Figure 2

Regularization Studies of Linear Discriminant Analysis in Small Sample Size Scenarios with Application to Face Recognition

Juwei Lu, K.N. Plataniotis, A.N. Venetsanopoulos

Bell Canada Multimedia Laboratory

The Edward S. Rogers Sr. Department of Electrical and Computer Engineering

University of Toronto, Toronto, M5S 3G4, ONTARIO, CANADA

Abstract

It is well-known that the applicability of Linear Discriminant Analysis (LDA) to high-dimensional pattern classification tasks such as face recognition often suffers from the so-called "*small sample size*" (SSS) problem arising from the small number of available training samples compared to the dimensionality of the sample space. In this paper, we propose a new LDA method that attempts to address the SSS problem using a regularized Fisher's separability criterion. In addition, a scheme of expanding the representational capacity of face database is introduced to overcome the limitation that the LDA-based algorithms require at least two samples per class available for learning. Extensive experiments performed on the FERET database indicate that the proposed methodology outperforms traditional methods such as Eigenfaces and some recently introduced LDA variants in a number of SSS scenarios.

Key words: Linear Discriminant Analysis, Small Sample Size, Regularization, Face Recognition

1 Introduction

Face recognition (FR) has a wide range of applications, such as face-based video indexing and browsing engines, biometric identity authentication, human-computer interaction, and multimedia monitoring/surveillance. Within the past two decades, numerous FR algorithms have been proposed, and detailed surveys of the developments in the area have appeared in the literature (see e.g. Samal and A.Iyengar, 1992; Valentin et al., 1994; Chellappa et al., 1995; Gong et al., 2000; Turk, 2001; Zhao et al., 2003). Among various FR methodologies used, the most popular are the so-called appearance-based approaches, which include two well-known FR methods, namely, Eigenfaces (Turk and Pentland, 1991) and Fisherfaces (Belhumeur et al., 1997). With focus on low-dimensional statistical feature extraction, the appearance-based approaches generally operate directly on the appearance images of face object and process them as 2D holistic patterns in order to avoid difficulties associated with 3D modelling, and shape or landmark detection (Turk, 2001).

Of the appearance-based FR methods, those utilizing *linear discriminant analysis* (LDA) techniques have shown promising results as it is demonstrated in (Belhumeur et al., 1997; Zhao et al., 1999; Chen et al., 2000; Yu and Yang, 2001; Liu and Wechsler, 2002; Lu et al., 2003a,b; Ye and Li, 2004). However, statistical learning methods including the LDA-based ones often suffer from the so-called "*small-sample-size*" (SSS) problem (Raudys and Jain, 1991), encountered in high-dimensional pattern recognition tasks where the number of training samples available for each subject is smaller than the dimensionality of the sample space. For example, only $L \in [1, 5]$ training samples per subject are available while the dimensionality is up to $J = 17154$ in the FR

experiments reported here. As a result, the sample-based estimation for the between- and within-class scatter matrices is often extremely ill-posed in the application of LDA into FR tasks. Briefly, there are two ways to address the problem. One option is to apply linear algebra techniques to solve the numerical problem of inverting the singular within-class scatter matrix. For example, Tian et al. (1986) utilize the pseudo inverse to complete this task. Also, some researchers (e.g. Hong and Yang, 1991; Zhao et al., 1999) recommended the addition of a small perturbation to the within-class scatter matrix so that it becomes nonsingular. The second option is a subspace approach, such as the one followed in the development of the Fisherfaces method (Belhumeur et al., 1997), where *principal component analysis* (PCA) is firstly used as a pre-processing step to remove the null space of $\mathbf{S}_w$, and then LDA is performed in the lower dimensional PCA subspace. However, it has been shown that the discarded null spaces may contain significant discriminatory information (Liu et al., 1992a,b, 1993). To prevent this from happening, solutions without a separate PCA step, called *direct* LDA (D-LDA) methods have been presented recently in (Chen et al., 2000; Yu and Yang, 2001; Lu et al., 2003b).

The basic premise behind the D-LDA approaches is that the information residing in (or close to) the null space of the within-class scatter matrix is more significant for discriminant tasks than the information out of (or far away from) the null space. Generally, the null space of a matrix is determined by its zero eigenvalues. However, due to insufficient training samples, it is very difficult to identify the true null eigenvalues. As a result, high variance is often introduced in the estimation for the zero (or very small) eigenvalues of the within-class scatter matrix. Note that the eigenvectors corresponding to these eigenvalues are considered to be the most significant feature bases in the

D-LDA approaches (Chen et al., 2000; Yu and Yang, 2001; Lu et al., 2003b).

To overcome the above problem, a new LDA method for FR tasks is proposed in this letter. The LDA method developed here is based on a novel regularized Fisher's discriminant criterion, which is particularly robust against the SSS problem compared to the original one. The purpose of regularization is to reduce the high variance related to the eigenvalue estimates of the within-class scatter matrix at the expense of potentially increased bias. It will be shown that by adjusting the regularization parameter, we can obtain a set of LDA variants, such as the D-LDA of Yu and Yang (2001) (hereafter YD-LDA) and the D-LDA of Lu et al. (2003b) (hereafter JD-LDA). The trade-off between the variance and the bias, depending on the severity of the SSS problem, is controlled by the strength of regularization. Extensive experiments indicate that there exists an optimal regularization solution for the proposed method, which outperforms some existing FR approaches including Eigenfaces, YD-LDA and JD-LDA. In addition, a scheme of expanding the representational capacity of face database is introduced to overcome a known limitation of the LDA style learning methods, which require at least two samples per class available for training. Furthermore, experimentation shows that the scheme also enhances the overall FR performance of the proposed LDA method.

2 Methods

2.1 The Small-Sample-Size (SSS) Problem

Given a training set, $\mathbf{Z} = \{\mathbf{Z}_i\}_{i=1}^{C}$, containing C classes with each class $\mathbf{Z}_i = \{\mathbf{z}_{ij}\}_{j=1}^{C_i}$ consisting of a number of localized face images $\mathbf{z}_{ij}$, a total

of $N = \sum_{i=1}^{C} C_i$ face images are available in the set. For computational convenience, each image is represented as a column vector of length $J(= I_w \times I_h)$ by lexicographic ordering of the pixel elements, *i.e.* $\mathbf{z}_{ij} \in \mathbb{R}^J$, where $(I_w \times I_h)$ is the image size, and $\mathbb{R}^J$ denotes the J-dimensional real space.

LDA finds a set of $M(\ll J)$ feature basis vectors, denoted as $\{\psi_m\}_{m=1}^{M}$, in such a way that the ratio of the between- and within-class scatters of the training sample is maximized (Fisher, 1936). The maximization problem is generally formulated as:

$$\Psi = \arg\max_{\Psi} \frac{\left|\Psi^T \mathbf{S}_b \Psi\right|}{\left|\Psi^T \mathbf{S}_w \Psi\right|}, \quad \Psi = [\psi_1, \cdots, \psi_M], \quad \psi_m \in \mathbb{R}^J \tag{1}$$

where $\mathbf{S}_b$ and $\mathbf{S}_w$ are the between- and within-class scatter matrices, having the following expressions,

$$\mathbf{S}_b = \frac{1}{N} \sum_{i=1}^{C} C_i (\bar{\mathbf{z}}_i - \bar{\mathbf{z}})(\bar{\mathbf{z}}_i - \bar{\mathbf{z}})^T = \sum_{i=1}^{C} \Phi_{b,i} \Phi_{b,i}^T = \Phi_b \Phi_b^T \tag{2}$$

$$\mathbf{S}_w = \frac{1}{N} \sum_{i=1}^{C} \sum_{j=1}^{C_i} (\mathbf{z}_{ij} - \bar{\mathbf{z}}_i)(\mathbf{z}_{ij} - \bar{\mathbf{z}}_i)^T \tag{3}$$

where $\Phi_{b,i} = (C_i/N)^{1/2}(\bar{\mathbf{z}}_i - \bar{\mathbf{z}})$, $\Phi_b = [\Phi_{b,1}, \cdots, \Phi_{b,c}]$, and $\bar{\mathbf{z}}_i = \frac{1}{C_i} \sum_{j=1}^{C_i} \mathbf{z}_{ij}$ is the mean of the class $\mathbf{Z}_i$. The optimization problem of Eq.1 is equivalent to the following generalized eigenvalue problem,

$$\mathbf{S}_b \psi_m = \lambda_m \mathbf{S}_w \psi_m, \quad m = 1 \cdots M \tag{4}$$

Thus, when $\mathbf{S}_w$ is non-singular, the basis vectors Ψ sought in Eq.1 correspond to the first M most significant eigenvectors of $(\mathbf{S}_w^{-1}\mathbf{S}_b)$, where the "significant" means that the eigenvalues corresponding to these eigenvectors are the first M largest ones. Due to the SSS problem, often a extremely degenerated $\mathbf{S}_w$

is generated in FR tasks. Let us assume that $\mathbf{A}$ and $\mathbf{B}$ represent the null spaces of $\mathbf{S}_b$ and $\mathbf{S}_w$ respectively, while $\mathbf{A}' = \mathbb{R}^J - \mathbf{A}$ and $\mathbf{B}' = \mathbb{R}^J - \mathbf{B}$ denote the orthogonal complements of $\mathbf{A}$ and $\mathbf{B}$. Traditional methods, for example Fisherfaces (Belhumeur et al., 1997), attempt to solve the problem by utilizing an intermediate PCA step to remove $\mathbf{A}$ and $\mathbf{B}$. Nevertheless, it should be noted at this point that the maximum of the ratio in Eq.5 can be reached only when $\left|\Psi^T \mathbf{S}_w \Psi\right| = 0$ and $\left|\Psi^T \mathbf{S}_b \Psi\right| \neq 0$. This means that the discarded null space $\mathbf{B}$ may contain the most significant discriminatory information. On the other hand, there is no significant information, in terms of the maximization in Eq.5, to be lost if $\mathbf{A}$ is discarded. It is not difficult to see at this point that when $\Psi \in \mathbf{A}$, the ratio $\frac{\left|\Psi^T \mathbf{S}_b \Psi\right|}{\left|\Psi^T \mathbf{S}_w \Psi\right|}$ drops to its minimum value, 0. Therefore, many researchers (e.g. Liu et al., 1992a,b, 1993; Chen et al., 2000) consider the intersection space ($\mathbf{A}' \cap \mathbf{B}$) to be spanned by the optimal discriminant feature bases.

Based on the above principle, Yu and Yang (2001) proposed the so-called direct LDA (YD-LDA) approach in order to prevent the removal of useful discriminant information contained in the null space $\mathbf{B}$. However, it has been recently found that the YD-LDA performance may deteriorate rapidly when the SSS problem becomes severe (Lu et al., 2003c). The deterioration should be attributed to the influence of the two factors, variance and bias. Firstly, it is well-known that the $\mathbf{S}_w$ estimate based on Eq.3 produces biased estimates of the eigenvalues. As a result, the largest ones are biased high and the smallest ones are biased toward values that are too low. Secondly, the estimate of the null space $\mathbf{B}$ can be highly unstable, giving rise to high variance. Both the variance and biasing degrees are determined by the degree of the SSS problem. A relevant method developed by Friedman (1989) in similar situations is the

regularized quadratic discriminant analysis, where each sample class covariance matrix estimate $\mathbf{S}_i$ could be highly ill-posed. The solution proposed by Friedman (1989) is to introduce a regularization term, which is a multiple of the identity matrix, $\gamma \cdot \mathbf{I}$, so as to have $\mathbf{S}_i = \mathbf{S}_i + \gamma\mathbf{I}$, where γ is the regularization parameter and $\mathbf{I}$ is the identity matrix. Such a regularization has the effect of decreasing the larger eigenvalues and increasing the smaller ones, thereby counteracting the biasing. Another effect of the regularization is to stabilize the smallest eigenvalues. Furthermore, it should be noted that the within-class scatter matrix $\mathbf{S}_w$ considered here is equivalent to the average of the individual class covariance matrices $\mathbf{S}_i$, *i.e.* $\mathbf{S}_w = \frac{1}{C}\sum_{i=1}^{C}\mathbf{S}_i$. This encourages us to conceive a similar solution to handle the SSS situations that the D-LDA type methods may encounter.

2.2 A Regularized Fisher's Discriminant Criterion

Motivated by the success of Friedman (1989), a variant of D-LDA is developed here by introducing a regularized Fisher's criterion, which can be expressed as follows:

$$\Psi = \arg\max_{\Psi} \frac{\left|\Psi^T\mathbf{S}_b\Psi\right|}{\left|\eta(\Psi^T\mathbf{S}_b\Psi) + (\Psi^T\mathbf{S}_w\Psi)\right|} \tag{5}$$

where $0 \leq \eta \leq 1$ is a regularization parameter. Although Eq.5 looks quite different from the conventional Fisher's criterion (Eq.1), it can be shown that they are exactly equivalent by the following theorem.

Theorem 1 *Let $\mathbb{R}^J$ denote the J-dimensional real space, and suppose that $\forall\psi \in \mathbb{R}^J$, $u(\psi) \geq 0$, $v(\psi) \geq 0$, $u(\psi) + v(\psi) > 0$ and $0 \leq \eta \leq 1$. Let $q_1(\psi) = \frac{u(\psi)}{v(\psi)}$ and $q_2(\psi) = \frac{u(\psi)}{\eta\cdot u(\psi)+v(\psi)}$. Then, $q_1(\psi)$ has the maximum (including*

positive infinity) at point $\psi^* \in \mathbb{R}^J$ iff $q_2(\psi)$ *has the maximum at point* ψ^*.

PROOF. Since $u(\psi) \geq 0$, $v(\psi) \geq 0$ and $0 \leq \eta \leq 1$, we have $0 \leq q_1(\psi) \leq +\infty$ and $0 \leq q_2(\psi) \leq \frac{1}{\eta}$.

(1) If $\eta = 0$, then $q_1(\psi) = q_2(\psi)$.

(2) If $0 < \eta \leq 1$ and $v(\psi) = 0$, then $q_1(\psi) = +\infty$ and $q_2(\psi) = 1/\eta$.

(3) If $0 < \eta \leq 1$ and $v(\psi) > 0$, then

$$q_2(\psi) = \frac{u(\psi)/v(\psi)}{1 + \eta u(\psi)/v(\psi)} = \frac{q_1(\psi)}{1 + \eta q_1(\psi)} = \frac{1}{\eta}\left(1 - \frac{1}{1 + \eta q_1(\psi)}\right).$$

It can be seen that in this case, $q_2(\psi)$ increases iff $q_1(\psi)$ increases.

Combining $(1) - (3)$, we have the theorem.

The modified Fisher's criterion is a function of the parameter η, which controls the strength of regularization. Within the variation range of η, two extremes should be noted. In one extreme where $\eta = 0$, the modified Fisher's criterion is reduced to the conventional one with no regularization. In contrast with this, rather strong regularization is introduced in another extreme where $\eta = 1$. In this case, Eq.5 becomes $\Psi = \arg\max_{\Psi} \frac{\left|\Psi^T \mathbf{S}_b \Psi\right|}{\left|\Psi^T (\mathbf{S}_b + \mathbf{S}_w) \Psi\right|}$, which as a variant of the original Fisher's criterion has been also widely used for example in the D-LDA method (JD-LDA) of Lu et al. (2003b) and others (see e.g. Liu et al., 1992a,b, 1993; Chen et al., 2000; Lu et al., 2003a). The advantages of introducing the regularization strategy will be seen during the development of the new LDA algorithm proposed below.

2.3 A Regularized LDA: R-LDA

In this work, we propose a *regularized* LDA (hereafter R-LDA) method, which attempts to optimize the regularized Fisher's criterion of Eq.5. The R-LDA method follows the D-LDA process of Yu and Yang (2001) and Lu et al. (2003b). To this end, we first solve the complement space of $\mathbf{S}_b$, $\mathbf{A}'$. Let $\mathbf{U}_m = [u_1, \cdots, u_m]$ be the eigenvectors of $\mathbf{S}_b$ corresponding to its first m largest nonzero eigenvalues Λ_b, where $m \leq C-1$. The complement space $\mathbf{A}'$ is spanned by $\mathbf{U}_m$, which is furthermore scaled by $\mathbf{H} = \mathbf{U}_m \Lambda_b^{-1/2}$ so as to have $\mathbf{H}^T \mathbf{S}_b \mathbf{H} = \mathbf{I}$, where $\mathbf{I}$ is the $(m \times m)$ identity matrix. In this way, it can be seen that the denominator of Eq.5 is naturally transformed to the regularization expression of Friedman's style, $\eta\mathbf{I} + \mathbf{H}^T \mathbf{S}_w \mathbf{H}$, in the subspace spanned by $\mathbf{H}$. We then seek a set of feature bases, which minimizes the regularized denominator. It is not difficult to see that the sought feature bases correspond to the $M(\leq m)$ eigenvectors of $\mathbf{H}^T \mathbf{S}_w \mathbf{H}$, $\mathbf{P}_M = [\mathbf{p}_1, \cdots, \mathbf{p}_M]$, with the smallest eigenvalues Λ_w. Combining these results, we can obtain the sought solution, $\Psi = \mathbf{H}\mathbf{P}_M(\eta\mathbf{I} + \Lambda_w)^{-1/2}$, which is considered a set of optimal discriminant feature basis vectors. The detailed process to implement the R-LDA method is depicted in Fig.1.

It can be seen from Fig.1 that R-LDA reduces to YD-LDA and JD-LDA when $\eta = 0$ and $\eta = 1$, respectively. Varying the values of η within $[0, 1]$ leads to a set of intermediate D-LDA variants between YD-LDA and JD-LDA. Since the subspace spanned by Ψ may contain the intersection space $(\mathbf{A}' \cap \mathbf{B})$, it is possible that there exist zero or very small eigenvalues in Λ_w, which have been shown to be high variance for estimation in the SSS environments (Friedman, 1989). As a result, any bias arising from the eigenvectors corresponding to

Input: A training set $\mathbf{Z}$ with C classes: $\mathbf{Z} = \{\mathbf{Z}_i\}_{i=1}^{C}$, each class containing $\mathbf{Z}_i = \{\mathbf{z}_{ij}\}_{j=1}^{C_i}$ face images, where $\mathbf{z}_{ij} \in \mathbb{R}^J$, and the regularization parameter η.

Output: An M-dimensional LDA subspace spanned by Ψ, a $J \times M$ matrix with $M \ll J$.

Algorithm:

Step 1. Express $\mathbf{S}_b = \Phi_b \Phi_b^T$, with $\Phi_b = [\Phi_{b,1}, \cdots, \Phi_{b,c}]$, $\Phi_{b,i} = (C_i/N)^{1/2}(\bar{\mathbf{z}}_i - \bar{\mathbf{z}})$, $\bar{\mathbf{z}}_i = 1/C_i \sum_{j=1}^{C_i} \mathbf{z}_{ij}$, and $\bar{\mathbf{z}} = 1/N \sum_{i=1}^{C} \sum_{j=1}^{C_i} \mathbf{z}_{ij}$.

Step 2. Find the m eigenvectors of $\Phi_b^T \Phi_b$ with non-zero eigenvalues, and denote them as $\mathbf{E}_m = [\mathbf{e}_1, \cdots, \mathbf{e}_m]$.

Step 3. Calculate the first m most significant eigenvectors ($\mathbf{U}_m$) of $\mathbf{S}_b$ and their corresponding eigenvalues (Λ_b) by $\mathbf{U}_m = \Phi_b \mathbf{E}_m$ and $\Lambda_b = \mathbf{U}_m^T \mathbf{S}_b \mathbf{U}_m$.

Step 4. Let $\mathbf{H} = \mathbf{U}_m \Lambda_b^{-1/2}$. Find eigenvectors of $\mathbf{H}^T \mathbf{S}_w \mathbf{H}$, $\mathbf{P} = [\mathbf{p}_1, \cdots, \mathbf{p}_m]$ sorted in increasing eigenvalue order.

Step 5. Choose the first $M(\leq m)$ eigenvectors in $\mathbf{P}$. Let $\mathbf{P}_M$ and Λ_w be the chosen eigenvectors and their corresponding eigenvalues, respectively.

Step 6. Return $\Psi = \mathbf{H}\mathbf{P}_M(\eta\mathbf{I} + \Lambda_w)^{-1/2}$.

Fig. 1. The pseudo code implementation of the R-LDA method

these eigenvalues is dramatically exaggerated due to the normalization process ($\mathbf{P}_M \Lambda_w^{-1/2}$). Against the effect, the introduction of the regularization helps to decrease the importance of these highly unstable eigenvectors, thereby re-

ducing the overall variance. Also, there may exist the zero eigenvalues in Λ_w, which are used as divisors in YD-LDA due to $\eta = 0$ so that the YD-LDA process can not be carried out. However, it is not difficult see that the problem can be avoided in the R-LDA solution, $\Psi = \mathbf{H}\mathbf{P}_M(\eta\mathbf{I} + \Lambda_w)^{-1/2}$, simply by setting the regularization parameter $\eta > 0$.

3 Discussion: A Different Viewpoint to the SSS Problem

The works described above are attempting to solve the SSS problem from the viewpoint of improving existing LDA algorithms. On the other hand, the problem can be addressed by expanding the representational capacity of the available training database. For example, given a pair of prototype images belonging to the same class, Li and Lu (1999) proposed a linear model, called the *nearest feature line* (NFL), to virtually generalize an infinite number of variants of the two prototypes under variations in illumination and expression. However, like LDA, the NFL method requires at least two training samples per subject to be available. To deal with the extreme case where only one training image per subject is available, Huang et al. (2003) recently proposed a method, which constructs more samples by rotating and translating the prototype image. Nevertheless, the method introduces bias inevitably when face recognition is performed on a set of well-aligned face images for example along with the centers of the eyes as did in the experiments reported here.

To avoid the bias, an alternative approach may be the use of the mirrored versions of the available training samples. Based on the symmetrical property of face object, intuitively it is reasonable to consider the mirrored view of a face image to be a real and bias-free sample of the face pattern. In this way, the

size of the training set can be doubled. In addition, the mirrored version of any test sample can also be utilized to enhance the performance of a FR system. For example, we can verify the classification result of a given query using its mirror. A recognition process is accepted only when the query and its mirror are given the same class label, otherwise the query is rejected to recognition. More sophisticated rules to combine the results from multiple classifiers can be found in (Kittler et al., 1998), but such a development is beyond the scope of this letter.

4 Experimental Results

4.1 The FR Evaluation Design

A set of experiments are included in the paper to assess the performance of the proposed R-LDA method. To show the high complexity of the face patterns' distribution, a medium-size subset of the FERET database (Phillips et al., 2000) is used in the experiments. The subset consists of 1147 gray-scale images of 120 people, each one having at least 6 samples so that we can generalize a set of SSS learning tasks. These images as depicted in Table 1 cover a wide range of variations in illumination, facial expression/details, and pose angles. We follow the preprocessing sequence recommended by Phillips et al. (2000), which includes four steps: (1) images are translated, rotated and scaled (to size 150×130) so that the centers of the eyes are placed on specific pixels; (2) a standard mask as shown in Fig.2:Middle is applied to remove the nonface portions; (3) histogram equalization is performed in the masked facial pixels; (4) face data are further normalized to have zero mean

and unit standard deviation. Fig.2:Right and Fig.3 depict some examples after the preprocessing sequence is applied. For computational requirement, each image is finally represented as a column vector of length $J = 17154$ prior to the recognition stage.

Table 1

The number of images divided into the standard FERET imagery categories, and the pose angle, α (degree), of each category.

Ct.	fa	fb	ba	bj	bk	ql	qr	rb	rc
No.	567	338	5	5	5	68	65	32	62
α	0	0	0	0	0	-22.5	+22.5	10	-10

Fig. 2. **Left**: Original samples in the FERET database. **Middle**: The standard mask. **Right**: The samples after the preprocessing sequence.

Fig. 3. Some samples of eight people come from the normalized FERET evaluation database.

The SSS problem is defined in terms of the number of available training sam-

ples per subject, L. Thus the value of L has a significant influence on the required strength of regularization. To study the sensitivity of the performance, in terms of *correct recognition rate* (CRR), to L, five tests were performed with various L values ranging from $L = 1$ to $L = 5$. For a particular L, the FERET subset is randomly partitioned into three datasets: a training set, a validation set and a test set. The training set is composed of $(L \times 120)$ samples: L images per person were randomly chosen. The validation set is composed of (2×120) samples: 2 images per person were randomly chosen. The remaining $(1147 - L \times 120 - 2 \times 120)$ images are used to form the test set. There is no overlapping between the three. To enhance the accuracy of the assessment, five runs of such a partition were executed, and all of the experimental results reported below have been averaged over the five runs.

4.2 CRR Performance with Varying Regularization Parameter

The first experiment is designed to test the CRR performance of R-LDA with varying regularization parameter in various SSS scenarios. To this end, the R-LDA method is applied to a testing grid of (η, M) values, defined by the outer product of $\eta = [10^{-4} : 0.01 : 1]$ and $M = [20 : 1 : 119]$, where the expression $[b_1 : b_2 : b_3]$ denotes a spaced vector consisting of $round((b_3 - b_1)/b_2)$ elements from b_1 to b_3 with step b_2, and η is initiated from 10^{-4} instead of zero to avoid numerical singularities in $(\mathbf{H}^T \mathbf{S}_w \mathbf{H})$. For every pair of (η, M) values in the grid, R-LDA is first trained with the training set. Since there is no requirement for parameter selection in this experiment, the learned R-LDA(η, M) machine is then directly applied to an evaluation dataset consisting of the validation and test sets. The CRRs obtained by R-LDA(η, M) on the combined evaluation

set are depicted in Fig.4.

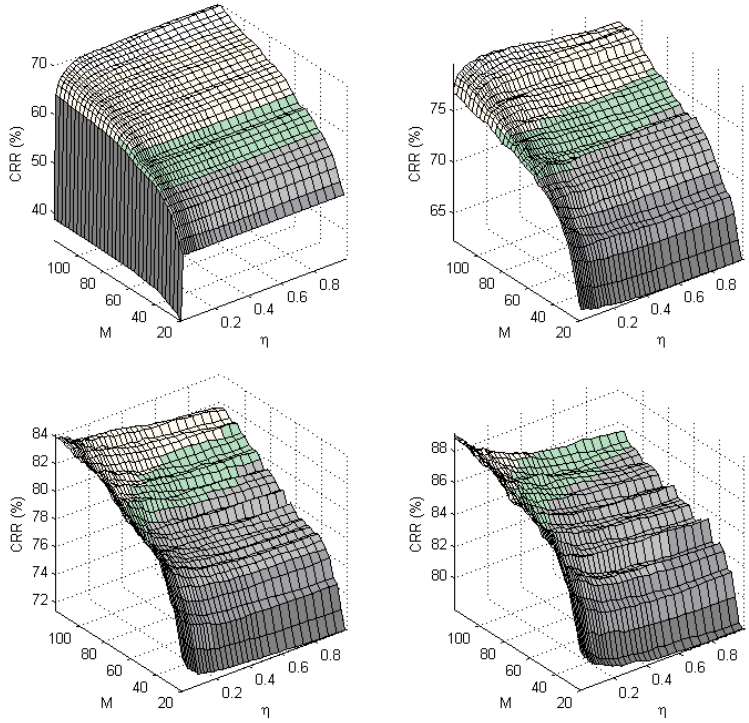

Fig. 4. CRRs obtained by R-LDA as a function of (M, η). **Top**: $L = 2, 3$; **Bottom**: $L = 4, 5$, where L is the number of training samples per subject.

The parameter η controls the strength of regularization, which balances the tradeoff between variance and bias in the estimation for the zero or small eigenvalues of the within-class scatter matrix. Varying the η values within $[0, 1]$ leads to a set of intermediate LDA variants between YD-LDA and JD-LDA. In theory, YD-LDA with no extra bias introduced through η should be the best performer among these variants if sufficient training samples are available. It can be observed at this point from Fig.4 that the CRR peaks gradually moved from the right side ($\eta = 1$) toward the left side ($\eta = 0$) that is the case of YD-LDA as L increases. Small values of η have been good enough for the regularization requirement in many cases ($L \geq 4$) as shown in Fig.4. However, it also can be seen from Fig.4 that YD-LDA performed poorly when $L = 2, 3$. This should be attributed to the high variance in the estimate of $\mathbf{S}_w$ due to

insufficient training samples. In these cases, even $\mathbf{H}^T\mathbf{S}_w\mathbf{H}$ is singular or close to singular, and the resulting effect is to dramatically exaggerate the importance associated with the eigenvectors corresponding to the smallest eigenvalues. Against the effect, the introduction of regularization helps to decrease the larger eigenvalues and increase the smaller ones, thereby counteracting for some extent the bias. This is also the reason why JD-LDA outperforms YD-LDA when L is small.

4.3 Quantitative Comparison with Other FR Methods

To further study the performance of the R-LDA method, we conducted a more strict experiment for a quantitative comparison among R-LDA, YD-LDA and JD-LDA in this section. The Eigenfaces method (Turk and Pentland, 1991) was also implemented to provide a performance baseline. For all the four methods compared here, the CRR is a function of the number of extracted feature vectors, M, and the number of available training examples per subject, L. In addition, R-LDA's performance depends critically on the regularization parameter, η. It has been shown by last experiment that R-LDA is capable of outperforming both YD-LDA and JD-LDA. However, it should be noted that the performance improvement is subject to the selection of the parameters (η, M). Thus, to make a fair comparison, the parameter selection process should be included in the experiment. To this end, we take advantages of the three splits: the training, validation and test sets. Each method compared here is first trained on the training set to generalize a set of models with various parameter configurations, for example, all the possible M values for Eigenfaces, YD-LDA and JD-LDA, and the (η, M) grid described in Section 4.2

for R-LDA. These models are evaluated on the validation set, and then the best found model is applied to the test set. The CRRs obtained by the four methods on the test set are reported in Table 2, where M^* and η^* denote the parameter values corresponding to the best configuration determined by using the validation set.

Table 2

Comparison of the CRRs (%) obtained on the test set and their corresponding parameter values without using mirrored samples. L is the number of training samples per subject.

$L =$	1	2	3	4	5
Eigenfaces	46.48	57.96	65.19	65.81	65.26
(M^*)	117	145	217	287	405
YD-LDA	–	17.42	75.69	83.61	88.73
(M^*)	–	114	116	108	106
JD-LDA	–	69.60	76.71	81.17	85.26
(M^*)	–	116	116	117	112
R-LDA	–	69.66	78.10	83.47	88.98
(M^*)	–	116	119	112	114
(η^*)	–	0.983	0.24	0.048	10^{-4}

From Table 2, it can be clearly seen that R-LDA is the top performer amongst all the methods compared here. Also, the characteristics of the three LDA-based methods are demonstrated again in this experiment. YD-LDA showed

excellent performance on one side ($L = 5$) of the SSS settings but failed on the other side ($L = 2$), while JD-LDA performed on the contrary. In contrast with this, by introducing the adjustable regularization parameter, R-LDA systematically combines the strengths of YD-LDA and JD-LDA while at the same time overcomes their shortcomings and limitations. On the other hand, it should be noted that compared to other methods, the determination of the optimal parameter values (M^*, η^*) for R-LDA is computationally demanding as it is based on an exhaust search in the preset (η, M) grid by using the validation set. Nevertheless, some heuristics may be applied to reduce the grid size. For example, it is not difficult to see that the optimal regularization parameter η^* decreases monotonously as the number of training samples per subject, L, increases. It seems that the relationship is not linear. This is the reason why the values of the best found η^* in Table 2 appear to be $[0.983, 0.24, 0.048, 10^{-4}]$ corresponding to $L = [2, 3, 4, 5]$. Also it should be noted that small values of η have been good enough for the regularization requirement in the cases of $L \geq 4$. In addition to L, our recent experiments indicated that the η^* value increases as the number of subjects, C, increases. It is similar at this point to the learning capacity of the LDA machines, which is generally considered to be directly proportional to the number of training samples per subject L, while reciprocally proportional to the number of the training subjects C. Therefore, further exploring the mathematical relationship among these parameters L, C, η, and the training/generalization errors may be an interesting future research direction to reveal the nature of discriminant learning under small sample size scenarios.

The LDA based algorithms require at least two training samples for each class. However, with the introduction of the mirrored training samples, it becomes possible to overcome the limitation. In this experiment, R-LDA is trained with a combined set consisting of the training samples and their mirrors. Same to the experiment described in Section 4.3, the model parameters are determined by using the validation set. The R-LDA classifier with the best found parameters is then applied to the test set in three ways. The resulting CRRs are depicted in Table 3, where R-LDA$^{\mathrm{m1}}$ and R-LDA$^{\mathrm{m2}}$ correspond to the results obtained by using the original test set and its mirrored version, respectively, while R-LDA$^{\mathrm{m3}}$ denotes the results from a combination of the two sets. As introduced in Section 3, a recognition process is accepted by R-LDA$^{\mathrm{m3}}$ only when the test sample and its mirror are identified as belonging to the same subject, otherwise the test sample is rejected to recognition.

Not surprisingly, as it can be seen from Table 3, R-LDA$^{\mathrm{m1}}$ and R-LDA$^{\mathrm{m2}}$ have similar performance. This means that to recognize a test sample, we can use either the sample or its mirror. Compared to R-LDA in Table 2, the performance improvement achieved by R-LDA$^{\mathrm{m1}}$ and R-LDA$^{\mathrm{m2}}$ is up to approximately 2% in average over the range $L = 2 \sim 5$. On the other hand, the reject rates obtained in R-LDA$^{\mathrm{m3}}$ indicate that the recognition is incorrect in most cases when the sample and its mirror are given different class labels. Therefore, compared to R-LDA$^{\mathrm{m1}}$ and R-LDA$^{\mathrm{m2}}$, an additional CRR improvement of approximately 2% in average over $L = 1 \sim 5$ is obtained by R-LDA$^{\mathrm{m3}}$. These results shown in Table 3 demonstrate that the mirrors of face images provide not only additional training samples, but also complemental

Table 3

Comparison of the CRRs (%) with the corresponding parameter values obtained by R-LDA on the test set using different mirror schemes.

$L =$	1	2	3	4	5
R-LDA$^{\mathrm{m1}}$	56.01	70.25	80.18	86.84	90.59
(M^*)	119	117	115	114	108
(η^*)	0.3	0.40	0.068	10^{-4}	10^{-4}
R-LDA$^{\mathrm{m2}}$	56.01	70.64	81.50	86.89	90.28
(M^*)	119	119	119	114	112
(η^*)	0.38	0.53	0.016	10^{-4}	10^{-4}
R-LDA$^{\mathrm{m3}}$	59.57	73.11	82.26	88.19	91.07
Reject Rate	7.85	6.21	3.14	2.06	0.87
(M^*)	119	117	115	112	107
(η^*)	1	1	0.064	10^{-4}	10^{-4}

information, which is useful to enhance the generalization performance of a LDA-based FR system.

5 Conclusions and Future Works

A new LDA method for face recognition has been introduced in this paper. The proposed method is based on a novel regularized Fisher's discriminant criterion, which is particularly robust against the SSS problem compared to

the traditional one used in LDA. It has been also shown that a series of traditional LDA variants including the recently introduced YD-LDA and JD-LDA can be derived from the proposed R-LDA framework by adjusting the regularization parameter. Also, a scheme to double the size of face databases is introduced, so that R-LDA can be carried out in the extreme case where only one training sample available for each subject. The effectiveness of the proposed method has been demonstrated through experimentation using the FERET database.

Our future work will concentrate on a continuing improvement of the R-LDA algorithm. Firstly, as discussed before, an immediate direction is to seek a fast and cost-effective parameter optimization method instead of the exhaust search. However, such a research is rather difficult, due to some unknown facts, for example,

(i) What is the actual distribution of the patterns?
(ii) How have the training data sampled the underlying distribution of the patterns?

Classical parameter estimation schemes such as leave-one-out may not work well, since the estimation will experience high variance as it is under small sample considerations. Alternatively, a further study on the mathematical relations between the number of training samples per subject L, the number of subjects C, the regularization parameter η, and the classification error seems more promising. Also, a kernel version of R-LDA is straightforward to develop a more general R-LDA framework, which is able to deal with both linear and nonlinear problems.

Acknowledgments

The authors would like to thank the reviewers, who have made interesting and constructive comments that resulted in an improved paper.

Portions of the research in this paper use the FERET database of facial images collected under the FERET program (Phillips et al., 1998). The authors would like to thank the FERET Technical Agent, the U.S. National Institute of Standards and Technology (NIST) for providing the FERET database.

This work is partially supported by the Bell University Laboratories at the University of Toronto.

References

Belhumeur, P. N., Hespanha, J. P., Kriegman, D. J., 1997. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7), 711–720.

Chellappa, R., Wilson, C., Sirohey, S., 1995. Human and machine recognition of faces: A survey. The Proceedings of the IEEE 83, 705–740.

Chen, L.-F., Liao, H.-Y. M., Ko, M.-T., Lin, J.-C., Yu, G.-J., 2000. A new LDA-based face recognition system which can solve the small sample size problem. Pattern Recognition 33, 1713–1726.

Fisher, R., 1936. The use of multiple measures in taxonomic problems. Ann. Eugenics 7, 179–188.

Friedman, J. H., 1989. Regularized discriminant analysis. Journal of the American Statistical Association 84, 165–175.

Gong, S., McKenna, S. J., Psarrou, A., May 2000. Dynamic Vision From Images to Face Recognition. Imperial College Press, World Scientific Publishing.

Hong, Z.-Q., Yang, J.-Y., 1991. Optimal discriminant plane for a small number of samples and design method of classifier on the plane. Pattern Recognition 24 (4), 317–324.

Huang, J., Yuen, P. C., Chen, W.-S., Lai, J. H., October 2003. Component-based lda method for face recognition with one training sample. In: Proceedings of the IEEE International Workshop on Analysis and Modeling of Faces and Gestures. Nice, France.

Kittler, J., Hatef, M., Duin, R. P., Matas, J., March 1998. On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (3), 226–239.

Li, S. Z., Lu, J., March 1999. Face recognition using the nearest feature line method. IEEE Transactions on Neural Networks 10, 439–443.

Liu, C., Wechsler, H., April 2002. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Transactions on Image Processing 11 (4), 467–476.

Liu, K., Cheng, Y., Yang, J., 1992a. A generalized optimal set of discriminant vectors. Pattern Recognition 25, 731–739.

Liu, K., Cheng, Y., Yang, J., 1993. Algebraic feature extraction for image recognition based on an optimal discriminant criterion. Pattern Recognition 26, 903–911.

Liu, K., Cheng, Y., Yang, J., Liu, X., 1992b. An efficient algorithm for foley-sammon optimal set of discriminant vectors by algebraic method. Int. J. Pattern Recog. Artif. Intell. 6, 817–829.

Lu, J., Plataniotis, K., Venetsanopoulos, A., January 2003a. Face recognition using kernel direct discriminant analysis algorithms. IEEE Transactions on Neural Networks 14 (1), 117–126.

Lu, J., Plataniotis, K., Venetsanopoulos, A., January 2003b. Face recognition using LDA based algorithms. IEEE Transactions on Neural Networks 14 (1), 195–200.

Lu, J., Plataniotis, K., Venetsanopoulos, A., December 2003c. Regularized discriminant analysis for the small sample size problem in face recognition. Pattern Recognition Letter 24 (16), 3079–3087.

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

Phillips, P. J., Wechsler, H., Huang, J., Rauss, P., 1998. The FERET database and evaluation procedure for face recognition algorithms. Image and Vision

Computing J 16 (5), 295–306.

Raudys, S. J., Jain, A. K., 1991. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (3), 252–264.

Samal, A., A.Iyengar, P., 1992. Automatic recognition and analysis of human faces and facial expressions: A survey. Pattern Recognition 25, 65–77.

Tian, Q., Barbero, M., Gu, Z., Lee, S., 1986. Image classification by the foley-sammon transform. Opt. Eng. 25 (7), 834–840.

Turk, M., December 2001. A random walk through eigenspace. IEICE Trans. Inf. & Syst. E84-D (12), 1586–1695.

Turk, M. A., Pentland, A. P., 1991. Eigenfaces for recognition. Journal of Cognitive Neuroscience 3 (1), 71–86.

Valentin, D., Alice, H. A., Toole, J. O., Cottrell, G. W., 1994. Connectionist models of face processing: A survey. Pattern Recognition 27 (9), 1209–1230.

Ye, J., Li, Q., April 2004. LDA/QR: an efficient and effective dimension reduction algorithm and its theoretical foundation. Pattern Recognition 37 (4), 851–854.

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

Zhao, W., Chellappa, R., Phillips, J., 1999. Subspace linear discriminant analysis for face recognition. Technical Report, CS-TR4009, University of Maryland .

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