	5.6	23.8
UFS	4.3	3.0	2.7
DSL	5.9	3.8	2.5
DVPE	2.6	2.0	2.0

the 200 subjects are used as probe genuine images. All four images of the remaining 56 subjects serve as probe imposter images. The size of the normalized image is 130×150, same as that in (Lu et al., 2006). For such a large image size, we first apply PCA to remove the null space of $\boldsymbol{S}^t$ and then apply the proposed DVPE approach on the 599-dimensional feature vectors. For this database we conducted four runs of training and testing with distinct probe genuine image set in each run. More specifically, the ith images ($i = 1, 2, 3, 4$) of all training subjects are chosen to form probe genuine set and the remaining three 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. 7 shows the average ROC curve that plots the CVR (%) against the FAR (%) of the four runs of training and testing.

The ROC curves of FLDA and BML do not appear in Fig. 7 because their CVRs and FARs are so low that their values are out of the range in Fig. 7. Although the second highest ROC curve is same for both the experiments on FERET databases, Fig. 7 shows once more that the proposed DVPE method consistently delivers the most accurate face verification for all different operating points.

For an accurate record, verification performance in terms of equal error rate (EER) obtained from the above three experiments are numerically recorded in Table 1. It is defined by EER=FAR=FRR at a specific threshold, serves as a single number indicator of a verification system's performance. Where FRR (false reject rate) is the ratio of the number of rejected genuine matches to the total number of genuine matches (or one minus the correct verification rate 1-CVR). Table 1 clearly demonstrates the superior performance of the proposed DVPE approach to all other approaches tested in the experiments on three different face databases.

4.4. Summary of the experimental results

We have performed four sets of experiments with three different face databases that evaluate nine subspace based approaches for face recognition 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.

From the above experiments, it is evident 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 extracts 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 four 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 DVPE method shows superior verification and identification performances to all the other eight subspace based approaches. In all three experiments on the different face databases, the proposed DVPE method consistently achieves the highest correct verification rates (or lowest EER) at many 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 CVR. In identification experiment, the proposed DVPE approach consistently achieves the lowest recognition error rate for all number of features. The superior verification and identification performances of the proposed method is attributed to the prediction of the eigenvalues and then regularizing eigenfeatures in the eigenspace. These enable a global optimization by the discriminant evaluation in the whole space and alleviate the over-fitting problem as the unreliable or noise sensitive small and zero eigenvalues are replaced by the predicted values.

4.5. Limitations and future work

Although the proposed eigenmodel scheme works well for face verification and identification problems, it might not be well suited for other computer vision and pattern recognition (CV & PR) tasks like general object recognition, fingerprint or palm print recognition. Our proposed algorithm uses the principal eigenvalues from the training dataset to predict the unreliable and unknown eigenvalues, its effectiveness on other CV & PR problems would be an interesting future research topic worth to investigate. The proposed eigenmodel uses a function form of $1/f$ which fits well to the decaying nature of the eigenspectrum of face images. Investigating other function forms which could even better fit into the decaying nature of the eigenspectrum is an interesting topic. In fact, the decaying nature of the eigenspectrum could vary from problem to problem in CV & PR research areas. Proposing a generic eigenmodel for various CV & PR problems is a challenging task. Currently, we are investigating the eigenspectrum modeling for human activity recognition tasks. In addition, the proposed algorithm has no free parameter as selection choice. This could be an advantage for the implementation and application. On the other hand, however, the algorithm may not be optimal for all training tasks, some of which may have very small training samples while others may have large number of training samples. If the algorithm is made to be adaptive to the number of training samples then better results can be expected.

5. Conclusions

Subspace based approaches such as FLDA, DLDA, NDA and UFS discard a subspace before the discriminant evaluation. The ex-

tracted 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. The DSL approach scales features in the complementary subspace by the average eigenvalue of within-class scatter matrix over this subspace. As the eigenvalues in this subspace are not well estimated, their average may not be a proper scaling factor relative to those in the principal subspace.

This work shows the problems of feature scaling and extraction from high-dimensional data such as face images for face verification and identification tasks. To alleviate these problems we decompose the eignespace into three subspaces using two control points and predict the unreliable eigenvalues. The forecasted eigenvalues are used for regularization of eigenfeatures in the eigenspace. These enable 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. Experiments on AR and FERET databases demonstrate that the proposed approach consistently outperforms other eight subspace based approaches for face verification and identification tasks.

Acknowledgement

This work was supported by the Institute for Infocomm Research, A*STAR, Singapore.

References

Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J., 1997. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Trans. Pattern Anal. Machine Intell. 19 (7), 711–720.

Beveridge, R., Bolme, D., Teixeira, M., Draper, B., 2003. The CSU Face Identification Evaluation System Users Guide: Version 5.0. Technical Report: <http://www.cs.colostate.edu/evalfacerec/data/normalization.html>.

Cevikalp, H., Neamtu, M., Wilkes, M., Barkana, A., 2005. Discriminative common vectors for face recognition. IEEE Trans. Pattern Anal. Machine Intell. 27 (1), 4–13.

Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification. John Wiley and Sons, New York.

Friedman, J.H., 1989. Regularized discriminant analysis. J. Am. Stat. Assoc. 84 (405), 165–175.

Fukunaga, K., 1991. Introduction to Statistical Pattern Recognition, second ed. Academic Press.

Gao, H., Davis, J.W., 2006. Why direct lda is not equivalent to lda. Pattern Recognition 39, 1002–1006.

Huang, R., Liu, Q., Lu, H., Ma, S., 2002. Solving the small size problem of lda. In: Proc. 16th Internat. Conf. on Pattern Recognition, vol. 3, pp. 29–32.

Jiang, X.D., 2009. Asymmetric principal component and discriminant analyses for pattern classification. IEEE Trans. Pattern Anal. Machine Intell. 31 (5), 931–937.

Jiang, X.D., Mandal, B., Kot, A., 2006. Enhanced maximum likelihood face recognition. IEE Electron. Lett. 42 (19), 1089–1090.

Jiang, X.D., Mandal, B., Kot, A., 2007. Face recognition based on discriminant evaluation in the whole space. In: IEEE 32nd International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007). Honolulu, Hawaii, USA, pp. 245–248.

Jiang, X.D., Mandal, B., Kot, A., 2008. Eigenfeature regularization and extraction in face recognition. IEEE Trans. Pattern Anal. Machine Intell. 30 (3), 383–394.

Jiang, X.D., Mandal, B., Kot, A., 2009. Complete discriminant evaluation and feature extraction in kernel space for face recognition. Machine Vision Appl. 20 (1), 35–46.

Liu, W., Wang, Y., Li, S.Z., Tan, T.N., 2004. Null space approach of fisher discriminant analysis for face recognition. In: ECCV Biomet Authen, pp. 32–44.

Lu, J., Plataniotis, K.N., Venetsanopoulos, A.N., Li, S.Z., 2006. Ensemble-based discriminant learning with boosting for face recognition. IEEE Trans. Neural Networks 17 (1), 166–178.

Mandal, B., Jiang, X.D., Kot, A., Dec 2008. Verification of human faces using predicted eigenvalues. In: 19th Internat. Conf. Pattern Recognition (ICPR). Tempa, Florida, USA, pp. 1–4.

Martinez, A.M., 2002. Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. IEEE Trans. Pattern Anal. Machine Intell. 24 (6), 748–763.

Moghaddam, B., Jebara, T., Pentland, A., 2000. Bayesian face recognition. Pattern Recognition 33 (11), 1771–1782.

Moghaddam, B., Pentland, A., 1997. Probabilistic visual learning for object representation. IEEE Trans. Pattern Anal. Machine Intell. 19 (7), 696–710.

Moghaddam, B., Wihid, W., Pentland, A., 1998. Beyond eigenfaces: Probabilistic matching for face recognition. In: IEEE 3rd Internat. Conf. on Automatic Face and Gesture Recognition, pp. 30–35.

Phillips, P.J., Moon, H., Rizvi, S., Rauss, P., 2000. The feret evaluation methodology for face recognition algorithms. IEEE Trans. Pattern Anal. Machine Intell. 22 (10), 1090–1104.

Shakhnarovich, G., Moghaddam, B., 2005. Face Recognition in Subspaces. Springer, 201 Broadway, Cambridge, Massachusetts 02139.

Stan, Z.L., Jain, A.K., 2004. Face Recognition in Subspaces: Handbook of Face Recognition. Springer, 201 Broadway, Cambridge, Massachusetts 02139.

Wang, X., Tang, X., 2004a. Dual-space linear discriminant analysis for face recognition. In: IEEE Conf. on Computer Vision and Pattern Recognition, vol. 2, pp. 564–569.

Wang, X., Tang, X., 2004b. A unified framework for subspace face recognition. IEEE Trans. Pattern Anal. Machine Intell. 26 (9), 1222–1228.

Xu, D., Yan, S., Lin, S., Huang, T., 2008. Convergent 2-d subspace learning with null space analysis. IEEE Trans. Ciruits Systems Video Technol. 18 (12), 1753–1759.

Yang, J., Frangi, A.F., Yang, J.Y., Zhang, D., Jin, Z., 2005. Kpca plus lda: A complete kernel fisher discriminant framework for feature extraction and recognition. IEEE Trans. Pattern Anal. Machine Intell. 27 (2), 230–244.

Yu, H., Yang, J., 2001. A direct lda algorithm for high-dimensional data with application to face recognition. Pattern Recognition 34 (10), 2067–2070.

Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A., 2003. Face recognition: A literature survey. ACM Comput. Surveys 35 (4), 399–458.

Zhuang, X., Dai, D., 2007. Improved discriminant analysis for high-dimensional data and its application to face recognition. Pattern Recognition 40, 1570–1578.