Kernel Fisher Discriminant Analysis in Full Eigenspace

Bappaditya Mandal Xudong Jiang Alex Kot

School of Electrical and Electronic Engineering,
Nanyang Technological University,
Block S1, 50 Nanyang Avenue, Singapore 639798.
Phone: +65 6790 5018; Fax: +65 6793 3318;
Email: {bapp0001, exdjiang, eackot}@ntu.edu.sg

Abstract*— This work proposes a method which enables us to perform kernel Fisher discriminant analysis in the whole eigenspace for face recognition. It employs the ratio of eigenvalues to decompose the entire kernel feature space into two subspaces: a reliable subspace spanned mainly by the facial variation and an unreliable subspace due to finite number of training samples. Eigenvectors are then scaled using a suitable weighting function. This weighting function circumvents undue scaling of projection vectors corresponding to the undependable small and zero eigenvalues. Eigenfeatures are only extracted after the discriminant evaluation in the whole kernel feature space. These efforts facilitate a discriminative and stable low-dimensional feature representation of the face image. Experimental results comparing other popular kernel subspace methods on FERET, ORL and GT databases show that our approach consistently outperforms others.

Keywords - Face recognition, kernel Fisher discriminant analysis, feature extraction, subspace methods**.**

I. INTRODUCTION

How the mammalian brain solves the problem of visual recognition has been a topic of study since the early days of brain science. Psychological experiments on human beings have shown that faces are recognized more holistically than other kinds of objects (e.g. houses, inverted faces, scrambled faces). Recently, there has been a growing interest in the holistic/appearance based approaches for face recognition. These appearance based approaches, in general, use statistical estimates for creating subspaces, which are utilized in subsequent face recognition. Although linear subspace methods have gained considerable attention, they cannot capture the nonlinearities and complex relationships among the input data that exist due to the large expression, illumination and pose variations. While nonlinear or kernel based subspace methods like kernel principal component analysis (KPCA) [1] and kernel Fisher discriminant analysis (KFDA) [2] have shown promising results. Good reviews on linear and nonlinear subspace based face recognition can be found in [3], [4], [5].

These kernel methods apply nonlinear mapping $\Phi : X \in \mathbb{R}^n \rightarrow \Phi(X) \in \mathbb{H}$ in the image space $\mathbb{R}^n$, followed by linear subspace methods like PCA and FDA in the mapped feature space $\mathbb{H}$. Examples include KPCA [1] and KFDA [2], [6]. Since the feature space $\mathbb{H}$ can be very high or possibly infinite dimensional and the orthogonality needs to be characterized in such a space, it is reasonable to view $\mathbb{H}$ as a Hilbert space. It is difficult to compute the dot products in the high dimensional feature space $\mathbb{H}$. Instead of mapping the data explicitly, the feature space can be computed by using the kernel trick, in which the inner products $\langle \Phi(X_i), \Phi(X_j) \rangle$ in $\mathbb{H}$ can be replaced with a kernel function $K(X_i, X_j)$, where $K(X_i, X_j) = \langle \Phi(X_i), \Phi(X_j) \rangle = \Phi(X_i)^T \cdot \Phi(X_j)$ and X_i, X_j are sample vectors in the image space $\mathbb{R}^n$. So, the nonlinear mapping Φ can be performed implicitly in image space $\mathbb{R}^n$ [7], [8]. Numerous studies [4], [9], [10] demonstrate that these kernel based approaches are very effective in many real-world applications. However, the basic subspace analysis has still outstanding challenging problems when applied to the face recognition due to the high dimensionality of the face image and the finite number of training samples in practice. Although, in this work, we take advantages from the nonlinear mapping, it does not explore the optimization of kernel mapping functions or any of its parameters. Nevertheless, throughout this paper, we assume a popular nonlinear mapping function with its parameters fixed for all the experiments and perform our proposed algorithm in this nonlinearly mapped feature space.

Over the last decade KFDA and its numerous variations have been applied in face recognition to solve the expression, pose and illumination problems [2], [6], [10]. Liu *et al.*[11] performed a good experimental analysis on KFDA and showed that KFDA gives better performance than that of KPCA. KFDA applies PCA first for dimensionality reduction so as to make the within-class scatter matrix nonsingular before the application of LDA. However, applying PCA for dimensionality reduction may lose important discriminative information [12], [13], [14], [15]. In fact, most of the nonlinear subspace based face recognition methods perform dimensionality reduction or discard a subspace before the discriminant evaluation. The null space approach, NKDA [16] eliminates the principal subspace and extracts eigenfeatures only from the eigenvectors corresponding to the zero eigenvalues. Therefore, NKDA assumes that the null space contains the most discriminative

information which is contradictory to KFDA.

To solve the small sample size problem Lu *et al.* [17] proposed kernel Direct-LDA (KDDA) method, which first removes the null space of the between-class scatter matrix and then extracts the eigenvectors corresponding to the smallest eigenvalues of the within-class scatter matrix. However, as pointed out in [18], the removal of the null components of between-class scatter matrix influence the projection of within-class matrix and hence they should not be discarded. Moreover, it is an open question of how to scale the extracted features as the smallest eigenvalues are very sensitive to noise. A common problem of KFDA, NKDA and KDDA approaches is that they all lose some discriminative information, either in the principal or in the null space because they perform the discriminant evaluation in a subspace.

In fact, the discriminative information resides in both subspaces. Recently, Yang *et al.* [19] proposed a complete kernel Fisher discriminant framework (CKFD), where features extracted from the two complementary subspaces are combined by a summed distance measures in the recognition phase [19]. Open questions of this approach are how to divide the space into the principal and the complementary subspaces and how to apportion a given number of features to the two subspaces. Furthermore, as the discriminative information resides in the both subspaces, it is inefficient or only suboptimal to extract features separately from the two subspaces.

This paper proposes a method which performs kernel Fisher discriminant analysis in full eigenspace (KFDAFE). Eigenratios (shown in Fig. 1) are used to decompose the nonlinear within-class eigenspace into two subspaces: a reliable subspace spanned mainly by the facial variation and an unreliable subspace due to limited number of training samples. To alleviate the problems of scaling eigenfeatures caused by unreliable small and zero eigenvalues we propose a weighting function (shown in Fig. 2). This weighting function circumvents undue scaling of projection vectors corresponding to the small and zero eigenvalues. Eigenfeatures are then extracted after the discriminant evaluation in the whole nonlinear eigenspace. In section 2, we first study the behavior of the unreliable small eigenvalues and eigenratios of within-class variation matrix, then propose a methodology to decompose the eigenspace into principal and unreliable subspaces. Eigenfeature scaling and extraction are presented in Section 3. Experimental results and discussions are presented in section 4. Conclusions are drawn in section 5.

II. Feature Scaling and Subspace Decomposition

A. Overview of Kernel Fisher Discriminant Analysis

For a nonlinear mapping Φ, the image data space $\mathbb{R}^n$ can be mapped into the feature space $\mathbb{H}$:

$$\Phi : X \in \mathbb{R}^n \rightarrow \Phi(X) \in \mathbb{H}. \tag{1}$$

Consequently, a pattern in the original image space $\mathbb{R}^n$ is mapped into a potentially much higher dimensional feature vector in the feature space $\mathbb{H}$. Given a set of properly normalized h-by-w face images, we can form a training set of column vectors $\{X_{ij}\}$, where $X_{ij} \in \mathbb{R}^{n=hw}$ is called image vector, by lexicographic ordering the pixel elements of image j of person i. Let the training set contain p persons and q_i sample images for person i. The total number of training samples is $l = \sum_{i=1}^{p} q_i$. The within-class scatter matrix is defined by

$$\mathbf{S}^w = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{q_i}\sum_{j=1}^{q_i}(\Phi(X_{ij}) - \overline{\Phi(X_i)})(\Phi(X_{ij}) - \overline{\Phi(X_i)})^T, \tag{2}$$

where $\overline{\Phi(X_i)} = \frac{1}{q_i}\sum_{j=1}^{q_i}\Phi(X_{ij})$. The between-class scatter matrix $\mathbf{S}^b$ is defined by

$$\mathbf{S}^b = \frac{1}{p}\sum_{i=1}^{p}(\overline{\Phi(X_i)} - \overline{\Phi(X)})(\overline{\Phi(X_i)} - \overline{\Phi(X)})^T, \tag{3}$$

where $\overline{\Phi(X)} = \frac{1}{p}\sum_{i=1}^{p}\overline{\Phi(X_i)}$, assuming all classes have equal prior probability.

The well known Fisher objective function [20], [21], [22] can be written in the mapped space $\mathbb{H}$ as

$$J(\mathbf{\Omega}) = \arg\max_{\Omega}\frac{\mid \mathbf{\Omega}^T\mathbf{S}^b\mathbf{\Omega} \mid}{\mid \mathbf{\Omega}^T\mathbf{S}^w\mathbf{\Omega} \mid}. \tag{4}$$

Because any solution $\mathbf{\Omega} \in \mathbb{H}$ must lie in the span of all the samples in $\mathbb{H}$, there exist coefficients ψ_i for $i = 1, 2, \ldots, l$, such that

$$\mathbf{\Omega} = \sum_{i=1}^{l}\psi_i\Phi(X_i). \tag{5}$$

As shown in [21], combining (4) and (5), we can write

$$\mathbf{\Omega}^T\mathbf{S}^w\mathbf{\Omega} = \mathbf{\Psi}^T\mathbf{S}_\Phi^w\mathbf{\Psi}, \tag{6}$$

$$\mathbf{\Omega}^T\mathbf{S}^b\mathbf{\Omega} = \mathbf{\Psi}^T\mathbf{S}_\Phi^b\mathbf{\Psi}, \tag{7}$$

where $\mathbf{\Psi} = \{\psi_i\}_{i=1}^{l}$ and

$$\begin{cases} \mathbf{S}_\Phi^w = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{q_i}\sum_{j=1}^{q_i}(\zeta_j - \mu_i)(\zeta_j - \mu_i)^T, \\ \zeta_j = (K(X_1, X_j), K(X_2, X_j), \ldots, K(X_l, X_j))^T \end{cases}, \tag{8}$$

$$\begin{cases} \mu_i = (\frac{1}{q_i}\sum_{j=1}^{q_i}K(X_1, X_j), \frac{1}{q_i}\sum_{j=1}^{q_i}K(X_2, X_j), \ldots, \\ \frac{1}{q_i}\sum_{j=1}^{q_i}K(X_l, X_j))^T, \\ \mathbf{S}_\Phi^b = \frac{1}{p(p-1)}\sum_{i=1}^{p}\sum_{j=1}^{p}(\mu_i - \mu_j)(\mu_i - \mu_j)^T \end{cases}. \tag{9}$$

So the solution of function (4) can be obtained by maximizing

$$J(\mathbf{\Psi}) = \arg\max_{\Psi}\frac{\mid \mathbf{\Psi}^T\mathbf{S}_\Phi^b\mathbf{\Psi} \mid}{\mid \mathbf{\Psi}^T\mathbf{S}_\Phi^w\mathbf{\Psi} \mid} \tag{10}$$

and the problem of kernel discriminant analysis is converted into finding the leading eigenvectors of $\mathbf{S}_\Phi^{w\,-1}\mathbf{S}_\Phi^b$. However, in practice, the inversion of $\mathbf{S}_\Phi^w$ is impossible as it is often singular due to the limited number of training samples. For a new image X, its projection onto $\mathbf{\Omega}$ in $\mathbb{H}$ can be calculated by

$$(\mathbf{\Omega} \cdot \Phi(X)) = \sum_{i=1}^{l}\psi_i K(X_i, X). \tag{11}$$

Let $\mathbf{S}_\Phi^g, g \in \{w, b\}$ represent one of the above scatter matrices. If we regard the elements of the image vector or

the class mean vector as features, these preliminary features will be de-correlated by solving the eigenvalue problem

$$\mathbf{\Lambda}^g = \mathbf{\Psi}^{gT}\mathbf{S}_\Phi^g\mathbf{\Psi}^g, \tag{12}$$

where $\mathbf{\Psi}^g = [\psi_1^g, ..., \psi_l^g]$ is the eigenvector matrix of $\mathbf{S}_\Phi^g$, and $\mathbf{\Lambda}^g$ is the diagonal matrix of eigenvalues $\lambda_1^g, ..., \lambda_l^g$ corresponding to the eigenvectors. We assume that the eigenvalues are sorted in descending order $\lambda_1^g \geq, ..., \geq \lambda_l^g$. The plot of eigenvalues λ_k^g against the index k is called eigenspectrum of the training data in the nonlinear plane. It plays a critical role in the subspace methods as the eigenvalues are used to scale and extract features.

B. Problems in Feature Scaling and Extraction of KFDA

Fisher's discriminant criteria (10) is known to be the Bayes optimal classifier for normal distributions with equal covariance. However, in all kernel subspace applications any of the scatter matrices (8) and (9) can be singular [20], [17], [23]. If we compute all the eigenvalues $diag(\mathbf{\Lambda}^w) = [\lambda_1^w, ..., \lambda_l^w]$ and eigenvectors $\mathbf{\Psi}^w = [\psi_1^w, ..., \psi_l^w]$ of the l-by-l dimensional matrix $\mathbf{S}_\Phi^w$ using (12), the projection matrix $\bar{\mathbf{\Psi}}^w = [\psi_1^w/\sigma_1^w, ..., \psi_l^w/\sigma_l^w]$ is so called whitened eigenvector matrix of $\mathbf{S}_\Phi^w$ with $||\psi_k^w|| = 1$ and $\sigma_k^w = \sqrt{\lambda_k^w}$. This implies that if any one of the eigenvalues in (12) of these matrices is zero then the corresponding eigenvector (10) gets an infinite weighting factor. In practice, most of the subspace based algorithms circumvent this problem by ignoring the eigenvectors corresponding to zero eigenvalues. However, as pointed out earlier that the null space of $\mathbf{S}_\Phi^w$ contain indispensable discriminative information essential for improving recognition accuracy.

The above argument can be viewed as an l-dimensional pattern vector $\Delta_{ij} = K(X_i, X_j)$ is first represented by an l-dimensional eigenfeature vector $Y_{ij} = \mathbf{\Psi}^{wT}\Delta_{ij}$, and then multiplied by a weighting function

$$w_k^w = \begin{cases} 1/\sqrt{\lambda_k^w}, & k \leq r_w \\ 0, & r_w < k \leq l \end{cases}, \tag{13}$$

as shown in Fig. 2, where r_w is the rank of $\mathbf{S}_\Phi^w$. It is apparent from (13) that the eigenvectors $\{\psi_k^w\}_{k=r_w+1}^l$ or the null space of $\mathbf{S}_\Phi^w$ are weighted by zero and thus the corresponding eigenvectors fail to contribute to the whole space discriminant evaluation, which is done in the later portion of the algorithm. This is unreasonable because features in the null space have zero within-class variances based on the training data and hence should be more heavily weighted. It seems anomalous that the weighting function increases with the decrease of the eigenvalues and then suddenly has a big drop from the maximum value to zero as shown in Fig. 2. Furthermore, weights determined by the inverse of σ_k^w is, though optimal in terms of the ML estimation, dangerous when σ_k^w is small $(m < k \leq r_w)$. The small and zero eigenvalues are training-set-specific and very sensitive to different training sets [24]. Adding new samples to the training set or using different training set may easily change some zero eigenvalues to nonzero and make some very small eigenvalues several times larger. Therefore, these eigenvalues of the within-class scatter matrix are unreliable.

C. Eigenratiospectrum and Subspace Decomposition

In order to alleviate the above problem we first work on the eigenspectrum of the within-class variation matrix. It is not difficult to estimate the rank of $\mathbf{S}_\Phi^w$, which is $r_w \leq (l-p)$. The eigenvalues whose indices are close to r_w $(m < k \leq r_w)$ are very small or close to zero, their inverses give undue overemphasis to the eigenvectors corresponding to this region (or indices) as shown in Fig. 2. To successfully differentiate the unreliable eigenvalues from the larger ones we propose to use the eigenratios of the eigenspectrum to decompose the whole eigenspace into two subspaces: a principal or reliable subspace spanned mainly by the facial variation, $\mathbf{P} = \{\psi_k^w\}_{k=1}^m$ and an unreliable or noise dominating subspace due to limited number of training samples, $\bar{\mathbf{P}} = \{\psi_k^w\}_{k=m+1}^l$. For a clearer illustration, we first define the eigenratios as $\mathbf{\Gamma}_\Phi^w = \{\gamma_1^w, ..., \gamma_{r_w-1}^w\}$, such that

$$\gamma_k^w = \frac{\lambda_k^w}{\lambda_{k+1}^w}, \quad 1 \leq k < r_w, \tag{14}$$

the plot of eigenratios γ_k^w of a typical real eigenspectrum against the index k is called kernel eigenratiospectrum of the training data as shown in Fig. 1.

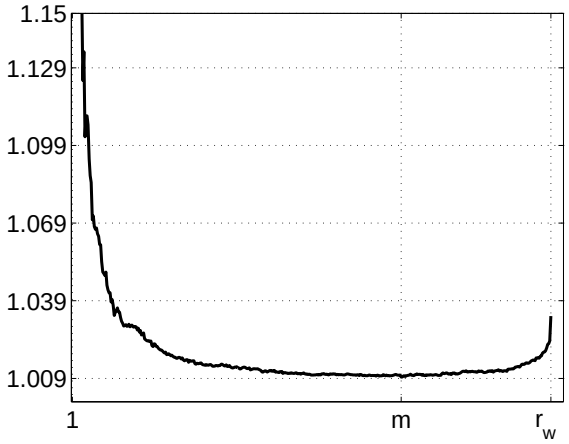

Fig. 1. Plot of eigenratios based on (14) and finding index m using (15) on a typical real kernel eigenratiospectrum of $\mathbf{S}_\Phi^w$ matrix.

For a robust training, the database size should be significantly larger than the (face or reliable) dimensionality m, although it could be and in practice is, much smaller than the number of training samples l. Also, in practical applications, there always exist large variations in face images and their dimensionality in various databases. Thus, one robust way of finding such a point would be finding the minimum of the eigenratios. The start point of the unreliable region $m+1$ is estimated by

$$\gamma_{m+1}^w = \min\{\forall\gamma_k^w, \quad 1 \leq k < r_w\}. \tag{15}$$

A typical such m value of a real kernel eigenspectrum is shown in Fig. 1.

The main purpose of finding the value of m using the eigenratios is to distinguish the reliable eigenvalues from the unreliable ones, which facilitates the decomposition of the entire eigenspace into reliable $\mathbf{P}$ and unreliable $\bar{\mathbf{P}}$ subspaces. Eigenvalues in the unreliable subspace $\bar{\mathbf{P}}$, spanned by $\{\psi_k^w\}_{k=m+1}^{l}$, will be replaced by a constant. This enables us to perform discriminant evaluation and feature extraction from the whole eigenspace of $\mathbf{S}_\Phi^w$ matrix (as described in section 3).

We conducted several experiments on different face image databases, the eigenratio plots shown in Fig. 1 is a general behavioral pattern that all the eigenratios of different databases portray. Thus, this depicts the general behavioral characteristics of eigenvalues. It is apparent from the graph that the eigenratios first decreases very rapidly, then stabilizes and finally increases. The increase of the eigenratios should not be the behavior of the true variances but occurs due to the limited number of training samples. The corresponding eigenvalues are therefore unreliable. Similarly, the zero eigenvalues are caused by the limited number of training samples and hence are unreliable. These unreliable eigenvalues may result in serious problems if their inverses are used to weight the eignefeatures as shown in Fig. 2. Therefore, we should not trust the eigenvalues, $\lambda_k, k > m$, where m is determined by the minimal eigenratios as given by (15).

III. Scaling of Kernel Eigenspectrum and Feature Extraction

A. Scaling of Kernel Eigenfeatures

As pointed out in [25], the largest sample-based eigenvalues are biased high and the smallest ones are biased low due to the finite number of training samples. The eigenspectrum in the principal/reliable subspace is dominated by the face structural component, hence, we keep the eigenvalues in the principal subspace unchanged. In the unreliable subspace $\bar{\mathbf{P}}$, however, the limited number of training samples results in faster decay of the eigenvalues than the true variances. Therefore, the decay of the eigenvalues should be slowed down to compensate the effect of the finite number of training samples.

From Fig. 2 it is evident that when the inverses of the eigenvalues $\{\lambda_k^w\}_{k=m+1}^{l}$ are used for feature weighting (13), the corresponding eigenvectors get undue over-scaling in this range. Therefore, we propose to replace the unreliable or noise dominating region eigenvalues $\{\lambda_k^w\}_{k=m+1}^{l}$ by choosing the eigenvalue corresponding to the value m obtained from (15), which is given by

$$\lambda_{const}^w = \max\{\forall \lambda_k^w, \quad m \leq k \leq r_w\}. \tag{16}$$

Thus, the final weighting function can be written as

$$\tilde{w}_k^w = \begin{cases} 1/\sqrt{\lambda_k^w}, & k \leq m \\ 1/\sqrt{\lambda_{const}^w}, & m < k \leq l \end{cases}. \tag{17}$$

Fig. 2 shows the proposed feature weighting function $\tilde{w}_k^w$ calculated by (14), (15), (16) and (17) comparing with that w_k^w of (13). Obviously, the new weighting function $\tilde{w}_k^w$ is identical to w_k^w in the principal space, remains constant along with k at a much lower value than w_k^w in the unreliable space and has maximal weights instead of zero of w_k^w in the null space.

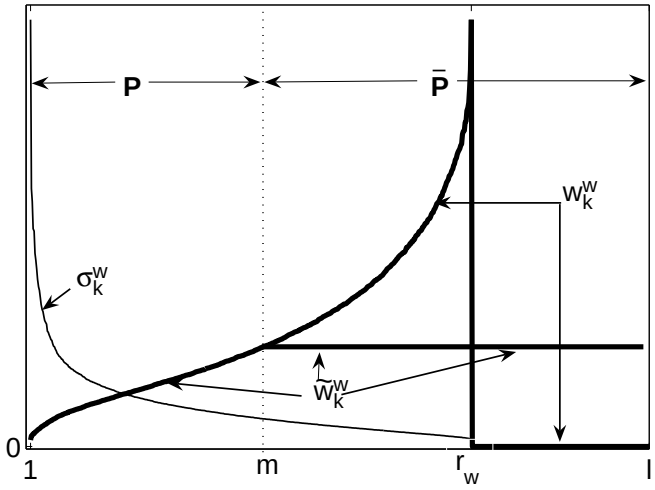

Fig. 2. Weighting functions of (13) and (17) in the principal- and unreliable-subspaces based on a typical real kernel eigenspectrum.

Using this weighting function and the eigenvectors ψ_k^w, training pattern data are transformed to

$$\tilde{Y}_{ij} = \tilde{\mathbf{\Psi}}_l^{w^T} \Delta_{ij}, \tag{18}$$

where

$$\tilde{\mathbf{\Psi}}_l^w = [\tilde{w}_k^w \psi_k^w]_{k=1}^{l} = [\tilde{w}_1^w \psi_1^w, ..., \tilde{w}_l^w \psi_l^w] \tag{19}$$

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

The problems of dimensionality reduction and discriminant evaluation in KFDA are also discussed in [22], where Chen *et al.* proposed kernel machine-based regularized Fisher discriminant (K1PRFD) algorithm. Although this method uses the full eigenspace, their approach regularize the pooled within-class scatter matrix equivalently by adding a constant to all eigenvalues. Although the largest sample-based eigenvalues are biased high and the smallest ones are biased low, as pointed out in [25], the bias is most pronounced when the population eigenvalues tend toward equality, and it is correspondingly less severe when their values are highly disparate. For the application of face recognition, it is well-known that the eigenspectrum first decays very rapidly and then stabilizes. Hence, adding a constant to the eigenspectrum may bias back the rapidly changing eigenvalues in principal space too much that introduces additional error source, and bias back the flat eigenvalues in null space too little at the same time [25].

B. Kernel Eigenfeature Extraction

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

$$\tilde{\mathbf{S}}_{\Phi}^{b} = \frac{1}{p}\sum_{i=1}^{p}(\overline{\tilde{Y}}_i - \overline{Y})(\overline{\tilde{Y}}_i - \overline{Y})^T, \tag{20}$$

where $\overline{\tilde{Y}}_i = \frac{1}{q_i}\sum_{j=1}^{q_i}\tilde{Y}_{ij}$ and $\overline{Y} = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{q_i}\sum_{j=1}^{q_i}\tilde{Y}_{ij}$. The weighted features $\tilde{Y}_{ij}$ will be de-correlated for $\tilde{\mathbf{S}}_{\Phi}^{b}$ by solving the eigenvalue problem as (12). Suppose that the eigenvectors in the eigenvector matrix $\tilde{\mathbf{\Psi}}_l^b = [\tilde{\psi_1^b}, ..., \tilde{\psi_l^b}]$ are sorted in descending order of the corresponding eigenvalues. The dimensionality reduction is performed here by keeping the eigenvectors with the d largest eigenvalues

$$\tilde{\mathbf{\Psi}}_d^b = [\tilde{\psi_k^b}]_{k=1}^d = [\tilde{\psi_1^b}, ..., \tilde{\psi_d^b}], \tag{21}$$

where d is the number of features usually selected by a specific application. Thus, the proposed feature scaling and extraction matrix $\mathbf{U}_{\Phi}$ is given by

$$\mathbf{U}_{\Phi} = \tilde{\mathbf{\Psi}}_l^w \tilde{\mathbf{\Psi}}_d^b. \tag{22}$$

This transforms a pattern image vector $\Delta_{ij} = K(X_i, X_j) \in \mathbb{H}$ of dimensionality l, into a feature vector F, $F \in \mathbb{H}$ of dimensionality d, by

$$F = \mathbf{U}_{\Phi}^T \Delta_{ij}. \tag{23}$$

It is apparent from the above equations that we perform the subspace decomposition to apply our weighting scheme for feature scaling. The discriminant evaluation (here the evaluation of the eigenvalues of $\tilde{\mathbf{S}}_{\Phi}^{b}$) is performed in the full kernel space $\mathbb{H}$. As a result of this, the feature extraction is not restricted to project a kernel vector into one of these subspaces. More specifically, any single feature in F is extracted from the whole kernel space $\mathbb{H}$ since any final projection vector $\mathbf{U}_{\Phi}$ may have nonzero components in all the other subspaces.

C. The Proposed Algorithm

The proposed kernel Fisher discriminant analysis in full eigenspace (KFDAFE) approach is summarized below:

At the training stage:

1) Given a training set of face image vectors $\{X_{ij}\}$ and a kernel function $K(X_i, X_j)$, compute $\Delta_{ij} = K(X_i, X_j)$.
2) Compute $\mathbf{S}_{\Phi}^{w}$ by (8) and solve the eigenvalue problem as (12).
3) Decompose the kernel eigenspace into principal- and unreliable-spaces by determining the m value using (14) and (15).
4) Transform the training pattern samples represented by Δ_{ij} into $\tilde{Y}_{ij}$ by (18) with the weighting function (17) determined by (14), (15) and (16).
5) Compute $\tilde{\mathbf{S}}_{\Phi}^{b}$ by (20) with $\tilde{Y}_{ij}$ and solve the eigenvalue problem as (12).
6) Obtain the final feature scaling and extraction matrix by (19), (21) and (22) with a predefined number of features d. □

At the recognition stage:

1) Transform each n-D face image vector X into l-D feature pattern vector Δ_{ij} using the kernel function K, such that $\Delta_{ij} = K(X_i, X_j)$.
2) Transform each l-D feature pattern vector Δ_{ij} into d-D feature vector F by (23) using the feature regularization and extraction matrix $\mathbf{U}_{\Phi}$ obtained in the training stage.
3) Apply a classifier trained on the gallery set to recognize the probe feature vectors. □

In the experiments of this work, a simple first nearest neighborhood classifier (1-NNK) is applied to test the proposed kernel Fisher discriminant analysis in full eigenspace (KFDAFE) approach for face recognition. Euclidean distance measure between a probe feature vector $\mathbf{F}_P$ and a gallery feature vector $\mathbf{F}_G$

$$dst(\mathbf{F}_P, \mathbf{F}_G) = \sqrt{(\mathbf{F}_P - \mathbf{F}_G)^T(\mathbf{F}_P - \mathbf{F}_G)} \tag{24}$$

is applied to the proposed approach.

IV. Experiments and Discussions

In all experiments reported in this work, images are preprocessed following the CSU Face Identification Evaluation System [26]. Three databases: ORL, GT and FERET are used for testing. Each database is partitioned into training and testing sets. For FERET databases, there is no overlap in subject between the training and testing sets. As ORL and GT databases have only a small number of subjects, both training and testing sets contain all subjects. However, there is no overlap in the sample image between the training and testing sets. In our experiments, the polynomial kernel function is chosen, $K(X_i, X_j) = \langle \Phi(X_i), \Phi(X_j) \rangle = (a\langle X_i \cdot X_j \rangle + b)^c$, since it gave good performances in the experiments of [11], [27], [10], [16], [28]. The kernel parameters are set same as that of in [11], [27], [28]. The recognition error rate given in this work is the percentage of the incorrect top 1 match on the testing set. The proposed KFDAFE method is tested and compared with KFDA [2], KDDA [17], NKDA [16], CKFD [19] and K1PRFD [22] approaches. The parameters of CKFD are applied that are mentioned in the experiments of [19].

A. Results on FERET Database

In FERET database, the face image variations include facial expression and other details (like glasses or no glasses), illumination, pose, and aging [29]. We select 2388 images comprising of 1194 subjects (two images per subject) from this database. Images are cropped into the size of 38×33. Images of 250 subjects are randomly selected for training and the remaining images of 944 subjects are used for testing. Hence, there is no overlap in subject between the training and testing sets. The recognition error rate given in this work is the percentage of the incorrect top 1 match on the testing set. The recognition error rates are shown in Fig. 3. Both KFDA and KDDA perform badly because the smaller number of training images does not well represent the variations of testing images. K1PRFD achieves higher accuracy gain than CKFD. This

shows that the summed distance used in CKFD does not well-handle the database with smaller number of training samples. The KFDAFE approach consistently outperforms all other approaches for all number of features. The gain is significant for smaller number of features.

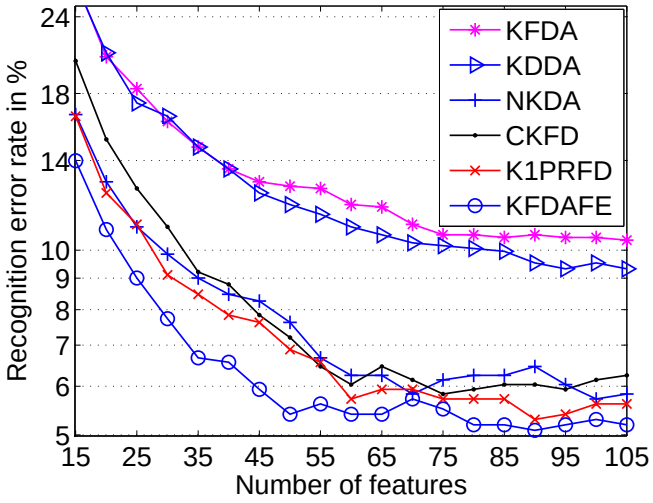

Fig. 3. Recognition error rate against the number of features used in the matching on the FERET database of 500 training images (250 subjects) and 1888 testing images (944 subjects).

B. Results on ORL Database

In the experiments on ORL database [30], images are cropped into the size of 57×50. The ORL database contains 400 images of 40 subjects (10 images per subject). Some images were captured at different times and have different variations including expression (open or closed eyes, smiling or nonsmiling) and facial details (glasses or no glasses). The images were taken with a tolerance for some tilting and rotation of the faces up to 20 degrees. In this experiment, we test various approaches using the first 5 samples per subject (200 images) for training and the remaining 5 samples per subject (200 images) for testing. Fig. 4 shows the recognition error rate on the testing set against the number of features. As the training set has only 200 images, it does not well represent the variations of testing images. Therefore, the small principal space does not capture the discriminative information well. This results in poor performance of KFDA and KDDA approaches. Similar to the previous experiment, CKFD and K1PRFD outperform KFDA, KDDA and NKDA approaches. CKFD, which extracts the features separately from the two subspaces outperforms K1PRFD significantly for larger number of features. Again, the proposed KFDAFE approach consistently outperforms all other approaches for all number of features.

C. Results on Georgia Tech Database

The Georgia Tech (GT) Face Database [31] consists 750 color images of 50 subjects (15 images per subject). For most

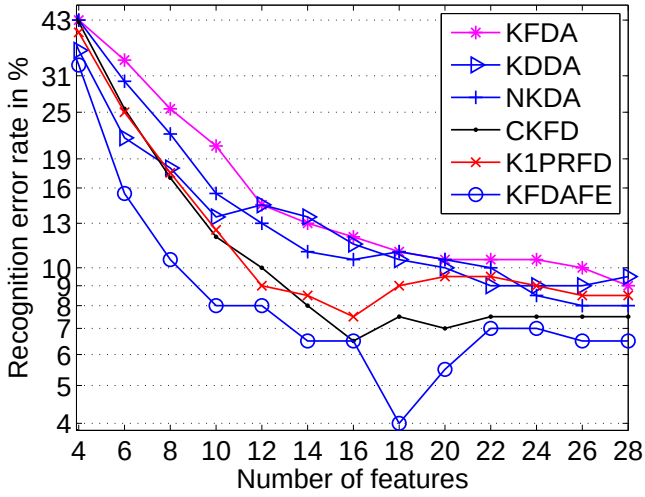

Fig. 4. Recognition error rate against the number of features used in the matching on the ORL database of 200 training images (40 subjects) and 200 testing images (40 subjects).

of the subjects the face images were taken in two or three sessions over a period of three months, allowing for strong variation in size, facial expression, illumination, and rotation in both the image plane and perpendicular to the image plane. These images are converted to gray-scale and cropped into the size of 112×92. The first 8 images of all the subjects are used in the training and the remaining 7 images serve as testing images. The testing results are numerically recorded in Table I. All the approaches perform relatively similar to the previous

TABLE I
RECOGNITION ERROR RATE OF DIFFERENT APPROACHES FOR DIFFERENT NUMBER OF FEATURES ON GT DATABASE.

Database	GT (400 / 350 training / testing images)						
# Feature	6	10	14	20	28	38	46
KFDA	38.57	26.57	22.00	18.00	14.86	12.86	12.57
KDDA	33.14	22.00	15.43	12.00	9.14	9.14	9.43
NKDA	35.43	20.00	17.14	13.43	13.43	12.29	12.29
CKFD	38.00	23.43	15.71	12.86	10.86	10.29	8.86
K1PRFD	31.71	19.43	17.14	13.71	13.14	11.71	12.57
KFDAFE	26.29	17.71	12.86	10.00	8.29	8.29	8.29

experiments. KDDA outperforms CKFD but not consistently. Both KDDA and CKFD outperform KFDA, NKDA and K1PRFD approaches but not consistently, this probably shows that KDDA and CKFD perform better when more number of samples per subject are present in the training database. The proposed KFDAFE approach consistently outperform all other approaches for all number of features. This demonstrates the effectiveness of the proposed kernel Fisher discriminant analysis and feature extraction in the full eigenspace of the within-class variation in alleviating the over-fitting problem or better discrimination ability.

V. Conclusions

In this paper, we have addressed the problems of eigenfeature scaling and its extraction from the principal and null subspaces. Information residing in the eigenratios of the within-class scatter matrix in the nonlinear space is used to decompose the eigenspace into two subspaces. Eigenfeatures are then scaled differently using a suitable weighting function. This weighting function circumvents undue scaling of projection vectors corresponding to the unreliable small and zero eigenvalues. Our proposed scaling and feature extraction method performs discriminant evaluation in the whole space and the feature extraction or dimensionality reduction occurs only at the final stage after the discriminant assessment. This facilitates a discriminative and stable low-dimensional feature representation of the image vectors. Experiments on the FERET, ORL and GT databases demonstrate that the proposed approach consistently outperforms other popular methods.

References

[1] B. Scholkopf, A. Smola, and K. R. Muller, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[2] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. R. Muller, "Fisher Discriminant Analysis with Kernels," *IEEE Workshop on Neural Networks for Signal Processing IX*, pp. 41–48, 1999.

[3] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face Recognition: A Literature Survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, December 2003.

[4] G. Shakhnarovich and B. Moghaddam, *Face Recognition in Subspaces*. 201 Broadway, Cambridge, Massachusetts 02139: Springer, 2004.

[5] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the Face Recognition Grand Challenge," in *IEEE Conf. Comp. Vis. & Patt. Recog.*, San Diego, June 2005, pp. 947–954.

[6] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, A. J. Smola, K. R. Muller, S. A. Solla, and T. K. Leen, "Invariant Feature Extraction and Classification in Kernel Spaces," *Advances in Neural Information Processing Systems, Cambridge, MA: MIT Press*, vol. 12, pp. 526–532, 2000.

[7] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[8] B. Scholkopf, S. Mika, C. J. Burges, P. Knirsch, K. R. Muller, G. Ratsch, and a. Smola, "Input Space Versus Feature Space in Kernel-based Methods," *IEEE Transc. Neural Network*, vol. 10, pp. 1000–1017, September 1999.

[9] H. Gupta, A. K. agarwal, T. Pruti, C. Shekhar, and R. Chellappa, "An Experiment evaluation of linear and kernel-based methods for Face Recognition," *IEEE Workshop on Applications on Computer Vision*, December 2002.

[10] M. H. Yang, "Kernel Eigenfaces vs. Kernel Fisherfaces: Face Recognition using Kernel Methods," *Proc. 5th IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 215–220, May 2002.

[11] Q. Liu, R. Huang, H. Lu, and S. Ma, "Face Recognition Using Kernel based Fisher Discriminant Analysis," *Proc. 5th IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 197–201, May 2002.

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

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

[14] X. Wang and X. Tang, "Dual-Space Linear Discriminant Analysis for Face Recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, Washington DC, USA, June 2004, pp. 564–569.

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

[16] W. Liu, Y. W. Wang, S. Z. Li, and T. N. Tan, "Null Space-based kernel Fisher Discriminant Analysis for Face Recognition," *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 369–374, 2004.

[17] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Face Recognition Using Kernel Direct Discriminant Analysis Algorithms," *IEEE Trans. Neural Networks*, vol. 14, no. 1, pp. 117–126, January 2003.

[18] H. Gao and J. W. Davis, "Why Direct LDA Is Not Equivalent to LDA," *Pattern Recognition*, vol. 39, pp. 1002–1006, 2006.

[19] J. Yang, A. F. Frangi, J. Y. Yang, D. Zhang, and Z. Jin, "KPCA Plus LDA: A Complete Kernel Fisher Discriminant Framework for Feature Extraction and Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, pp. 230–244, February 2005.

[20] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. San Diego, CA, USA: Academic Press, INC, 1990.

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

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

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

[24] X. D. Jiang, B. Mandal, and A. Kot, "Enhanced maximum likelihood face recognition," *IEE Electronics Letters*, vol. 42, no. 19, pp. 1089–1090, September 2006.

[25] J. H. Friedman, "Regularized Discriminant Analysis," *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 165–175, March 1989.

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

[27] Q. S. Liu, H. Q. Lu, and S. D. Ma, "Improving Kernel Fisher Discriminant Analysis for Face Recognition," *IEEE Trans. Circuits Sys. Video Techno.*, vol. 14, no. 1, pp. 42–49, January 2004.

[28] Q. S. Liu, . Tang, H. Lu, and S. D. Ma, "Face Recognition Using Kernel Scatter-Difference-Based Discriminant Analysis," *IEEE Trans. Neural Network*, vol. 17, no. 4, pp. 1081–1085, July 2006.

[29] P. J. Phillips, H. Moon, S. Rizvi, and P. Rauss, "The FERET evaluation methodology for face recognition algorithms," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, October 2000.

[30] F. Samaria and A. Harter, "Parameterization of a Stochastic Model for Human Face Identification," in *Proceedings of 2nd IEEE Workshop on Applications of Computer Vision*, Sarasota, FL, USA, December 1994, pp. 138–142.

[31] Georgia tech face database. [Online]. Available: http://www.anefian.com/face_reco.htm