Contents lists available at ScienceDirect

Pattern Recognition

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

# SVM-based feature extraction for face recognition

Sang-Ki Kim, Youn Jung Park, Kar-Ann Toh, Sangyoun Lee *

*Biometrics Engineering Research Center, School of Electrical & Electronic Engineering, Yonsei University, 134 Shinchon-dong, Seodaemun-gu, Seoul 120-749, Republic of Korea*

ARTICLE INFO

*Article history:*
Received 22 July 2009
Received in revised form
21 January 2010
Accepted 4 March 2010

*Keywords:*
Face recognition
Identity verification
Discriminant analysis
Support vector

ABSTRACT

The primary goal of linear discriminant analysis (LDA) in face feature extraction is to find an effective subspace for identity discrimination. The introduction of kernel trick has extended the LDA to nonlinear decision hypersurface. However, there remained inherent limitations for the nonlinear LDA to deal with physical applications under complex environmental factors. These limitations include the use of a common covariance function among each class, and the limited dimensionality inherent to the definition of the between-class scatter. Since these problems are inherently caused by the definition of the Fisher's criterion itself, they may not be solvable under the conventional LDA framework. This paper proposes to adopt a margin-based between-class scatter and a regularization process to resolve the issue. Essentially, we redesign the between-class scatter matrix based on the SVM margins to facilitate an effective and reliable feature extraction. This is followed by a regularization of the within-class scatter matrix. Extensive empirical experiments are performed to compare the proposed method with several other variants of the LDA method using the FERET, AR, and CMU-PIE databases.

## 1. Introduction

Over the past few decades, subspace projection-based representation methods, such as principal component analysis (PCA) [1], independent component analysis (ICA) [2], local feature analysis (LFA) [3], and non-negative matrix factorization (NMF) [4], have been extensively adopted as effective feature extraction means for face recognition. Via a set of basis vectors, these methods define a subspace from the training samples. By projecting the face image samples onto these basis vectors, one can utilize the projected weights as representation features for face recognition. The goal of these subspace projection methods is to reduce the dimensionality while maximizing the essential information for classification as well as minimizing possible redundancies.

Among them, the linear discriminant analysis (LDA) [5,6] has achieved remarkable progress in terms of its recognition performance. The main distinction of LDA from other subspace methods is its supervised nature. In other words, LDA utilizes the class information (the identity of faces) associated with each pattern sample to calculate a discriminant subspace. LDA simultaneously emphasizes the between-class separability and deemphasizes the within-class variation by searching a set of basis vectors that maximizes the *Fisher's criterion* [7]. The basis vectors of LDA is obtained from finding the eigenvectors of $\mathbf{S}_W{}^{-1}\mathbf{S}_B$, the product of the inverse of the within-class scatter matrix ($\mathbf{S}_W$) and the between-class scatter matrix ($\mathbf{S}_B$) [8]. Recently, the LDA has confronted a potential performance improvement via introduction of a nonlinear extension based on a *kernel trick* [9]. This kernel trick allows the conventional linear subspace methods to be directly applied for nonlinear feature mapping. As part of this scheme, the kernel Fisher discriminant (KFD) [10], the generalized discriminant analysis (GDA) [11], and the kernel direct discriminant analysis (KDDA) [12] have been proposed.

Nevertheless, the above variants of LDA appear to be confined within the inherent limitations of LDA. These limitations include the use of a common covariance function which is shared by each class [13,14], the small sample size problem [15,5,6] and a limited dimensionality inherent to the definition of the between-class scatter. Particularly, in multi-class problems, most LDA methods assume a homoscedastic distribution of data, i.e., every class shares a common covariance function in the scatters computation. Although such assumption simplifies both mathematical and computational treatments, the assumption may not be adequate for problems with heteroscedastically distributed data, i.e. data where each class owns a separate distribution property [13].

In view of the above limitations of LDA and motivated by the success of SVMs [20], particularly in terms of using it to enhance traditional classifiers [31,32], we propose a new feature extraction algorithm, called SVM-based discriminant analysis (SVM-DA) in this paper. By implanting an SVM margin to the framework of LDA, we can make the feature extraction applicable to heteroscedastic data while alleviating the SSS and the dimensionality problems. Several empirical experiments were performed to observe the effectiveness of the proposed method using FERET [16], AR [17], and CMU-PIE [18] databases.

* Corresponding author. Tel.: +82 2 2123 5768; fax: +82 2 313 2879.
*E-mail address:* syleee@yonsei.ac.kr (S. Lee).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.03.008

The rest of this paper is organized as follows: in Section 2, the problems of applying LDA in face recognition are addressed in a greater detail. In Section 3, some preliminaries are provided for immediate reference. The proposed methodology is described and illustrated in Section 4. This is followed by an introduction of the databases, the setup of our experiments and the results in Section 5. Finally, some concluding remarks are given in Section 6.

## 2. Problem definition and background

Given training face images $x_k \in \mathbb{R}^D$ (vectors obtained from reshaping two-dimensional images into a single column vector form and indexed by $k$), LDA tries to find a set of linear projection vectors $v$, which maximizes the Fisher's criterion [7]:

$$v = \underset{v}{\operatorname{argmax}} \frac{|v^T \mathbf{S}_B v|}{|v^T \mathbf{S}_W v|} \tag{1}$$

where $\mathbf{S}_B = \sum_{i=1}^{c} p_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T$ and $\mathbf{S}_W = \sum_{i=1}^{c} p_i \mathbf{S}_i$ are the between-class scatter matrix and the within-class scatter matrix, respectively. $\mathbf{S}_i = \sum_{x_k \in C_i} (x_k - \mathbf{m}_i)(x_k - \mathbf{m}_i)^T$ is the within-class scatter matrix of class $i$ where $C_i$ is its sample set. $c$ is the number of classes. $\mathbf{m}_i$ and $\mathbf{m}$ are the mean vector of the samples of class $i$ and mean of the entire training set, respectively. $p_i$ is the a priori probability of class $i$ which equals $N_i/N$, where $N_i$ and $N$ are the number of samples of class $i$ and that of entire training set.

In physical face recognition applications, the LDA faces the *small sample size* problem (SSS problem [15]). Since the majority of face recognition scenarios are based on snapshot face images [19], the data dimensionality $D$ typically exceeds the number of training samples in the database. Thus, very often the dimensionality of the scatter matrices is much larger than the number of their rank. This makes the within-class scatter matrix singular, leading to direct computation of its inverse intractable.

A common way to solve this SSS problem is to collect enough number of training images so that the rank of the within-class scatter, which is at most $N-c$, equals the size of scatter matrix which equals $D$. However, due to the inherent characteristic of the kernel trick, this solution does not apply for kernel Fisher discriminant. Kernel-based expansion assumes an implicit non-linear mapping of the input space to a high-dimensional feature space which possibly has infinite dimensionality. The kernel-trick transforms the scatter matrices into different forms, and finds an eigen-solution only within the sample space [20,11,12]. In this situation, the transformed scatter matrices have dimensionality of $N$, which equals the number of training samples, and, thus the within-class scatter matrix becomes always singular.

To avoid the SSS problem in LDA, a priori dimensionality reduction schemes were introduced. Belhumeur et al. [5] and Zhao et al. [6] applied the PCA before application of LDA (Fisherface method). A preceding PCA projection reduces the dimension of the data while minimizing the loss of information so that the within-class scatter matrix becomes non-singular in the transformed space.

However, it is argued that the PCA projection in the Fisherface method loses information residing in the zone of either the small or the zero eigenvalues. Worse still, the lost data usually contains a lot of discriminative information [21]. Under this viewpoint, a direct LDA (D-LDA [22]) and a null space-based LDA (N-LDA [21]) were proposed. The main strategies of D-LDA and N-LDA are to find the most discriminant set of eigenvectors of the between-class scatter matrix which is projected into the eigenspace (a subspace of nonzero eigenvalues) or the null space (a subspace of zero eigenvalues) of the within-class scatter matrix. However, these schemes are opposite extremes to each other, and lose information residing in the complementary subspace of the within-class scatter matrix.

For D-LDA and N-LDA, it is obvious that certain discriminating information resides in both the eigenspace and the null space. However, utilizing both subspaces requires a balanced weighting between the two subspaces. Recently, an eigen-feature regularization and extraction (ERE) method has been proposed by Jiang et al. [23]. Unlike previous two methods, ERE utilizes the entire sample space which consists of both the eigenspace and the nullspace. To provide an appropriate weighting of both subspaces, a modeling approach is adopted which regularizes the small and zero eigenvalues of the within-class scatter matrix via a reciprocal function. This eigenvalue regularization not only utilizes both subspaces but also reduces the noise factor which is dominant particularly at the region of small eigenvalues.

Another issue is the lack of dimensionality in LDA. Since the maximum rank of $\mathbf{S}_B$ is bounded to one less than the number of classes, $c-1$, so is the number of extracted features. This small dimensionality of $\mathbf{S}_B$ is often not sufficient for extracting discriminative information from the training data, and this confines the recognition performance. Noting that $\mathbf{S}_B$ describes the scatter of class means, this problem can be regarded as another SSS problem on $\mathbf{S}_B$. Although collecting more images can somewhat relieve the problem, it remains a limitation to the Fisher's criterion in the aspect of extracting useful information efficiently.

The essence of finding a discriminant subspace is to reduce the dimensionality while preserving the classification structure of the training data. A theoretic lower bound for the dimensionality can be specified by the *intrinsic dimensionality* of data [8]. The intrinsic dimensionality conceptually describes the minimum dimensionality of a subspace which can hold all the classification structure, i.e. discriminative information. For convenience sake, we will use the term *intrinsic subspace* to identify such a subspace. Consequently, the effectiveness of a projection subspace can be determined from the ratio of the preserved intrinsic subspace to the dimensionality of extracted features. Since the efficiency of LDA subspace is pretty much limited by the sub-optimality of its criterion and the linearity (or the kernel space for the nonlinear case), the rank of a conventional between-class scatter is usually too small to contain the intrinsic dimension.

The last problematic issue is the assumption regarding *homoscedasticity* and *Gaussianity* of all data. Conventional LDA assumes that all data classes share the same density function and are normally distributed. This assumption is often inconsistent with data used in face recognition applications. The variation in a face image is known to be caused by external factors such as illumination, view point (head pose), facial expression, occlusion, etc. [19,24]. These external factors affect the appearance of a three-dimensional object. The resulting face image, which is a two-dimensional projection of reflected light from the face, thus heavily depends on the geometric shape of the face which is unique for each identity. Consequently, we can induce that the variation of face images is class-specific (specific to each identity).

Under the Bayesian framework, a classification cannot achieve an optimum separability without a precise estimation of the density function [8,25]. Noting that LDA is a special case of the Bayes classifier assuming a homoscedastic and Gaussian density function [13,8], the LDA would thus be suboptimal for heteroscedastic data which frequently occurs in face recognition. Although such assumption may contribute to a generalized performance when only a small number of samples per class is available, the invaluable information regarding the real data distribution has been abandoned. In order to find a good discriminative subspace for face recognition, the conventional LDA needs to be modified to consider the distributions of respective pattern classes.

Taking all the above shortcomings of LDA into account, we propose a method to explicitly deal with the above issues and subsequently extend it to nonlinear feature extraction via a kernel trick. We adopted the eigenspectrum regularization (ER) scheme from ERE [23] to relieve the former SSS problem. The ER procedure is briefly described in the following section, and its application to our proposed nonlinear feature extraction will be presented in Section 4. The last two issues are deeply related to each other and are inherited from the Fisher's criterion. Thus we modified the criterion itself with the SVM margin instead of the class mean separation as seen in the conventional Fisher's criterion. This modification is graphically illustrated at the beginning of Section 4 and then followed by a description of its kernel-based implementation.

## 3. Preliminaries

### 3.1. Support vector machine

To provide an immediate reference to the adopted concept, we include a brief introduction to SVM in this section. The interested readers are referred to [20] for greater details regarding SVM.

Without loss of generality, consider a two-category classification problem. The SVM adopts a supervised learning paradigm. Hence, similar to LDA discussed above, SVM can be considered as a classifier which constructs a decision boundary according to training examples [8].

Suppose we have some indexed ($k=1,\dots,M$) training data denoted by $x_k \in \mathbb{R}^D$, each with a corresponding label $y_k \in +1,-1$. A decision boundary in the form of a *hyperplane* can generally be described as follows:

$$\omega \cdot x + b = 0 \tag{2}$$

As illustrated in Fig. 1, the hyperplane separates samples of class 1 from those of class 2. If $x_1$ and $x_2$ are both on this boundary surface, they satisfy following equation:

$$\omega \cdot x_1 + b = \omega \cdot x_2 + b = 0 \tag{3}$$

or

$$\omega \cdot (x_1 - x_2) = 0 \tag{4}$$

Eq. (4) implies that $\omega$ is a vector perpendicular to any vector lying on the decision hyperplane [27].

The major difference between SVM and LDA lies in their optimization criteria: LDA maximizes a discriminative projection while SVM maximizes a discrimination *margin*. Here, a margin is defined as the distance between the decision boundary and the training samples nearest to it as shown in Fig. 1 (i.e. the width of the shaded area). Those training samples which determine the decision boundary are called *support vectors* and they satisfy following equation:

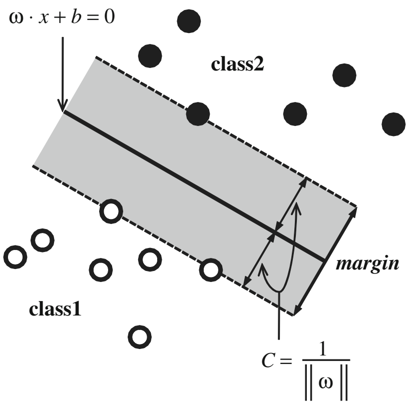

**Fig. 1.** Support vector machine. The optimal separating hyperplane is the solid line, while the margin is the gap between the dashed lines of which the width is $2C=2/||\omega||$.

$$y_k(\omega \cdot x_k + b) = 1 \tag{5}$$

Let $x_c$ be a support vector corresponding to $y_c=1$, and $x_c'$ be its projection to the hyperplane. Then, $x_c$ can be represented as

$$x_c = x_c' + \rho \frac{\omega}{\|\omega\|} \tag{6}$$

where $\rho$ is the distance between $x_c$ and the hyperplane. Substituting Eqs. (6)–(5), we have

$$y_c(\omega \cdot x_c + b) = \omega \cdot x_c' + b + \rho \frac{\|\omega\|^2}{\|\omega\|} = \rho\|\omega\| = 1 \tag{7}$$

Thus, the margin is given as $2\rho = 2/||\omega||$.

When the data is linearly separable, there exists a boundary hyperplane which optimally separates the two classes without error. SVM tries to find such a hyperplane based on following optimization criterion [26]:

$$\min_{\omega,b} \|\omega\| \text{ subject to } y_k(\omega \cdot x_k + b) \geq 1, \quad \forall k \tag{8}$$

where the margin is given by $2/||\omega||$. Thus, minimizing $||\omega||$ is equivalent to maximizing the margin. Solving this quadratic problem gives the hyperplane parameter as follows:

$$\omega = \sum_{\forall x_k \in S} \alpha_k y_k x_k \tag{9}$$

where $S$ is a set of support vectors for both classes, and $\alpha_k$ is a trained weight on the corresponding support vectors. Based on this solution, one can classify an arbitrary new input $x$ using

$$f(x) = \text{sign}(\omega \cdot x + b) = \text{sign}\left(\sum_{\forall x_k \in S} \alpha_k y_k x_k \cdot x + b\right) \tag{10}$$

The entire platform can be generalized to a nonlinear case. This generalization can be accomplished by mapping the samples to a certain high-dimensional space $\mathcal{H}$:

$$\begin{aligned} &\phi : \mathbb{R}^D \mapsto \mathcal{H} \\ &x \mapsto \phi(x) \end{aligned} \tag{11}$$

Under such a high-dimensional space, usually called the *feature space*, the original overlapping data could become linearly separable. Constructing a separating hyperplane in that space yields a nonlinear decision boundary in the input space (face space). However, since the dimensionality of this new feature space could be very high (possibly infinite), a direct data mapping often becomes intractable. Nevertheless, by adopting a kernel function $k(x_i x_j)$, the nonlinear SVM can be formulated in a tractable manner without explicitly carrying out the mapping into the feature space:

$$\begin{aligned} f(x) &= \text{sign}(\Omega \cdot \phi(x) + b) \\ &= \text{sign}\Big(\sum_{\forall x_k \in S} \alpha_k y_k \phi(x_k) \cdot \phi(x) + b\Big) = \text{sign}\Big(\sum_{\forall x_k \in S} \alpha_k y_k k(x_k \cdot x) + b\Big) \end{aligned} \tag{12}$$

### 3.2. Eigenspectrum regularization

As introduced previously, the conventional within-class scatter is often singular due to the SSS problem where the small eigenvalues are likely to be caused by lack of samples rather than due to the statistical property of data distribution. Hence, the resulting solution might be over-fitted to the training data. On the other hand, solving the problem by discarding the subspace of small eigenvalues such as GDA and KDDA do, yields a solution which is not unique and may lose much information.

The eigenspectrum regularization (ER) scheme [23] regularizes the within-class scatter matrix making it a full-rank matrix while relieving those sample noises. The first step in the procedure of ER is to decompose the within-class scatter matrix into eigenvalues and eigenvectors via *eigen-decomposition*. Next, the span of these scatter matrix is sectioned into three subspaces namely, a reliable subspace (face space), an unstable subspace (noise space), and a null subspace. These subspaces are specified by the following two variables:

$$r = \max\{\forall l | \lambda_l^W > \varepsilon\}$$
$$m = \min\left\{\forall l | \lambda_l^W < \left(\lambda_{med} + \mu(\lambda_{med} - \lambda_r^W)\right)\right\} \tag{13}$$

where $\lambda_l{}^W$ are the decomposed eigenvalues sorted in decreasing order, $\varepsilon$ is a threshold having a very small value compared to $\lambda_1^W$, $\lambda_{med} = \mathrm{median}(\forall \lambda_l^W | l \le r)$, and $\mu$ is the tuning parameter which is set to one according to the original reference [23]. Subsequently, the parameters of reciprocal function are calculated based on the first and the last eigenvalues of face space ($\lambda_1^W$ and $\lambda_r{}^W$):

$$\alpha_{er} = \frac{\lambda_1^W \lambda_m^W (m-1)}{\lambda_1^W - \lambda_m^W}, \quad \beta_{er} = \frac{m\lambda_m^W - \lambda_1^W}{\lambda_1^W - \lambda_m^W} \tag{14}$$

Based on the model function $\alpha_{er}/(l+\beta_{er})$, the regularized eigenvalue $\tilde{\lambda}_l^W$ for each respective subspace is given as follows:

$$\tilde{\lambda}_l^W = \begin{cases} \lambda_l^W, & l < m \\ \alpha_{er}/(l+\beta_{er}), & m \le l \le r \\ \alpha_{er}/(r+1+\beta_{er}), & r < l \le N_W \end{cases} \tag{15}$$

where $N_W$ is the size of given scatter matrix. Then the regularized within-class scatter matrix is given as

$$\tilde{\mathbf{S}}_W = \mathbf{V}_W \tilde{\mathbf{\Lambda}}_W \mathbf{V}_W^T \tag{16}$$

where $\tilde{\mathbf{\Lambda}}_W = \mathrm{diag}[\tilde{\lambda}_1^W, \ldots, \tilde{\lambda}_{N_W}^W]$ is a matrix with the regularized eigenvalues in its diagonal, and $\mathbf{V}_W = [v_1^W, \ldots, v_{N_W}^W]$ is a matrix of corresponding eigenvectors.

## 4. Proposed methodology

### 4.1. SVM-based modification of Fisher's criterion

Since an optimal separating hyperplane of SVM can be constructed based on several supporting vectors which arise from a fraction of the training data, the SVM does not need either a parametric modeling (Gaussianity and Homoscedasticity [13]) or a prediction of data distribution. Leveraging on these properties of SVM, we will reformulate the Fisher's criterion to cater for heteroscedastic and non-Gaussian data.

A conventional LDA defines the between-class scatter based on the class centers' displacements. The solid arrow in Fig. 2(a) shows a graphical description of the conventional LDA. For convenience sake, let us call the vector connecting the centers of two classes a *between-class vector* $\Delta_{ij}$. Although LDA finds an optimal projection via a relation between the between-class scatter and the within-class scatter, the definition of the between-class scatter involves a parametric assumption of Gaussianity and homoscedasticity [13]. Hence a conventional LDA might be ineffective when the assumption is violated.

Considering the goal of LDA is to find an optimal projection for classification, it would be desirable to replace $\Delta_{ij}$ by a vector which orientates according to the boundaries of cluster distributions that appear best separated. To find this vector, we adopt SVM search for an optimal projection from the separation margin perspective. As shown in Fig. 2(b), SVM finds a *separating hyperplane* which maximizes the margin between two classes. In other words, the normal vector of the SVM hyperplane is the one where the margin is maximized. From the OSH equation of Eq. (2), the hyperplane parameter $\omega$ is a vector which is perpendicular to the boundary hyperplane (see Fig. 1 and Eq. (4)). Thus, we get the normal vector as: $\omega/||\omega||$. Instead of using the original between-class vector, we shall use this SVM-based normal vector to redefine the between-class scatter. This constitutes a novel and useful idea to define the between-class scatter for real world data applications.

Since SVM is originally a two-category classifier, it is required to modify the scatter matrix for multi-class problems. Recalling that the between-class scatter matrix originally depicts the displacements of class centroids [27], a multi-class $\mathbf{S}_B$ can be modified as follows:

$$\mathbf{S}_B = \sum_{i=1}^{c} p_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T = \sum_{i=1}^{c-1}\sum_{j=i+1}^{c} 2p_ip_j(\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T$$
$$= \sum_{i=1}^{c-1}\sum_{j=i+1}^{c} 2p_ip_j\Delta_{ij}\Delta_{ij}^T \tag{17}$$

The above matrix is also known as a *pair-wise scatter matrix* [28]. To define a new between-class scatter matrix, we replace the between-class vector, $\Delta_{ij}$, by the normal vector $\omega_{ij}/||\omega_{ij}||$ where $\omega_{ij}$ is the hyperplane parameter of the SVM classifier which separates class $i$ from class $j$:

$$\mathbf{S}_B' = \sum_{i=1}^{c-1}\sum_{j=i+1}^{c} 2p_ip_j\left(\frac{\omega}{\|\omega\|}\right)\left(\frac{\omega}{\|\omega\|}\right)^T \tag{18}$$

As illustrated in Fig. 2, a key difference between $\Delta_{ij}$ and $\omega_{ij}/||\omega_{ij}||$ is that the latter one takes into account the orientation of data distribution whereas $\Delta_{ij}$ does not.

In the conventional LDA, the pair-wise scatters are summed up without giving a certain weight to each pair of class means. This results in a projection in favor of a wide between-class separation. A large $\omega_{ij}/||\omega_{ij}||$ implies that both classes are well separated. Nonetheless, such $\omega_{ij}/||\omega_{ij}||$ contributes to a dominating (larger) amount to the overall eigenvalue [25], and deemphasizes smaller

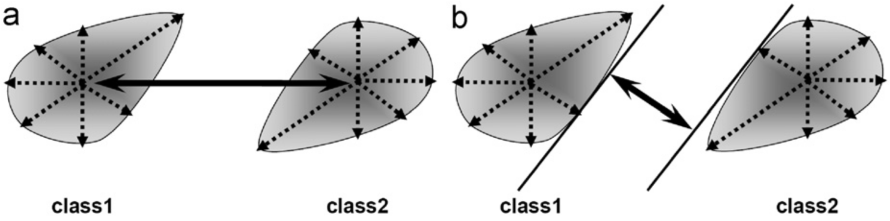

Fig. 2. (a) Conventional definition of the between-class vector and (b) the normal vector of hyperplane of the proposed SVM-DA (solid arrows).

ones which may actually have more discriminating information. In this situation, precise (discriminant) information of closely neighboring classes (small between-class vector) can be lost. Hence, we propose to weight the normal vectors of each hyperplane with an inverse of the corresponding margin $2/||\omega_{ij}||$. Thus, the new between-class scatter matrix is finally given as follows:

$$\mathbf{S}'_B = \sum_{i=1}^{c-1}\sum_{j=i+1}^{c} 2p_i p_j \frac{1}{2/\|\omega\|}\left(\frac{\omega}{\|\omega\|}\right)\left(\frac{\omega}{\|\omega\|}\right)^T = \sum_{i=1}^{c}\sum_{j=1}^{c} p_i p_j \frac{\omega\omega^T}{\|\omega\|} \tag{19}$$

Based on this newly defined between-class scatter matrix, we then solve the following modified criterion:

$$v = \operatorname*{argmax}_{v} \frac{|v^T \mathbf{S}'_B v|}{|v^T \mathbf{S}_W v|} \tag{20}$$

The optimum projection of this criterion can be found as the eigenvectors corresponding to the largest eigenvalues.

With this newly defined criterion, there is a conceptual modification from the Fisher's criterion: *maximizing the between-class margin* while minimizing the within-class scatter. Besides the nonparametric generalization, $\mathbf{S}'_B$ relaxes the dimensionality problem of LDA. The maximum rank of $\mathbf{S}'_B$ is bounded by $\min(c(c-1)/2, N_{SV})$ instead of $c-1$, where $N_{SV}$ is the total number of support vectors. This dimensionality covers the subspace spanned by the training data, hence loosens the dimensional bound for feature extraction.

### 4.2. Extraction of nonlinear features

#### 4.2.1. Between-class scatter in feature space

Since the SVM supports a nonlinear expansion using the kernel trick, the proposed method can be readily extended to its nonlinear version. The nonlinear SVM [29] allows an effective discriminant analysis in the proposed framework.

To begin with the newly defined between-class scatter matrix, we first train the binary SVM classifiers for all combinatorial pairs of pattern classes. Let $\Omega_{ij}$ denotes the parameter of the SVM hyperplane which separates class $i$ and class $j$ in the feature space $\mathcal{H}$. We can then express $\Omega_{ij}$ as follows:

$$\Omega_{ij} = \sum_{\forall x_k \in S_{ij}} \alpha_{ij}^k y_{ij}^k \phi(x_k) = \mathbf{\Phi}\alpha_{ij} \tag{21}$$

where $S_{ij}(\subset \mathbf{\Phi})$ is the set of support vectors, ${\alpha_{ij}}^k$ and ${y_{ij}}^k$ are the weight and the label for corresponding support vector $x_k$, $\mathbf{\Phi} = [\phi(x_1)\ldots\phi(x_N)]$ is the training data set mapped to the feature space, $\alpha_{ij}$ is a $N$-dimensional vector in which ${\alpha_{ij}}^k {y_{ij}}^k$ are located at the positions of corresponding support vectors in $\mathbf{\Phi}$. Then, the proposed SVM-based between-class scatter matrix in feature space is given as

$$\mathbf{S}'_B = \sum_{i=1}^{c-1}\sum_{j=i+1}^{c} p_i p_j \frac{\Omega_{ij}\Omega_{ij}^T}{\|\Omega_{ij}\|} = \frac{1}{N^2}\sum_{i=1}^{c-1}\sum_{j=i+1}^{c} N_i N_j \frac{\Omega_{ij}\Omega_{ij}^T}{\sqrt{\Omega_{ij}^T\Omega_{ij}}} \tag{22}$$

#### 4.2.2. Kernel formulation

In terms of kernel representation, we need to transform the above formulation in the form of dot product between two vectors. Whenever a dot product is used, it is replaced with the kernel function. First, we express both scatter matrices in a matrix form as follows:

$$\mathbf{S}'_B = \mathbf{\Omega}\mathbf{A}\mathbf{\Omega}^T = \mathbf{\Phi}\boldsymbol{\alpha}\mathbf{A}\boldsymbol{\alpha}^T\mathbf{\Phi}^T \tag{23}$$

where $\mathbf{\Omega} = [\Omega_{1,2},\ldots,\Omega_{1,c},\ldots,\Omega_{(c-1),c}]$, $\mathbf{A} = \mathrm{diag}[(N_1 \times N_2/N^2),\ldots,(N_{c-1} \times N_c/N^2)]$, and $\boldsymbol{\alpha} = [a_{1,2},\ldots,a_{(c-1),c}]$. The matrix-form expression for the within-class scatter can be obtained as follows:

$$\mathbf{S}_W = \sum_{i=1}^{c}\sum_{\forall x_j \in C_i}\left(\phi(x_j)-\overline{\phi}_i\right)\left(\phi(x_j)-\overline{\phi}_i\right)^T = \mathbf{\Phi}(\mathbf{I}-\mathbf{B}-\mathbf{B}^T+\mathbf{B}\mathbf{B}^T)\mathbf{\Phi}^T \tag{24}$$

where $\overline{\phi}_i$ is the mean vector of the samples of $i$th class, $\mathbf{B}$ is a $(N \times N)$ block diagonal matrix given by $(\mathbf{B}_i)_{i=1,\ldots,c}$ and $\mathbf{B}_i$ is a $(N_i \times N_i)$ matrix with terms all equal to $1/N_i$. Then, substituting Eq. (23) and (24) into Eq. (20), we have

$$\lambda = \frac{v^T\mathbf{\Phi}\boldsymbol{\alpha}\mathbf{A}\boldsymbol{\alpha}^T\mathbf{\Phi}^T v}{v^T\mathbf{\Phi}(\mathbf{I}-2\mathbf{B}-\mathbf{B}\mathbf{B}^T)\mathbf{\Phi}^T v} \tag{25}$$

Consider a desired basis vector $v$ with nonzero eigenvalues should lie within the span of training vectors $\mathbf{\Phi}$ in the feature space [11,12], then there exist coefficients $\beta$ such that

$$v = \sum_{k=1}^{N} \beta_k \phi(x_k) = \mathbf{\Phi}\beta \tag{26}$$

where $\beta = [\beta_1 \ldots \beta_N]^T$ is a $N$-dimensional vector. Substituting Eq. (26) into Eq. (25), we can express our criterion as the following quotient:

$$\lambda = \frac{\beta^T\mathbf{\Phi}^T\mathbf{\Phi}\boldsymbol{\alpha}\mathbf{A}\boldsymbol{\alpha}^T\mathbf{\Phi}^T\mathbf{\Phi}\beta}{\beta^T\mathbf{\Phi}^T\mathbf{\Phi}(\mathbf{I}-2\mathbf{B}-\mathbf{B}\mathbf{B}^T)\mathbf{\Phi}^T\mathbf{\Phi}\beta} \tag{27}$$

Since calculating $\mathbf{\Phi}^T\mathbf{\Phi}$ only requires a dot product in $\mathcal{H}$, it can be replaced by a kernel function. Certainly, this kernel function should be the same one used for the SVM hyperplanes in Eq. (12). Let $\mathbf{K}$ be a $N \times N$ matrix which consists of a dot product of training samples in the feature space $\mathcal{H}$:

$$\mathbf{K} = (k_{ij})_{\substack{i=1,\ldots,N\\ j=1,\ldots,N}}, \quad \text{where } k_{ij} = k(x_i,x_j) = \phi^T(x_i)\phi(x_j) \tag{28}$$

Through this replacement of $\mathbf{\Phi}^T\mathbf{\Phi}$ with $\mathbf{K}$, we can express Eq. (27) with tractable matrices:

$$\lambda = \frac{\beta^T\mathbf{K}\boldsymbol{\alpha}\mathbf{A}\boldsymbol{\alpha}^T\mathbf{K}\beta}{\beta^T\mathbf{K}(\mathbf{I}-2\mathbf{B}-\mathbf{B}\mathbf{B}^T)\mathbf{K}\beta} = \frac{\beta^T\mathbf{S}'^K_B\beta}{\beta^T\mathbf{S}^K_W\beta} \tag{29}$$

resulting in an eigenvalue problem expressed in terms of the newly defined scatter matrices, $\mathbf{S}'^K_B$ and $\mathbf{S}^K_W$, with eigenvector $\beta$.

#### 4.2.3. Eigenvalue resolution and regularization

In this section, we shall tackle the eigenvalue problem given by Eq. (29) and avoid the SSS problem at the same time. Firstly, we decompose and regularize the matrix $\mathbf{S}^K_W$ following the ER procedure given in Section 3.2:

$$\tilde{\mathbf{S}}^K_W = \mathbf{V}_W\tilde{\mathbf{\Lambda}}_W\mathbf{V}_W^T \tag{30}$$

Then, using the regularized eigenvalues $\tilde{\mathbf{\Lambda}}_W$ and $\mathbf{V}_W$, we define a whitening (diagonalizing and scaling) matrix:

$$\mathbf{P} = \mathbf{V}_W{\tilde{\mathbf{\Lambda}}_W}^{-1/2}, \quad \text{where } \mathbf{P}^T\tilde{\mathbf{S}}^K_W\mathbf{P} = ({\tilde{\mathbf{\Lambda}}_W}^{-1/2}\mathbf{V}_W^T)(\mathbf{V}_W\tilde{\mathbf{\Lambda}}_W\mathbf{V}_W^T)(\mathbf{V}_W{\tilde{\mathbf{\Lambda}}_W}^{-1/2}) = \mathbf{I}_N \tag{31}$$

where $\mathbf{I}_N$ is a $N$-dimensional identity matrix. Since $\mathbf{P}$ is a full-rank matrix, there exists a unique solution $\beta$ satisfying $\beta = \mathbf{P}\beta'$.

Substituting these representations into Eq. (29), we can diagonalize $\mathbf{S}^K_W$ and collapse the denominator using $\beta'^T\beta' = 1$:

$$\lambda = \frac{\beta'^T\mathbf{P}^T\mathbf{S}'^K_B\mathbf{P}\beta'}{\beta'^T\mathbf{P}^T\tilde{\mathbf{S}}^K_W\mathbf{P}\beta'} = \frac{\beta'^T\mathbf{P}^T\mathbf{S}'^K_B\mathbf{P}\beta'}{\beta'^T\mathbf{I}_N\beta'} = \beta'^T\mathbf{P}^T\mathbf{S}'^K_B\mathbf{P}\beta' \tag{32}$$

which is a standard form of the eigenvalue problem. By solving this problem, we $\beta'$ as the eigenvector of $\mathbf{P}^T\mathbf{S}'^K_B\mathbf{P}$ and subsequently $\beta$.

Finally, we have the desired vectors computed as $v = \mathbf{\Phi}\beta$. In order for $v$ to have a unit length in $\mathcal{H}$, we normalize $\beta$ so that the following conditions are met:

$$||v|| = v^T v = \beta^T \mathbf{\Phi}^T \mathbf{\Phi} \beta = \beta^T \mathbf{K} \beta = 1 \tag{33}$$

Here, we divide $\beta$ by $\sqrt{\beta^T \mathbf{K} \beta}$ to have the corresponding vectors $v$ normalized.

For any arbitrary test input $z$, we can calculate its projections using

$$v^T z = \beta^T \mathbf{\Phi}^T z = \beta^T \gamma(z) \tag{34}$$

where $\gamma(z) = [k(x_1, z), \ldots, k(x_N, z)]^T$ is a $(N \times 1)$ kernel vector.

The procedure of implementing the proposed SVM-DA is summarized as follows:

1. Train $c(c-1)/2$ SVM classifiers separating each and every pair of class samples based on the kernel matrix $\mathbf{K}$ and find the SVM parameter $\boldsymbol{\alpha}$.
2. Based on Eq. (29), calculate $\mathbf{S}'^K_B$ and $\mathbf{S}^K_W$ using the obtained $\boldsymbol{\alpha}$ plus $\mathbf{A}$ and $\mathbf{B}$ of Eqs. (23) and (24).
3. Decompose $\mathbf{S}^K_W$ into its eigenvalues $\mathbf{\Lambda}$ and eigenvectors $\mathbf{V}_W$ via eigen-decomposition.
4. Calculate the regularized eigenvalues $\tilde{\mathbf{\Lambda}}$ based on Eq. (15).
5. Calculate the whitening matrix $\mathbf{P}$ based on Eq. (31).
6. Compute $\beta'$ by solving the eigenvalue problem of Eq. (32) and calculate $\beta$.
7. Compute projections of test sample onto the eigenvector $v$ based on Eq. (34).

## 5. Experiments

### 5.1. Databases and experiment settings

To evaluate the proposed method, we perform experiments on three publicly available databases namely, FERET [16], AR [17], and CMU-PIE [18] (image samples are shown in Fig. 3). The face of each image was located manually by clicking a mouse at the center of each eye. All images were normalized to 56 × 46 pixels according to the eye centers, by rotating and resizing. Then, the images were masked revealing only the face region, histogram-equalized, and scaled so that the pixels are normalized to have zero mean and unit variances.

The data set for FERET database is constructed using 1702 images taken from 256 identities where there are at least 4 images per identity. For the AR database, the data set comprises of 1680 images taken from 120 identities. Each identity participated in two sessions, which are separated by a two-week interval. For each session, 7 images are chosen which were captured under different states by varying illumination and facial expression. For the CMU-PIE database, 1840 images taken from 68 identities are used for the experimentation. The data samples of each identity contain variation of head pose, facial expression, illumination, light source, and eye glasses.

Table 1 summarizes the specifications of data sets used in the above three databases. A similar number of face images were used in all experiments, while the number of identities used was nearly halved in each consecutive experiment. Although each data set contains images taken under different conditions, an analysis of the relative performances and their trends shall reveal significant information regarding the relation among the types of scatter matrices and the number of samples.

We perform experiments on the proposed SVM-DA, a conventional LDA (Fisherface) based on [5], ERE [23], KFD [10], GDA [11], and KDDA [12] in this comparative study. As for those nonlinear methods (including the proposed one) a polynomial kernel will be adopted:

$$k(x_i, x_j) = (w_p(x_i x_j) + b_p)^{D_p} \tag{35}$$

For simplicity, we fixed the bias parameter ($b_p$) to one and the polynomial degree parameter ($D_p$) to three, and changed only the weight parameter ($w_p$) as in [12]. When implementing the Fisherface, we heuristically determined the number of PCA bases for dimensionality reduction. We presented the best result for each respective database. The GDA and KDDA were implemented using the Matlab sources from [30] which are provided by Baudat and Anouar [11] and Lu et al. [12].

The distinguishing characteristics of all compared methods are listed in Table 2. Here we see that the proposed SVM-DA uses a novel pair-wise margin-based between-class scatter to deal with the non-Gaussian and heteroscedastic data while inheriting advantages offered by ERE. The advantages include a nonlinear kernel for data with complex decision boundary, and a stable solution for scatter matrix inverse offered by regularization.

All the experiments are carried out under the verification scenario, and the resulting equal error rates (EER) are recorded as the performance measure. The comparative results are presented in two phases. In the first phase, a graphical comparison of the best EERs of nonlinear methods is shown over the kernel variable $w_p$. For each setting of the $w_p$ value, the optimal numbers of projection bases are selected for each method respectively. In the second phase, a graphical comparison of the EERs of all methods is presented by varying the number of projection bases. For those nonlinear methods, the kernel parameter is set according to the optimal setting of each individual algorithm.

To prevent a biased evaluation, we perform a 4-fold cross-validation for all experiments. The original data set is partitioned into 4 subsets so that there is no common identity among them. A single subset is retained for testing, and the other 3 subsets are used for training. After running 4 folds, the test results are averaged.

### 5.2. Summary of results and discussion

Fig. 4 shows the comparative results of nonlinear methods by varying the kernel parameter $w_p$ and Fig. 5 shows the results varying the number of features $d$. The best result of each method is reported in Table 3. Below each EER, the number of features $d$ and the kernel parameter $w_p$ where respective method shows the best result are written in the parenthesis.

As shown in Table 2, the main distinction between GDA and KDDA is the scatter matrix in the denominator of the optimization criterion function: the total scatter matrix and the within-class scatter matrix. Noting that the total scatter matrix is equivalent to the sum of between-class scatter and the within-class scatter matrices [15], $\mathbf{S}_T = \mathbf{S}_B + \mathbf{S}_W$, maximizing the modified criterion of GDA, $\max(\mathbf{S}_B/\mathbf{S}_T)$, is actually equivalent to maximizing the conventional Fisher criterion:

$$\max(\mathbf{S}_B/\mathbf{S}_T) \Leftrightarrow \min(\mathbf{S}_T/\mathbf{S}_B) = \min(\mathbf{S}_W/\mathbf{S}_B + \mathbf{I}) \Leftrightarrow \max(\mathbf{S}_B/\mathbf{S}_W) \tag{36}$$

Thus, if all the scatter matrices are assumed to be invertible, the GDA yields exactly identical eigenvectors to those of KDDA but different eigenvalues as $\lambda_{KDDA}/(\lambda_{KDDA}+1)$, where $\lambda_{KDDA}$ is the eigenvalue of KDDA. In other words, the discriminant bases of GDA and KDDA are of exactly the same direction but different in their scale. Thus, GDA works substantially as eigenvalue regularization like ERE. Since the modified eigenvalue of the GDA asymptotically approaches to 1 as $\lambda_{KDDA}$ gets larger, the GDA reduces large $\lambda_{KDDA}$ more than those small $\lambda_{KDDA}$ values.

A main problem with this modified criterion is that the features corresponding to large eigenvalues tend to become less conspicuous. A large eigenvalue of Fisher's criterion can be contributed by two different sources: large between-class separation and small within-class variation. Based on the finding

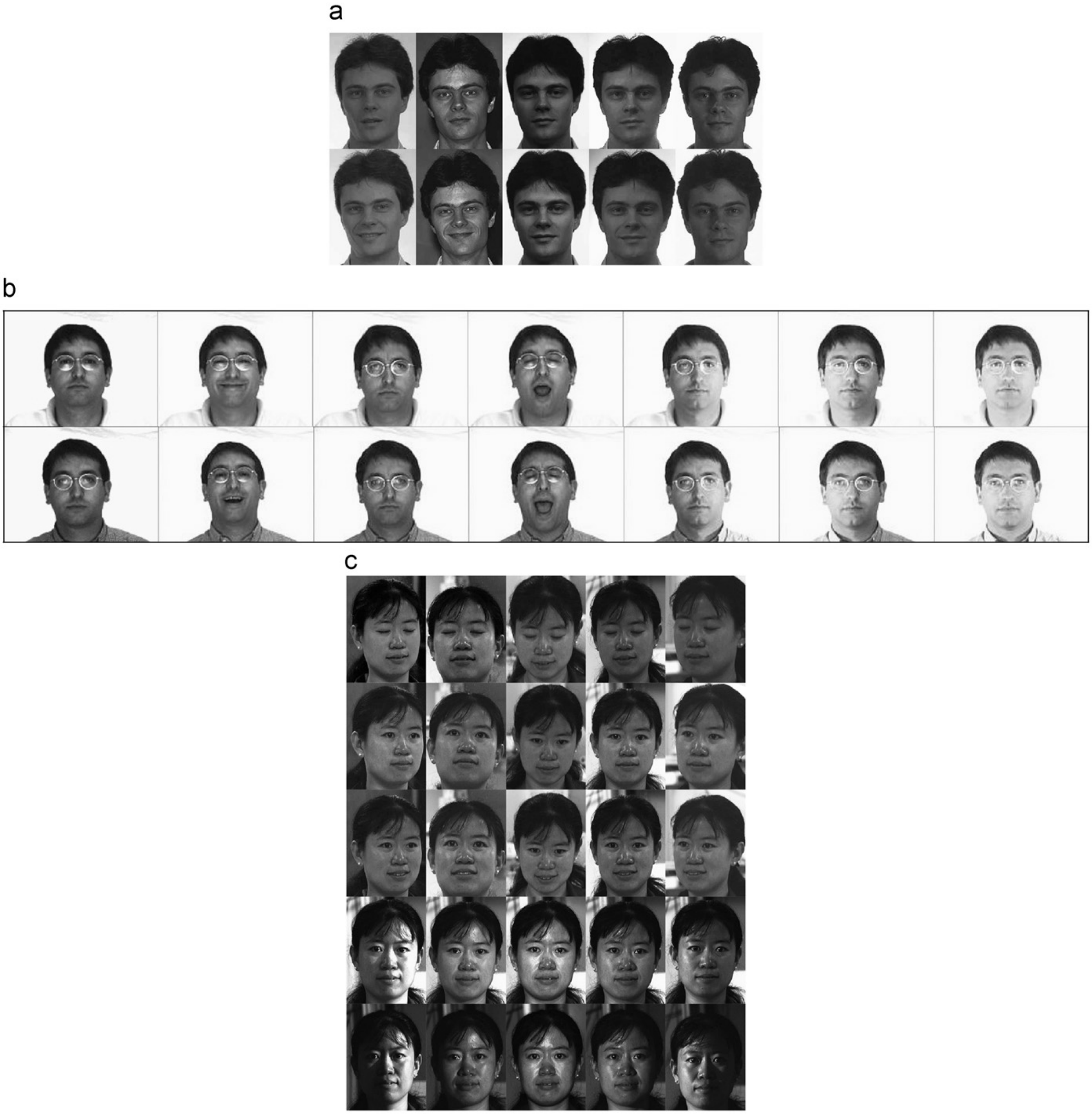

**Fig. 3.** Examples of the databases: (a) FERET database, (b) AR database (2 sessions in each row) and (c) CMU-PIE database.

of Jiang et al. in [23], those small eigenvalues of the within-class scatter might deteriorate more severely due to sample noise, and hence the ER scheme only regularizes small eigenvalues of the within-class scatter matrix. However, GDA regularizes any large eigenvalue of the Fisher's criterion regardless of its original source. Thus, when the between-class scatter is reliable, the regularization of GDA might deemphasize useful feature vectors which comprise of large between-class variation. On the other hand, when the between-class scatter is unreliable, GDA can achieve more generalizable feature extraction than KDDA.

The main peculiarity of statistical analysis in face recognition and other biometrics applications is that the test data mostly comes from unseen class (identity). Thus, any kind of scatter matrix (especially the class-specific ones such as within-class and between-class scatter matrices) is only an estimate of real data correlation. Taking this point into account, we can deduce that collecting more samples can result in a better confidence to

**Table 1**
Summary of the data sets used in experiments.

| | Exp. 1 | Exp. 2 | Exp. 3 |
|---|---|---|---|
| Database | FERET [16] | AR [17] | CMU-PIE [18] |
| No. of images | 1702 | 1680 | 1840 |
| No. of people | 256 | 120 | 68 |
| Avg. no. of images per a person | 6.65 | 14 | 27.06 |

**Table 2**
Main characteristics of compared methods.

| | LDA | ERE | KFD | GDA | KDDA | SVM-DA |
|---|---|---|---|---|---|---|
| $\mathbf{S}_B$ | Between-class scatter | Total-scatter | Kernel between-class scatter | Kernel between-class scatter | Kernel between-class scatter | Pair-wise margin-based between-class scatter |
| Solving inverse problem | Dimension reduction by PCA | Eigenvalue regularized inverse | Inverse of $(\mathbf{S}_W+\tau\mathbf{I})$ | Pseudo-inverse of $\mathbf{S}_T$ discarding small eigenvalues | Pseudo-inverse of $\mathbf{S}_B$ discarding small eigenvalues | Eigenvalue regularized inverse |
| Linearity | Linear | Linear | Nonlinear | Nonlinear | Nonlinear | Nonlinear |

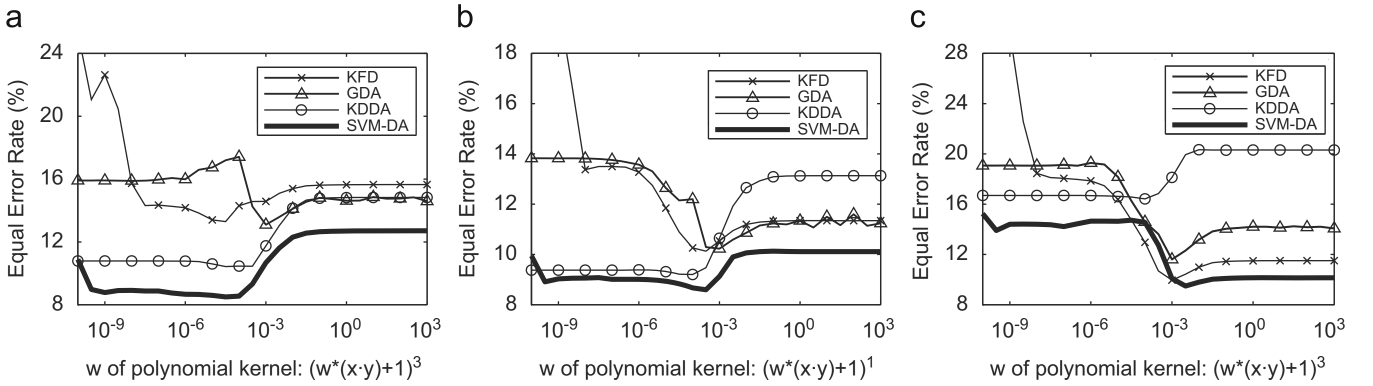

**Fig. 4.** Experimental results of nonlinear methods on (1) FERET, (b) AR, and (c) CMU-PIE Databases varying parameter $w_p$.

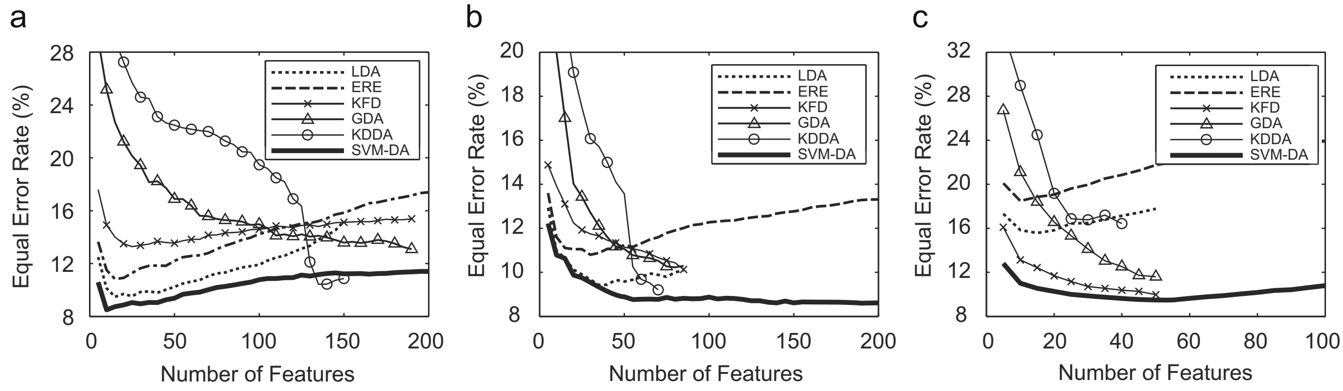

**Fig. 5.** Experimental results on (1) FERET, (b) AR, and (c) CMU-PIE databases varying number of features.

**Table 3**
Best results of all three experiments.

| | LDA | ERE | KFD | GDA | KDDA | SVM-DA |
|---|---|---|---|---|---|---|
| FERET | 9.4833 | 10.8294 | 13.2889 | 13.0916 | 10.4338 | 8.4989 |
| | ($d$=15) | ($d$=15) | ($d$=25, $w_p$ =1e−4.5) | ($d$=190, $w_p$ =1e−3) | ($d$=135, $w_p$ =1e−4.5) | ($d$=10, $w_p$ =1e−4.5) |
| AR | 9.4282 | 10.7965 | 10.1346 | 10.2279 | 9.2039 | 8.5996 |
| | ($d$=35) | ($d$=30) | ($d$=85, $w_p$ =1e−3.5) | ($d$=80, $w_p$ =1e−3) | ($d$=70, $w_p$ =1e−4) | ($d$=190, $w_p$ =1e−3.5) |
| CMU-PIE | 15.5775 | 18.4862 | 9.9675 | 11.6111 | 16.4276 | 9.4900 |
| | ($d$=15) | ($d$=10) | ($d$=50, $w_p$ =1e−3) | ($d$=50, $w_p$ =1e−3) | ($d$=40, $w_p$ =1e−4) | ($d$=55, $w_p$ =1e−2.5) |

estimating the scatter matrices. Likewise, as the between-class scatter takes the class means as its samples, a data set with more pattern classes can make the between-class scatter more reliable and generalizable to unseen class data than that for the case with small number of pattern classes.

The above deduction can be evidenced by comparing the results of GDA and KDDA in the three experiments. As shown in Table 3, KDDA shows a better performance than that of GDA in the FERET experiment. In the AR experiment, the performance gap between the two methods is reduced, and in the CMU-PIE experiment GDA performs better than KDDA. The numbers of classes/identities in the training data sets are 192, 90, and 51 for FERET, AR, and CMU-PIE experiments, respectively (using 4-fold validation). Based on the above observations, the between-class scatter matrices would be more, moderate, and less reliable in the respective experiments. Thus, it is understandable that GDA performs relatively better than KDDA as the number of classes decreases, and vice versa for KDDA.

Regarding the implementation of KFD, there can be two ways to circumvent the noninvertibility (SSS problem) of the within-class scatter matrix: one is to take a pseudo-inverse and the other is to regularize the matrix. As we have discussed earlier, GDA and KDDA take the former way which results in loss of dimensionality. Thus, we took the latter for implementation of KFD as suggested in [10]. Adding a certain scalar $\tau$, which is called the *conditioning coefficient* [10], to the diagonal of $\mathbf{S}_W$, one can introduce an invertible matrix $\tilde{\mathbf{S}}_W$ as

$$\tilde{\mathbf{S}}_W = \mathbf{S}_W + \tau \mathbf{I} \tag{37}$$

This method substantially suppresses the divergence of the eigenvectors corresponding to small eigenvalues when $\tilde{\mathbf{S}}_W$ is inverted, and is actually an even stronger regularization than that of GDA.

However, considering the objective of regularization is to improve the generalization (avoid overfitting) of feature extraction, the regularizations in GDA and KFD are deemed an overkill of data fitting since they does more harm than good when the between-class scatter comprises enough information as seen in the FERET experiment.

Besides regularization, data fitting is also affected by the complexity of the method. There are several factors which can increase the complexity, such as small $k$ in $k$-nearest neighbor ($k$-NN), large number of units in the hidden layer of neural network, small variance of radial basis function (RBF) kernel, high degree of polynomial kernel, and etc. In our experiments, the kernel parameter $w_p$ plays a certain role. Although we fixed the degree of polynomial to three, a large $w_p$ substantially emphasizes those higher order terms and results in having a good fit of the extracted features to training data.

Through the three experimental results as shown in Fig. 4, KDDA, GDA, and KFD show a certain trend: KDDA performs better as $w_p$ gets larger, while GDA and KFD performs in a reverse manner. These trends are the results of data fitting. Since the GDA and KFD features tend to be less fitted to the training data, the under-fitting problem occurs at small $w_p$ values. On the contrary, the overfitting problem occurs in KDDA at large $w_p$ values. In spite of the opposing trajectories, all methods showed their best performance at a certain midway point (around $w_p = 1\mathrm{e}-3$). This is the point where the fitting property and the complexity of the kernel functions are well balanced.

Unlike the other kernel-based methods, the proposed SVM-DA does not show any trend of performance transition as seen in Fig. 4. Apparently, it shows a somewhat adaptive adjustment according to the extent of data fitting. In the FERET experiment where a great extent of data fitting is advantageous, the performance graph of SVM-DA shows a similar trend to that of KDDA.In the CMU-PIE experiment where a strong regularization is advantageous, the performance of SVM-DA approaches the graph of GDA. This adaptability substantially increases the reliability of SVM-DA which is particularly well illustrated in the AR experiment. Although GDA and KDDA show large fluctuation of EER over $w_p$, SVM-DA shows only marginal variation meanwhile maintaining a superior performance over most of the operating range. This is at the mercy of increasing the reliability of the within-class scatter matrix via the ER procedure. However, SVM-DA does not owe its reliability solely to ER. The pair-wise design of the new between-class scatter matrix contributes to the reliability as well. In the CMU-PIE experiment, the number of pattern classes is extremely small and this yields a very unreliable estimate of the between-class scatter. Here, as shown in Table 3, the proposed SVM-DA outperforms all other methods. This result evidences a good reliability of the proposed between-class scatter matrix.

Contrary to the general belief, our experimental results do not suggest that nonlinear methods are more efficient than linear methods under all scenarios. Particularly at small $d$ values, GDA and KDDA failed to find an effective subspace as seen in Fig. 5. Considering the kernel expansion can potentially give a better representation of data, the effectiveness of GDA and KDDA appears controversial. Recall the intrinsic dimension issue discussed in Section 2, an effective subspace should include useful information and discard insignificant information. Since a subspace based on Fisher's criterion includes more and more insignificant information as the dimensionality grows, it is important to pay particular attention to discriminative information at small dimensionality. Particularly, both GDA and KDDA do not prevent loss of information in terms of the Gaussianity and homoscedasticity assumptions. On the other hand, the proposed SVM-DA extracts discrimination information from a thorough SVM search. It is thus not surprising to observe a significantly superior performance of SVM-DA when compared with other LDA variants.

Another problem of GDA and KDDA is that their resulting subspaces carry too much insignificant information. Since the dimensionality of the kernel space can be infinitely large, each and every sample can thus have a significant influence in terms of discrimination. This makes the nonlinear methods sensitive to the sample noise. This explains why GDA, which emphasizes reliability based on regularization, shows a better performance than KDDA at small $d$ in all three experiments (see Fig. 5).

Another noteworthy observation in our experiments is that the conventional LDA consistently outperforms ERE. The only difference between our experiments and the experiments of Jiang et al. [23] is that we exhaustively searched the PCA dimensionality for LDA in the entire possible range whereas Jiang et al. did only in the range from 70% to 100% for the rank of the within-class scatter matrix. Noting that we reported the best results of LDA using the PCA dimensionality around 100–150, the results of LDA reported in [23] were not the best in view of the available dimensions. Nevertheless, this does not imply that LDA always outperforms ERE since our experiment design may not have fully exploited the strength of ERE. The eigenvalue regularization is advantageous when the number of samples per person is small. However, in our experimental setup, the number of classes becomes large at the same time. As shown previously, the between-class scatter of LDA performs better than the total scatter when there are a large number of pattern classes. Hence, our experiments on the face data sets may not be applicable in terms of assessing the generic capabilities of LDA and ERE where the issue remains an open area for discussion.

### 5.3. *Computational complexity*

As mentioned, the proposed SVM-DA includes an additional training phase of SVM classifiers. Although training $c(c-1)/2$ pairs of classifiers seems to be computationally intensive, the resulted computing time is not far beyond most methods. Tables 4 and 5 show the computing time $T$ for training each method using AR database with varying number of classes ($c$) and number of samples per class ($\overline{N}_c$). The computing time for the SVM training

**Table 4**
Computational Time With Varying Number Of Classestime with varying number of classes (s).

| | LDA | ERE | KFD | GDA | KDDA | SVM-DA | SVM phase |
|---|---|---|---|---|---|---|---|
| No. of classes | | | | | | | |
| 15 | 0.41 | 0.61 | 0.25 | 0.27 | 0.03 | 0.73 | 0.50 |
| 30 | 1.60 | 3.96 | 2.04 | 2.15 | 0.26 | 3.87 | 2.00 |
| 60 | 9.92 | 28.26 | 16.56 | 15.17 | 1.33 | 23.09 | 8.06 |
| 90 | 29.63 | 91.49 | 56.43 | 43.27 | 3.63 | 68.962 | 17.99 |
| 120 | 69.99 | 222.71 | 136.73 | 94.85 | 7.67 | 156.64 | 32.12 |
| Growth rate | 2.49 | 2.83 | 3.03 | 2.82 | 2.33 | 2.57 | 2.00 |

**Table 5**
Computational Time With Varying Number Of Samples Per Classtime with varying number of samples per class (s).

| | LDA | ERE | KFD | GDA | KDDA | SVM-DA | SVM phase |
|---|---|---|---|---|---|---|---|
| No. of classes | | | | | | | |
| 15 | 0.58 | 0.91 | 0.33 | 0.38 | 0.17 | 2.10 | 1.13 |
| 30 | 2.35 | 5.95 | 2.68 | 3.01 | 0.47 | 7.68 | 4.05 |
| 60 | 14.47 | 43.62 | 23.52 | 20.85 | 2.07 | 38.20 | 13.82 |
| 90 | 98.55 | 309.47 | 207.15 | 147.08 | 10.67 | 226.28 | 42.08 |
| Growth rate | 2.49 | 2.71 | 3.10 | 2.86 | 2.01 | 2.26 | 1.74 |

phase in SVM-DA is also included in the tables (labeled as "SVM phase"). Since the computational time grows exponentially as $c$ and $\overline{N}_c$ increase, we approximated it using a power growth rate as follows:

$$T(c,\overline{N}_c) = O(c^{g_c} \cdot (\overline{N}_c)^{g_{\overline{N}_c}}) \tag{38}$$

The growth rates, $g_c$ and $g_{\overline{N}_c}$ are estimated from the resulting computing times and these results are shown in the bottom rows of Tables 4 and 5, respectively. Due to the nature of picking up only the support vectors, SVM shows the slowest growth rate, and, owing to this, SVM-DA shows a relatively slow growth rate comparing with other classifiers.

## 6. Conclusion

In this paper, we addressed two inherent issues of Fisher's criterion for face recognition. The first issue of having an underlying homoscedastic assumption of data distribution was addressed by adoption of a margin-based between-class scatter. Essentially, the displacement computation of the class means in the conventional between-class scatter was replaced by a SVM (margin-based) projection which took the structural information of data distribution into account. The second issue of unreliable estimate caused by small training sample size was addressed by eigenspectrum regularization. Here, unlike previous attempts which addressed the problem via data trimming, the regularization process had made use of all available pair-wise data for discrimination analysis. This in turn preserved the necessary ranks for a stable inverse computation, and hence a reliable estimation.

Extensive experiments were performed on several benchmark data sets (FERET, AR, and CMU-PIE) with a comparison of the proposed method to several other related methods such as LDA, KFD, GDA, KDDA and ERE. Our empirical results showed that the proposed method outperformed all the compared methods under a wide range of parameter settings in all datasets that portrayed different data scenarios, thus supporting the claimed effectiveness and reliability.

## Acknowledgments

This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the Biometrics Engineering Research Center (BERC) at Yonsei University (Grant number: R11-2002-105-07003-0(2009)) and by Korea Research Foundation Grant funded by Korean Government (KRF-2008-313-D00774).

## References

[1] M. Turk, A. Pentland, Eigenfaces for recognition, Cognitive Neuroscience 3 (1) (1991) 72–86.

[2] M.S. Bartlett, J.R. Movellan, T.J. Sejnowski, Face recognition by independent component analysis, IEEE Transactions of Neural Networks 13 (6) (2002).

[3] P. Penev, J. Atick, Local feature analysis: a general statistical theory for object representation, Network: Computation in Neural Systems 7 (3) (1996) 477–500.

[4] D.D. Lee, H.S. Seung, Learning the parts of objects by nonnegative matrix factorization, Nature 401 (1999) 788–791.

[5] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (1997) 711–720.

[6] W. Zhao, A. Krishnaswamy, R. Chellappa, D.L. Swets, J. Weng, Discriminant analysis of principal components for face recognition, in: H. Wechsler, P.J. Phillips, V. Bruce, F.F. Soulie, T.S. Huang (Eds.), Face Recognition: From Theory to Applications, Springer-Verlag, Berlin, 1998, pp. 73–85.

[7] R.A. Fisher, The use of multiple measurements in taxonomic problems, Annals of Eugenics 7 (1936) 179–188.

[8] A.K. Jain, R.P.W. Duin, J. Mao, Statistical pattern recognition: a review, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (1) (2000) 4–37.

[9] M. Aizerman, E. Braverman, L. Rozonoer, Theoretical foundations of the potential function method in pattern recognition learning, Automation and Remote Control 25 (1964) 821–837.

[10] J. Ma, J.L. Sancho-Gómez, S.C. Ahalt, Nonlinear multiclass discriminant analysis, IEEE Signal Processing Letters 10 (7) (2003).

[11] G. Baudat, F. Anouar, Generalized discriminant analysis using a kernel approach, Neural Computation 12 (10) (2000) 2385–2404.

[12] J. Lu, K.N. Plataniotis, A. Venetsanopoulos, Face recognition using kernel direct discriminant analysis algorithms, IEEE Transactions of Neural Networks 14 (1) (2003) 117–126.

[13] N. Kumar, A.G. Andreou, Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition, Speech Communication 26 (4) (1998) 283–297.

[14] M. Loog, R.P.W. Duin, Linear dimensionality reduction via a heteroscedastic extension of LDA: the Chernoff criterion, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (6) (2004) 732–739.

[15] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, New York, 1990.

[16] P.J. Phillips, H. Moon, S. Rizvi, P. Rauss, The FERET evaluation methodology for face recognition algorithms, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (10) (2000) 1090–1104.

[17] A.R. Martinez, R. Benavente, The AR face database, Technical Report 24, Computer Vision Center (CVC), June 1998.

[18] T. Sim, S. Baker, M. Bsat, The CMU pose, illumination, and expression database, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (12) (2003) 1615–1618.

[19] W. Zhao, R. Chellappa, P.J. Phillips, Face recognition: a literature survey, ACM Computing Surveys 12 (2003) 399–458.

[20] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery 2 (1998) 121–167.

[21] L. Chen, H. Liao, M. Ko, J. Lin, G. Yu, A new LDA-based face recognition system which can solve the small sample size problem, Pattern Recognition 33 (10) (2000) 1713–1726.

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

[23] X. Jiang, B. Mandal, A. Kot, Eigenfeature regularization and extraction in face recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (3) (2008).

[24] S.-K. Kim, K.-A. Toh, S. Lee, Face recognition incorporating ancillary information, EURASIP Journal on Advances in Signal Processing (2008) (Article ID 312849) 11, doi:10.1155/2008/312849.

[25] C. Lee, D.A. Landgrebe, Feature extraction based on decision boundaries, IEEE Transactions on Pattern Analysis and Machine Intelligence 15 (4) (1993) 388–400.

[26] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer Series in Statistics, Springer, New York, 2001.

[27] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, 2nd ed, John Wiley and Sons, Inc., New York, 2001.

[28] M. Loog, Approximate Pairwise Accuracy Criteria for Multiclass Linear Dimension Reduction: Generalisations of the Fisher Criterion, Delft University Press, 1999.

[29] B. Scholkopf, K. Sung, C. Burges, F. Girosi, P. Niyogi, T. Poggio, V. Vapnik, Comparing support vector machines with Gaussian kernels to radial basis function classifiers, IEEE Transactions on Signal Processing 45 (11) (1997) 2758–2765.

[30] 〈http://www.kernel-machines.org/software〉.

[31] R. Fransens, Jan De Prins, SVM-based nonparametric discriminant analysis, an application to face detection, in: Ninth IEEE International Conference on Computer Vision, vol. 2, October 13–16, 2003.

[32] M. Vatsa, R. Singh, A. Ross, A. Noore, Likelihood ratio in a SVM framework: fusing linear and non-linear face classifiers, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 23–28, 2008.

**About the Author**—SANG-KI KIM received his B.S. degree in Electrical and Electronic Engineering from Yonsei University, Seoul, Korea. Currently, he is a candidate of Ph.D. degree in Electrical and Electronic Engineering from Yonsei University, Seoul, Korea. His research interests include biometrics and pattern recognition.

**About the Author**—YOUNJUNG PARK received her B.S. degree in Electrical and Electronic Engineering from Yonsei University, Seoul, Korea. She is presently a candidate of M.S. degree in Electrical and Electronic Engineering from Yonsei University, Seoul, Korea. Her research interests include pattern recognition, artificial intelligence, image processing and biometrics applications.

**About the Author**—KAR-ANN TOH received his Ph.D. degree from Nanyang Technological University (NTU), Singapore in 1999. He worked for 2 years in the aerospace industry prior to his post-doctoral appointments at research centers in NTU from 1998 to 2002. He was then affiliated with Institute for Infocomm Research, Singapore from 2002 to 2005. Currently, he is a faculty member of the School of Electrical and Electronic Engineering, Yonsei University, Korea. His research interests include biometrics, pattern classification, optimization and neural networks. He has made several PCT filings related to biometric applications, and has actively published his works in the above areas of interest. He has served as a member of technical program committee and as a reviewer for international conferences and journals related to biometrics and pattern recognition.

**About the Author**—SANGYOUN LEE received his B.S. and M.S. degrees in Electronic Engineering from Yonsei University, Seoul, South Korea in 1987, 1989, respectively. He received his Ph.D. degree in Electrical and Computer Engineering from Georgia Tech., Atlanta, Georgia, in 1999. He was a Senior Researcher in Korea Telecom from 1989 to 2004. He is now a faculty member of the School of Electrical and Electronic Engineering, Yonsei University, Korea. His research interests include pattern recognition, computer vision, video coding and biometrics.