Brief Papers

Face Recognition Using LDA-Based Algorithms

Juwei Lu, Kostantinos N. Plataniotis, and Anastasios N. Venetsanopoulos

***Abstract*—Low-dimensional feature representation with enhanced discriminatory power is of paramount importance to face recognition (FR) systems. Most of traditional linear discriminant analysis (LDA)-based methods suffer from the disadvantage that their optimality criteria are not directly related to the classification ability of the obtained feature representation. Moreover, their classification accuracy is affected by the "small sample size" (SSS) problem which is often encountered in FR tasks. In this short paper, we propose a new algorithm that deals with both of the shortcomings in an efficient and cost effective manner. The proposed here method is compared, in terms of classification accuracy, to other commonly used FR methods on two face databases. Results indicate that the performance of the proposed method is overall superior to those of traditional FR approaches, such as the Eigenfaces, Fisherfaces, and D-LDA methods.**

***Index Terms*—Direct LDA, Eigenfaces, face recognition, Fisherfaces, fractional-step LDA, linear discriminant analysis (LDA), principle component analysis (PCA).**

I. Introduction

FEATURE selection for face representation is one of central issues to face recognition (FR) systems. Among various solutions to the problem (see [1], [2] for a survey), the most successful seems to be those appearance-based approaches, which generally operate directly on images or appearances of face objects and process the images as two-dimensional (2-D) holistic patterns, to avoid difficulties associated with three-dimensional (3-D) modeling, and shape or landmark detection [2]. Principle component analysis (PCA) and linear discriminant analysis (LDA) are two powerful tools used for data reduction and feature extraction in the appearance-based approaches. Two state-of-the-art FR methods, Eigenfaces [3] and Fisherfaces [4], built on the two techniques, respectively, have been proved to be very successful.

It is generally believed that, when it comes to solving problems of pattern classification, LDA-based algorithms outperform PCA-based ones, since the former optimizes the low-dimensional representation of the objects with focus on the most discriminant feature extraction while the latter achieves simply object reconstruction [4]–[6]. However, the classification performance of traditional LDA is often degraded by the fact that their separability criteria are not directly related to their classification accuracy in the output space [7]. A solution to the problem is to introduce weighting functions into LDA. Object classes that are closer together in the output space, and thus can potentially result in misclassification, should be more heavily weighted in the input space. This idea has been further extended in [7] with the introduction of the fractional-step linear discriminant analysis algorithm (F-LDA), where the dimensionality reduction is implemented in a few small fractional steps allowing for the relevant distances to be more accurately weighted. Although the method has been successfully tested on low-dimensional patterns whose dimensionality is $D \leq 5$, it cannot be directly applied to high-dimensional patterns, such as those face images used in this paper [it should be noted at this point that a typical image pattern of size (112×92) (Fig. 2) results to a vector of dimension $D = 10304$], due to two factors: 1) the computational difficulty of the eigen-decomposition of matrices in the high-dimensional image space; 2) the degenerated scatter matrices caused by the so-called "small sample size" (SSS) problem, which widely exists in the FR tasks where the number of training samples is smaller than the dimensionality of the samples [4]–[6].

The traditional solution to the SSS problem requires the incorporation of a PCA step into the LDA framework. In this approach, PCA is used as a preprocessing step for dimensionality reduction so as to discard the null space of the within-class scatter matrix of the training data set. Then LDA is performed in the lower dimensional PCA subspace [4]. However, it has been shown that the discarded null space may contain significant discriminatory information [5], [6]. To prevent this from happening, solutions without a separate PCA step, called direct LDA (D-LDA) methods have been presented recently [5], [6]. In the D-LDA framework, data are processed directly in the original high-dimensional input space avoiding the loss of significant discriminatory information due to the PCA preprocessing step.

In this paper, we introduce a new feature representation method for FR tasks. The method combines the strengths of the D-LDA and F-LDA approaches, while at the same time overcomes their shortcomings and limitations. In the proposed framework, hereafter DF-LDA, we first lower the dimensionality of the original input space by introducing a new variant of D-LDA that results in a low-dimensional SSS-free subspace where the most discriminatory features are preserved. The variant of D-LDA developed here utilizes a modified Fisher's criterion to avoid a problem resulting from the wage of the zero eigenvalues of the within-class scatter matrix as possible divisors in [6]. Also, a weighting function is introduced into the proposed variant of D-LDA, so that a subsequent F-LDA

Manuscript received January 15, 2001; revised April 16, 2002.

The authors are with Multimedia Laboratory, Edward S. Rogers, Sr. Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4, Canada (e-mail: kostas@dsp.toronto.edu).

Digital Object Identifier 10.1109/TNN.2002.806647

step can be applied to carefully reorient the SSS-free subspace resulting in a set of optimal discriminant features for face representation.

II. Direct Fractional-Step LDA (DF-LDA)

The problem of low-dimensional feature representation in FR systems can be stated as follows. Given a set of L training face images $\{\mathbf{z}_i\}_{i=1}^{L}$, each of which is represented as a vector of length $N(= I_w \times I_h)$, i.e., $\mathbf{z}_i \in \mathbb{R}^N$ belonging to one of C classes $\{\mathbf{Z}_i\}_{i=1}^{C}$, where $(I_w \times I_h)$ is the image size and $\mathbb{R}^N$ denotes a N-dimensional real space, the objective is to find a transformation φ, based on optimization of certain separability criteria, to produce a representation $\mathbf{y}_i = \varphi(\mathbf{z}_i)$, where $\mathbf{y}_i \in \mathbb{R}^M$ with $M \ll N$. The representation $\mathbf{y}_i$ should enhance the separability of the different face objects under consideration.

A. Where are the Optimal Discriminant Features?

Let $\mathbf{S}_{\mathrm{BTW}}$ and $\mathbf{S}_{\mathrm{WTH}}$ denote the between- and within-class scatter matrices of the training image set, respectively. LDA-like approaches such as the Fisherface method [4] find a set of basis vectors, denoted by Ψ that maximizes the ratio between $\mathbf{S}_{\mathrm{BTW}}$ and $\mathbf{S}_{\mathrm{WTH}}$

$$\Psi = \arg\max_{\Psi} \frac{\left|\left(\Psi^T \mathbf{S}_{\mathrm{BTW}} \Psi\right)\right|}{\left|\left(\Psi^T \mathbf{S}_{\mathrm{WTH}} \Psi\right)\right|}. \tag{1}$$

Assuming that $\mathbf{S}_{\mathrm{WTH}}$ is nonsingular, the basis vectors Ψ correspond to the first M eigenvectors with the largest eigenvalues of $(\mathbf{S}_{\mathrm{WTH}}^{-1}\mathbf{S}_{\mathrm{BTW}})$. The M-dimensional representation is then obtained by projecting the original face images onto the subspace spanned by the M eigenvectors. However, a degenerated $\mathbf{S}_{\mathrm{WTH}}$ in (1) may be generated due to the SSS problem widely existing in most FR tasks. It was noted in the introduction that a possible solution is to apply a PCA step in order to remove the null space of $\mathbf{S}_{\mathrm{WTH}}$ prior to the maximization in (1). Nevertheless, it recently has been shown that the null space of $\mathbf{S}_{\mathrm{WTH}}$ may contain significant discriminatory information [5], [6]. As a consequence, some of significant discriminatory information may be lost due to this preprocessing PCA step.

The basic premise of the D-LDA methods that attempt to solve the SSS problem without a PCA step is, that the null space of $\mathbf{S}_{\mathrm{WTH}}$ contains significant discriminant information if the projection of $\mathbf{S}_{\mathrm{BTW}}$ is not zero in that direction, and that no significant information will be lost if the null space of $\mathbf{S}_{\mathrm{BTW}}$ is discarded. Assuming that $\mathcal{A}$ and $\mathcal{B}$ represent the null space of $\mathbf{S}_{\mathrm{BTW}}$ and $\mathbf{S}_{\mathrm{WTH}}$, while $\mathcal{A}' = \mathbb{R}^N - \mathcal{A}$ and $\mathcal{B}' = \mathbb{R}^N - \mathcal{B}$ are the complement spaces of $\mathcal{A}$ and $\mathcal{B}$, respectively, the optimal discriminant subspace sought by D-LDA is the intersection space $(\mathcal{A}' \cap \mathcal{B})$. The method in [6] first diagonalizes $\mathbf{S}_{\mathrm{BTW}}$ to find $\mathcal{A}'$ when seek the solution of (1), while [5] diagonalizes $\mathbf{S}_{\mathrm{WTH}}$ to find $\mathcal{B}$. Although it appears that the two methods are not significantly different, it may be intractable to calculate $\mathcal{B}$ when the size of $\mathbf{S}_{\mathrm{WTH}}$ is large, which is the case in most FR applications. For example, a typical face pattern of (112×92) results to $\mathbf{S}_{\mathrm{WTH}}$ and $\mathbf{S}_{\mathrm{BTW}}$ matrices with dimensionality $(10\,304 \times 10\,304)$. Fortunately, the rank of $\mathbf{S}_{\mathrm{BTW}}$ is determined by $\mathrm{rank}(\mathbf{S}_{\mathrm{BTW}}) = \min(N, C-1)$, with C the number of image classes, which is usually a small value in most of FR tasks, e.g., $C = 40$ in the ORL database, resulting in $\mathrm{rank}(\mathbf{S}_{\mathrm{BTW}}) = 39$. $\mathcal{A}'$ can be easily found by solving eigenvectors of a (39×39) matrix rather than the original $(10\,304 \times 10\,304)$ matrix through an algebraic transformation [3], [6]. Then $(\mathcal{A}' \cap \mathcal{B})$ can be obtained by solving the null space of projection of $\mathbf{S}_{\mathrm{WTH}}$ into $\mathcal{A}'$, while the projection is a small matrix of size (39×39).

Based on the analysis given above, it can be known that the most significant discriminant information exist in the intersection subspace $(\mathcal{A}' \cap \mathcal{B})$, which is usually low-dimensional so that it becomes possible to further apply some sophisticated techniques, such as the rotation strategy of the LDA subspace used in F-LDA, to derive the optimal discriminant features from the intersection subspace.

B. Variant of D-LDA

The maximization process in (1) is not directly linked to the classification error which is the criterion of performance used to measure the success of the FR procedure. Modified versions of the method, such as the F-LDA approach, use a weighting function in the input space, to penalize those classes that are close and can potentially lead to misclassifications in the output space. Thus, the weighted between-class scatter matrix can be expressed as:

$$\hat{\mathbf{S}}_{\mathrm{BTW}} = \sum_{i=1}^{C} \phi_i \phi_i^T \tag{2}$$

where $\phi_i = (L_i/L)^{1/2} \sum_{j=1}^{C} (w(d_{ij}))^{1/2} (\overline{\mathbf{z}}_i - \overline{\mathbf{z}}_j)$, $\overline{\mathbf{z}}_i$ is the mean of class $\mathbf{Z}_i$, L_i is the number of elements in $\mathbf{Z}_i$, and $d_{ij} = \|\overline{\mathbf{z}}_i - \overline{\mathbf{z}}_j\|$ is the Euclidean distance between the means of class i and class j. The weighting function $w(d_{ij})$ is a monotonically decreasing function of the distance d_{ij}. The only constraint is that the weight should drop faster than the Euclidean distance between the means of class i and class j with the authors in [7] recommending weighting functions of the form $w(d_{ij}) = (d_{ij})^{-2p}$ with $p = 2, 3, \ldots$.

Most LDA based algorithms including Fisherfaces [4] and D-LDA [6] utilize the conventional Fisher's criterion denoted by (1). In this work we propose the utilization of a variant of the conventional metric. The proposed metric can be expressed as follows:

$$\Psi = \arg\max_{\Psi} \frac{\left|\left(\Psi^T \hat{\mathbf{S}}_{BTW} \Psi\right)\right|}{\left|\left(\Psi^T \mathbf{S}_{TOT} \Psi\right)\right|} \tag{3}$$

where $\mathbf{S}_{\mathrm{TOT}} = \mathbf{S}_{\mathrm{WTH}} + \hat{\mathbf{S}}_{\mathrm{BTW}}$, and $\hat{\mathbf{S}}_{\mathrm{BTW}}$ is the weighted between-class scatter matrix defined in (2). This modified Fisher's criterion can be proven to be equivalent to the conventional one by introducing the analysis of [11] where it was shown that in $\mathbb{R}^N$ $\forall x \in \mathbb{R}^N$, if $f(x) \geq 0$, $g(x) > 0$ and $f(x) + g(x) > 0$, and $h_1(x) = f(x)/g(x)$, $h_2(x) = f(x)/(f(x) + g(x))$, the function $h_1(x)$ has the maximum (including positive infinity) at point $x_0 \in \mathbb{R}^N$ iff $h_2(x)$ has the maximum at point x_0.

For the reasons explained in Section II-A, we start by solving the eigenvalue problem of $\hat{\mathbf{S}}_{\mathrm{BTW}}$. It is intractable to directly compute eigenvectors of $\hat{\mathbf{S}}_{\mathrm{BTW}}$ which is a large size $(N \times N)$ matrix. Fortunately, the first $m\,(\leq C-1)$ most significant eigenvectors of $\hat{\mathbf{S}}_{\mathrm{BTW}}$, which correspond to nonzero eigenvalues,

Input: A set of training face images $\{\mathbf{z}_i\}_{i=1}^{L}$, each of which is represented as a N-dimensional vector.

Output: A low-dimensional representation $\mathbf{y}$ of $\mathbf{z}$ with enhanced discriminatory power, after a transformation $\mathbf{y} = \varphi(\mathbf{z})$.

Algorithm:

Step 1. Calculate those eigenvectors of $\Phi_b^T\Phi_b$ with non-zero eigenvalues: $\mathbf{E}_m = [e_1 \dots e_m]$, where $m \leq C-1$ and Φ_b is from $\hat{\mathbf{S}}_{BTW} = \Phi_b\Phi_b^T$.

Step 2. Calculate the first m most significant eigenvectors and their corresponding eigenvalues of $\hat{\mathbf{S}}_{BTW}$ by $\mathbf{V} = \Phi_b\mathbf{E}_m$ and $\Lambda_b = \mathbf{V}^T\hat{\mathbf{S}}_{BTW}\mathbf{V}$.

Step 3. Let $\mathbf{U} = \mathbf{V}\Lambda_b^{-1/2}$. Calculate eigenvectors of $\mathbf{U}^T\mathbf{S}_{TOT}\mathbf{U}$, $\mathbf{P}$.

Step 4. Optionally discard those eigenvectors in $\mathbf{P}$ with the largest eigenvalues. Let $\mathbf{P}_{M'}$ and Λ_w be the $M'(\leq m)$ selected eigenvectors and their corresponding eigenvalues.

Step 5. Map all face images $\{\mathbf{z}_i\}_{i=1}^{L}$ to the M'-dimensional subspace spanned by $\Gamma = \mathbf{U}\mathbf{P}_{M'}\Lambda_w^{-1/2}$, and have $\{\mathbf{x}_i\}_{i=1}^{L}$, where $\mathbf{x}_i = \Gamma^T\mathbf{z}_i$.

Step 6. Further reduce the dimensionality of $\mathbf{x}_i$ from M' to M by performing a F-LDA on $\{\mathbf{x}_i\}_{i=1}^{L}$, and let W(size $M' \times M$) be the bases of the output space.

Step 7. The optimal discriminant feature representation of $\mathbf{z}$ can be obtained by $\mathbf{y} = \varphi(\mathbf{z}) = (\Gamma W)^T\mathbf{z}$.

Fig. 1. Pseudocode for the computation of the DF-LDA algorithm.

can be indirectly derived from the eigenvectors of the matrix $(\Phi_b^T\Phi_b)$ with size $(C \times C)$, where $\Phi_b = [\phi_1 \cdots \phi_c]$ [3]. Let λ_i and $\mathbf{e}_i$ be the ith eigenvalue and its corresponding eigenvector of $(\Phi_b^T\Phi_b)$, $i = 1 \cdots C$, sorted in *decreasing* eigenvalue order. Since $(\Phi_b\Phi_b^T)(\Phi_b\mathbf{e}_i) = \lambda_i(\Phi_b\mathbf{e}_i)$, $\mathbf{v}_i = \Phi_b\mathbf{e}_i$ is the eigenvector of $\hat{\mathbf{S}}_{\mathrm{BTW}}$.

To remove the null space of $\hat{\mathbf{S}}_{\mathrm{BTW}}$, the first m $(\leq C-1)$ eigenvectors: $\mathbf{V} = [\mathbf{v}_1 \cdots \mathbf{v}_m] = \Phi_b\mathbf{E}_m$, whose corresponding eigenvalues are greater than 0, are used, where $\mathbf{E}_m = [\mathbf{e}_1 \cdots \mathbf{e}_m]$. It is not difficult to see that $\mathbf{V}^T\hat{\mathbf{S}}_{\mathrm{BTW}}\mathbf{V} = \Lambda_b$, with $\Lambda_b = \mathrm{diag}[\lambda_1^2 \cdots \lambda_m^2]$, a $(m \times m)$ diagonal matrix. Let $\mathbf{U} = \mathbf{V}\Lambda_b^{-1/2}$. Projecting $\hat{\mathbf{S}}_{\mathrm{BTW}}$ and $\mathbf{S}_{\mathrm{TOT}}$ into the subspace spanned by $\mathbf{U}$, we have $\mathbf{U}^T\hat{\mathbf{S}}_{\mathrm{BTW}}\mathbf{U} = \mathbf{I}$ and $\mathbf{U}^T\mathbf{S}_{\mathrm{TOT}}\mathbf{U}$. Then, we diagonalize $\mathbf{U}^T\mathbf{S}_{\mathrm{TOT}}\mathbf{U}$ which is a tractable matrix with size $(m \times m)$. Let $\mathbf{p}_i$ be the ith eigenvector of $\mathbf{U}^T\mathbf{S}_{\mathrm{TOT}}\mathbf{U}$, where $i = 1 \cdots m$, sorted in *increasing* order according to corresponding eigenvalues λ_i'. In the set of ordered eigenvectors, those that correspond to the smallest eigenvalues maximize the ratio in (1) and they should be considered as the most discriminatory features. We can discard the eigenvectors with the largest eigenvalues, and denote the $M'(\leq m)$ selected eigenvectors as $\mathbf{P} = [\mathbf{p}_1 \cdots \mathbf{p}_{M'}]$. Defining a matrix $\mathbf{Q} = \mathbf{U}\mathbf{P}$, we can obtain $\mathbf{Q}^T\mathbf{S}_{\mathrm{TOT}}\mathbf{Q} = \Lambda_w$, with $\Lambda_w = \mathrm{diag}[\lambda_1' \cdots \lambda_{M'}']$, a $(M' \times M')$ diagonal matrix.

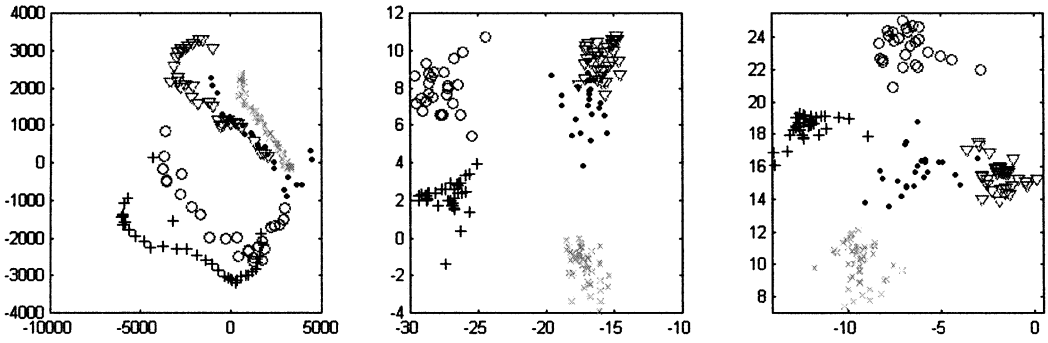
Based on the derivation presented above, a set of optimal discriminant feature basis vectors can be derived through $\Gamma = \mathbf{Q}\Lambda_w^{-1/2}$. To facilitate comparison, it should be mentioned at this point that the D-LDA method of [6] uses the conventional Fisher's criterion of (1) with $\mathbf{S}_{\mathrm{TOT}}$ replaced by $\mathbf{S}_{\mathrm{WTH}}$. However, since the subspace spanned by Γ contains the intersection space $(\mathcal{A}' \cap \mathcal{B})$, it is possible that there exist zero eigenvalues in Λ_w. To prevent this from happening, a heuristic threshold was introduced in [6]. A small threshold value ϵ was set and any value below ϵ was adjusted to ϵ. Obviously, performance heavily depends on the proper choice of the value for the artificial threshold ϵ, which is done in a heuristic manner [6]. Unlike the method in [6], due to the modified Fisher's criterion of (3), the nonsingularity of $\Lambda_w = \mathbf{Q}^T\mathbf{S}_{\mathrm{TOT}}\mathbf{Q}$ can be guaranteed by the following lemma.

Lemma 1: Suppose $\mathbf{B}$ is a real matrix of size $(N \times N)$. Furthermore, let us assume that it can be represented as $\mathbf{B} = \Phi\Phi^T$ where Φ is a real matrix of size $(N \times M)$. Then, the matrix $(\mathbf{I} + \mathbf{B})$ is positive definite, i.e., $\mathbf{I} + \mathbf{B} > 0$, where $\mathbf{I}$ is the $(N \times N)$ identity matrix.

Proof: Since $B^T = B$, $I + B$ is a real symmetric matrix. Let x be any $N \times 1$ nonzero real vector, we have $x^T(I+B)x = x^Tx + x^TBx = x^Tx + (\Phi^Tx)^T(\Phi^Tx) > 0$. According to [12], the matrix $I + B$ that satisfies the above condition is positive definite, i.e., $I + B > 0$. ■

Similar to $\hat{\mathbf{S}}_{\mathrm{BTW}}$, $\mathbf{S}_{\mathrm{WTH}}$ can be expressed as $\mathbf{S}_{\mathrm{WTH}} = \Phi_w\Phi_w^T$, and then $\mathbf{U}^T\mathbf{S}_{\mathrm{WTH}}\mathbf{U} = (\mathbf{U}^T\Phi_w)(\mathbf{U}^T\Phi_w)^T$. Since $\mathbf{U}^T\hat{\mathbf{S}}_{\mathrm{BTW}}\mathbf{U} = \mathbf{I}$ and $(\mathbf{U}^T\mathbf{S}_{\mathrm{WTH}}\mathbf{U})$ is real symmetric it can be easily seen that $(\mathbf{U}^T\mathbf{S}_{\mathrm{TOT}}\mathbf{U})$ is positive definite, and thus $\Lambda_w = \mathbf{Q}^T\mathbf{S}_{\mathrm{TOT}}\mathbf{Q}$ is nonsingular.

Fig. 2. Some sample images of three persons randomly chosen from the two databases. (Left): ORL. (Rright): UMIST.



(a) (b) (c)

Fig. 3. Distribution of 170 face images of five subjects (classes) randomly selected from the UMIST database in (a) PCA-based subspace, (b) D-LDA-based subspace, and (c) DF-LDA-based subspace.

C. Rotation and Reorientation of the D-LDA Subspace

Through the enhanced D-LDA step discussed above, a low-dimensional SSS-free subspace spanned by Γ has been derived without losing the most important, for discrimination purposes, information. In this subspace, $\mathbf{S}_{\mathrm{TOT}}$ is nonsingular and has been whitened due to $\Gamma^T\mathbf{S}_{\mathrm{TOT}}\Gamma = \mathbf{I}$. Thus, an F-LDA step can be safely applied to further reduce the dimensionality from M' to the required M now.

To this end, we firstly project the original face images into the M'-dimensional subspace, obtaining a representation $\mathbf{x}_i = \Gamma^T\mathbf{z}_i$ where $i = 1, 2, \ldots, L$. Let $\mathbf{S}_b$ be the between-class scatter matrix of $\{\mathbf{x}_i\}_{i=1}^{L}$, and $\gamma_{M'}$ be the M'th eigenvector of $\mathbf{S}_b$ which corresponds to the smallest eigenvalue of $\mathbf{S}_b$. This eigenvector will be discarded when dimensionality is reduced from M' to $(M'-1)$. A problem may be encountered during the dimensionality reduction procedure. If classes $\mathbf{Z}_i$ and $\mathbf{Z}_j$ are well separated in the M'-dimensional input space, this will produce a very small $w(d_{ij})$. As a result, the two classes may heavily overlap in the $(M'-1)$-dimensional output space which is orthogonal to $\gamma_{M'}$. To avoid the problem, a kind of "automatic gain control" is introduced to the weighting procedure in F-LDA [7], where dimensionality is reduced from M' to $(M'-1)$ at $r \geq 1$ fractional steps instead of one step directly. In each step, $\mathbf{S}_b$ and its eigenvectors are recomputed based on the changes of $w(d_{ij})$ in the output space, so that the $(M'-1)$-dimensional subspace is reoriented and severe overlap between classes in the output space is avoided. $\gamma_{M'}$ will not be discarded until r iterations are done.

It should be noted at this point that the approach of [7] has only been applied in small dimensionality pattern spaces. To the best of the author's knowledge the work reported here constitutes the first attempt to introduce fractional reorientation in a realistic application involving large dimensionality spaces. This becomes possible due to the integrated structure of the DF-LDA algorithm, the pseudocode implementation of which can be found in Fig. 1.

The effect of the above rotation strategy of the D-LDA subspace is illustrated in Fig. 3, where the first two most significant features of each image extracted by PCA, D-LDA (the variant proposed in Section II-B) and DF-LDA, respectively, are visualized. The PCA-based representation shown in Fig. 3(a) is optimal in terms of image reconstruction, thereby provides some insight on the original structure of image distribution, which is highly complex and nonseparable. Although the separability of subjects is greatly improved in the D-LDA-based subspace, some classes still overlap as shown in Fig. 3(b). It can be seen from Fig. 3(c) that the separability is further enhanced, and different classes tend to be equally spaced after a few fractional (reorientation) steps.

III. Experimental Results

Two popular face databases, the ORL [8] and the UMIST [13], are used to demonstrate the effectiveness of the proposed DF-LDA framework. The ORL database contains 40 distinct persons with ten images per person. The images are taken at different time instances, with varying lighting conditions, facial expressions and facial details (glasses/no glasses). All persons are in the upright, frontal position, with tolerance for some side movement. The UMIST repository is a multiview database, consisting of 575 images of 20 people, each covering a wide range of poses from profile to frontal views. Fig. 2 depicts some samples contained in the two databases, where each image is scaled into (112×92), resulting in an input dimensionality of $N = 10\,304$.

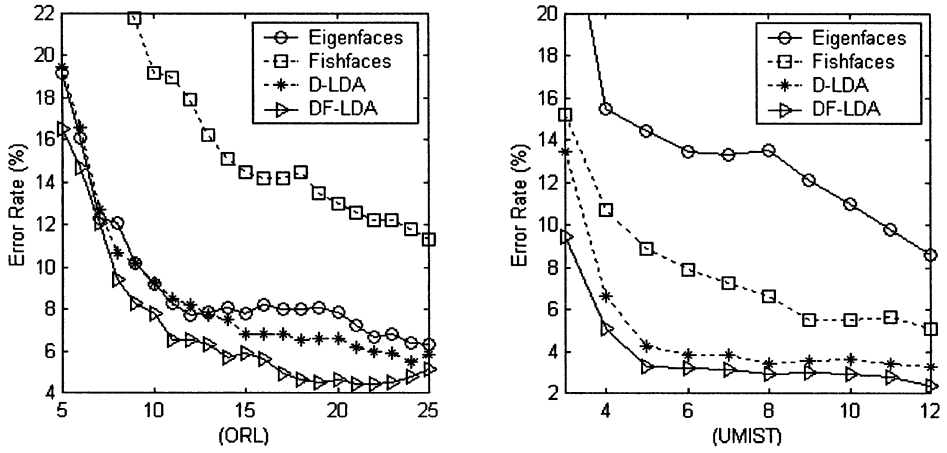


Fig. 4. Comparison of error rates obtained by the four FR methods as functions of the number of feature vectors, where $w(d) = d^{-12}$ is used in DF-LDA for the ORL, $w(d) = d^{-8}$ for the UMIST, and $r = 20$ for both.

To start the FR experiments, each one of the two databases is randomly partitioned into a training set and a test set with no overlap between the two. The partition of the ORL database is done following the recommendation of [14], [15] which call for five images per person randomly chosen for training, and the other five for testing. Thus, a training set of 200 images and a test set with 200 images are created. For the UMIST database, eight images per person are randomly chosen to produce a training set of 160 images. The remaining 415 images are used to form the test set. In the following experiments, the figures of merit are error rates averaged over five runs (four runs in [14] and three runs in [15]), each run being performed on such random partitions in the two databases. It is worthy to mention here that both experimental setups introduce SSS conditions since the number of training samples are in both cases much smaller than the dimensionality of the input space. Also, we do have observed some partition cases, where zero eigenvalues occurred in Λ_w as discussed in Section II-B. In these cases, in contrast with the failure of D-LDA [6], DF-LDA was still able to perform well.

In addition to D-LDA [6], DF-LDA is compared against two popular feature selection methods, namely: Eigenfaces [3] and Fisherfaces [4]. For each of the four methods, the FR procedure consists of 1) a feature extraction step where four kinds of feature representation of each training or test sample are extracted by projecting the sample onto the four feature spaces generalized by Eigenface, Fisherface, D-LDA, and DF-LDA, respectively, and 2) a classification step in which each feature representation obtained in the first step is fed into a simple nearest neighbor classifier. It should be noted at this point that, since the focus in this short paper is on feature extraction, a very simple classifier, namely nearest neighbor, is used in step 2). We anticipate that the classification accuracy of all four methods compared here will improve if a more sophisticated classifier is used instead of the nearest neighbor. However, such an experiment is beyond the scope of this short paper.

The error rate curves obtained for the four methods are shown in Fig. 4 as functions of the number of feature vectors. The number of fractional steps used in DF-LDA is $r = 20$ and the weighted function utilized is $w(d) = d^{-8}$. From Fig. 4, it can be seen that the performance of DF-LDA is overall superior to that of the other three methods on both databases. Let α_i and β_i be the error rates of the DF-LDA and one of the other three methods respectively, where i is the number of feature vectors. We can obtain the average percentage of the error rate of DF-LDA over that of the other methods by $\mathcal{E}_{\text{orl}} = \sum_{i=5}^{25} (\alpha_i/\beta_i)$ for the ORL database and $\mathcal{E}_{\text{umist}} = \sum_{i=3}^{12} (\alpha_i/\beta_i)$ for the UMIST database. The results summarized in Table I indicate that the average error rate of DF-LDA is approximately 50.5%, 43% and 80% of that of Eigenface, Fisherface and D-LDA, respectively. It is of interest to observe the performance of Eigenfaces vs that of Fisherfaces. Not surprisingly, Eigenfaces outperform Fisherfaces in the ORL database, because Fisherfaces may lost significant discriminant information due to the intermediate PCA step. The similar observation has also been found in [10], [16].

TABLE I
AVERAGE PERCENTAGE OF ERROR RATES OF DF-LDA OVER THAT OF OTHERS

Methods	Eigenfaces	Fisherfaces	D-LDA
$\mathcal{E}_{orl}$	74.18%	38.51%	80.03%
$\mathcal{E}_{umist}$	26.75%	47.68%	79.6%
$(\mathcal{E}_{orl} + \mathcal{E}_{umist})/2$	50.47%	43.1%	79.82%

The weighting function $w(d_{ij})$ influences the performance of the DF-LDA method. For different feature extraction tasks, appropriate values for the weighting exponent function should be determined through experimentation using the available training set. However, it appears that there is a set of values for which good results can be obtained for a wide range of applications. Following the recommendation in [7] we examine the performance of the DF-LDA method for $w(d_{ij}) \in \{d^{-4}, d^{-8}, d^{-12}, d^{-16}\}$. Results obtained through the utilization of these weighting functions are depicted in Fig. 5 where error rates are plotted against the feature vectors selected (output space dimensionality). The lowest error rate on the ORL database is approximately 4.0% and it is obtained using a weighting function of $w(d) = d^{-16}$ and a set of $M = 22$ feature basis vectors, a result comparable to the best results reported previously in the literatures [14], [15].

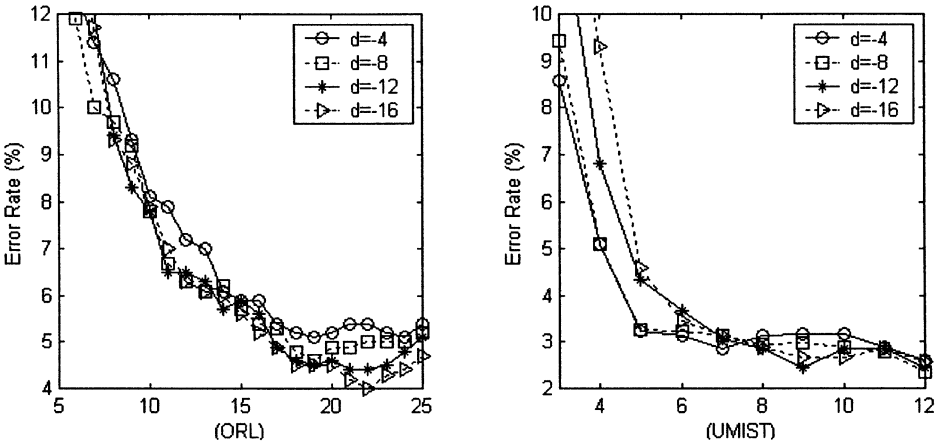


Fig. 5. Error rates of DF-LDA as functions of the number of feature vectors with $r = 20$ and different weighting functions.

IV. Conclusion

In this paper, a new feature extraction method for face recognition tasks has been proposed. The method introduced here utilizes the well-known framework of linear discriminant analysis and it can be considered as a generalization of a number of techniques which are commonly in use. The new method utilizes a new variant of D-LDA to safely remove the null space of the between-class scatter matrix and applies a fractional step LDA scheme to enhance the discriminatory power of the obtained D-LDA feature space. The effectiveness of the proposed method has been demonstrated through experimentation using two popular face databases.

The DF-LDA method presented here is a linear pattern recognition method. Compared with nonlinear models, a linear model is rather robust against noises and most likely will not overfit. Although it has been shown that distribution of face patterns is highly non convex and complex in most cases, linear methods are still able to provide cost effective solutions to the FR tasks through integration with other strategies, such as the principle of "divide and conquer," in which a large and nonlinear problem is divided into a few smaller and local linear subproblems. The development of mixtures of localized DF-LDA to be used in the problem of large size face recognition as well as the development of a nonlinear DF-LDA through the utilization of kernel machine techniques are research topics under current investigation.

Acknowledgment

The authors would like to thank Dr. D. Graham and Dr. N. Allinson for providing the UMIST face database, and thank AT&T Laboratories Cambridge for providing the ORL face database.

References

[1] R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey," *Proc. IEEE*, vol. 83, pp. 705–740, May 1995.
[2] M. Turk, "A random walk through eigenspace," *IEICE Trans. Inform. Syst.*, vol. E84-D, pp. 1586–1695, Dec. 2001.
[3] M. Turk and A. P. Pentland, "Eigenfaces for recognition," *J. Cognitive Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.
[4] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 711–720, May 1997.
[5] L.-F. Chen, H.-Y. Mark Liao, M-.T. Ko, J.-C. Lin, and G.-J. Yu, "A new LDA-based face recognition system which can solve the small sample size problem," *Pattern Recognition*, vol. 33, pp. 1713–1726, 2000.
[6] H. Yu and J. Yang, "A direct LDA algorithm for high-dimensional data with application to face recognition," *Pattern Recognition*, vol. 34, pp. 2067–2070, 2001.
[7] R. Lotlikar and R. Kothari, "Fractional-step dimensionality reduction," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 623–627, June 2000.
[8] ORL face database.. AT&T Laboratories, Cambridge, U.K.. [Online]. Available: http://www.cam-orl.co.uk/facedatabase.html
[9] D. L. Swets and J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, pp. 831–836, Aug. 1996.
[10] C. Liu and H. Wechsler, "Evolutionary pursuit and its application to face recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 570–582, June 2000.
[11] K. Liu, Y. Q. Cheng, J. Y. Yang, and X. Liu, "An efficient algorithm for Foley–Sammon optimal set of discriminant vectors by algebraic method," *Int. J. Pattern Recog. Artificial Intell.*, vol. 6, pp. 817–829, 1992.
[12] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1992.
[13] D. B. Graham and N. M. Allinson, "Characterizing virtual eigensignatures for general purpose face recognition," in *Face Recognition: From Theory to Applications*, H. Wechsler, P. J. Phillips, V. Bruce, F. Fogelman-Soulie, and T. S. Huang, Eds., 1998, vol. 163, NATO ASI Series F, Computer and Systems Sciences, pp. 446–456.
[14] S. Z. Li and J. Lu, "Face recognition using the nearest feature line method," *IEEE Trans. Neural Networks*, vol. 10, pp. 439–443, Mar. 1999.
[15] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural network approach," *IEEE Trans. Neural Networks*, vol. 8, pp. 98–113, Jan. 1997.
[16] A. M. Martnez and A. C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, pp. 228–233, Feb. 2001.