Brief Papers

Regularized Kernel Discriminant Analysis With a Robust Kernel for Face Recognition and Verification

Stefanos Zafeiriou, Georgios Tzimiropoulos, Maria Petrou,
and Tania Stathaki

Abstract—We propose a robust approach to discriminant kernel-based feature extraction for face recognition and verification. We show, for the first time, how to perform the eigen analysis of the within-class scatter matrix directly in the feature space. This eigen analysis provides the eigenspectrum of its range space and the corresponding eigenvectors as well as the eigenvectors spanning its null space. Based on our analysis, we propose a kernel discriminant analysis (KDA) which combines eigenspectrum regularization with a feature-level scheme (ER-KDA). Finally, we combine the proposed ER-KDA with a nonlinear robust kernel particularly suitable for face recognition/verification applications which require robustness against outliers caused by occlusions and illumination changes. We applied the proposed framework to several popular databases (Yale, AR, XM2VTS) and achieved state-of-the-art performance for most of our experiments.

Index Terms—Eigenspectrum regularization, face recognition, kernel discriminant analysis, robust kernel.

I. INTRODUCTION

Learning algorithms for face recognition/verification have generated a wealth of scientific research within the computer vision community for more than two decades [1]. This research has primarily revolved around providing efficient solutions to the following problem: given samples of a high-dimensional space (facial images of our training set), estimate a low-dimensional face space that preserves the intrinsic structure of the input data [2]–[15]. Classification is then typically performed by projecting the probe face onto this low-dimensional space and applying off-the-shelf classifiers. Efficient estimation of the face space is usually hindered by two factors. First, in a practical setting, very few samples (usually 1–5 for each class) are available for training.

Second, the estimated subspace may inefficiently represent the probe face because of possible facial variations or different capturing conditions. Typical examples include various facial expressions, nonuniform illumination changes, occlusion, and misalignment errors.

To cope with such phenomena, recent research on feature extraction for face recognition and verification has focused on nonlinear kernel-based extensions to linear discriminant analysis (LDA) [4], [8], [10], [11]. Kernel discriminant analysis (KDA) optimization problems find a set of projection bases that maximize the between-class distances while minimizing the within-class distances in the feature space implied by the kernel. Similar to LDA, the main problem is how to exploit the discrimination capabilities of both the range and null space of the within-class scatter matrix. Popular methods discard discriminative information, either in one space or the other. For example, Kernel Fisherface [5] applies dimensionality reduction so as to make the within-class scatter matrix invertible before the application of LDA and therefore discards the discriminative information of the null space [8]. The authors in [10] and [11] proposed a KDA method that finds discriminant projection bases only in the null space of the within-class scatter matrix. Kernel direct LDA [4] removes the null space of the between-class scatter matrix and hence removes information from the null space of the within-class scatter matrix. Finally, methods that exploit the null space in an iterative manner by using successive projections have been proposed in [16]–[18].

The above problems, to some extent, have been circumvented by the complete kernel Fisher discriminant (CKFD) framework [8], which attempts to harness the discriminative information in both subspaces and is considered the state-of-the-art method. In [8], the authors prove that KDA is equivalent to first performing a kernel principal component analysis (KPCA) for dimensionality reduction and then the LDA. Features extracted from the two complementary subspaces are then combined by a weighted distance measure in the recognition phase [8].

However, some important problems and open questions related to complete KDA still need to be addressed. In particular, is it possible to perform complete KDA directly and without performing the KPCA step? How can feature scaling techniques, such as eigenspectrum regularization, be applied in a kernel-based fashion? How can score-level and feature-level schemes, as suggested in [14], [19]–[21], be applied in a kernel-based fashion? How efficient is it to extract features separately from the two subspaces and then use a score-level fusion scheme?

In this brief, we address the above problems by showing, for the first time, how to perform the eigen analysis of the within-class scatter matrix directly in the feature space. We

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

2

IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS

show that, while CKFD indeed provides the eigenvectors of the two complementary subspaces, it does not provide the eigenspectrum of the range space and thus does not allow for the application of eigenspectrum regularization techniques or feature-level schemes. On the contrary, the proposed eigen analysis of the within-class scatter matrix in the feature space provides, for the first time, the eigenspectrum of its range space and the corresponding eigenvectors as well as the eigenvectors spanning its null space. A by-product of our analysis is that the application of eigenspectrum regularization as well as feature-level schemes becomes now straightforward. Based on these techniques, we then propose an eigenspectrum-based regularized KDA algorithm (ER-KDA). Finally, we propose to combine the proposed ER-KDA with the nonlinear extension of the robust gradient-based kernel of [22], [23]. The proposed nonlinear kernel is particularly suitable for face recognition/verification applications which require robustness against outliers caused by occlusions and illumination changes. We applied the proposed framework to several popular databases (extended Yale B, AR, XM2VTS) and achieved state-of-the-art performance for most of our experiments.

The rest of this brief is organized as follows. Section II presents the necessary definitions and discusses prior work in KDA. In Section III, we propose our kernel-based discriminant feature extraction method based on the eigenspectrum regularization of the within-class scatter matrix (ER-KDA). In Section IV, we propose a robust radial basis function (RBF) kernel based on the cosine of the gradient orientations. Section V presents our experimental results. Finally, we draw conclusions in Section VI.

II. DEFINITIONS AND PRIOR WORK

We assume that we are given N training samples $\mathbf{X}^{\mathcal{X}} = [\mathbf{x}_1 | \cdots | \mathbf{x}_N]$, where $\mathbf{x}_i \in \Re^K$ and $\mathcal{X}$ denotes our input vector space. Let k be a positive definite kernel that satisfies the Mercer's conditions [24]. The kernel k defines an arbitrary-dimensional Hilbert space $\mathcal{H}$ (hereinafter, referred to as feature space) through an implicit mapping $\phi(.) \colon \Re^K \to \mathcal{H}$ such that $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$. We denote by $\mathbf{X}^{\mathcal{H}} \triangleq [\phi(\mathbf{x}_1) | \cdots | \phi(\mathbf{x}_N)]$ the data matrix in $\mathcal{H}$ and by $\tilde{\mathbf{X}}^{\mathcal{H}} \triangleq [\phi(\mathbf{x}_1) - \mathbf{m}^{\mathcal{H}} | \cdots | \phi(\mathbf{x}_N) - \mathbf{m}^{\mathcal{H}}]$ its centralized version, where $\mathbf{m}^{\mathcal{H}} \triangleq (1/N) \sum_{i=1}^{N} \phi(\mathbf{x}_i)$ is the mean vector in $\mathcal{H}$. Let us also define the symmetric positive (semi) definite kernel matrix $\mathbf{K} \triangleq [k(\mathbf{x}_i, \mathbf{x}_j)] \in \mathbb{R}^{N \times N}$ and its centralized version $\tilde{\mathbf{K}} \triangleq (\tilde{\mathbf{X}}^{\mathcal{H}})^T \tilde{\mathbf{X}}^{\mathcal{H}} = (\mathbf{I} - (1/N)\mathbf{E}_N)\mathbf{K}(\mathbf{I} - (1/N)\mathbf{E}_N)$, where $\mathbf{E}_N$ is an $N \times N$ matrix of ones.

KDA aims at finding discriminant projection bases by exploiting class-label information in the feature space. We assume that our training set consists of C classes $\mathcal{C}_1, \ldots, \mathcal{C}_C$. N_c denotes the cardinality of set $\mathcal{C}_c$. We define the between-class, within-class, and total scatter matrices $\mathbf{S}_b^{\mathcal{H}}$, $\mathbf{S}_w^{\mathcal{H}}$, and $\mathbf{S}_t^{\mathcal{H}}$ in $\mathcal{H}$ as

$$\mathbf{S}_b^{\mathcal{H}} \triangleq \sum_{c=1}^{C} N_c \left(\mathbf{m}_c^{\mathcal{H}} - \mathbf{m}^{\mathcal{H}} \right) \left(\mathbf{m}_c^{\mathcal{H}} - \mathbf{m}^{\mathcal{H}} \right)^T \qquad (1)$$

$$\mathbf{S}_w^{\mathcal{H}} \triangleq \sum_{c=1}^{K} \sum_{\mathbf{x}_i \in \mathcal{C}_c} \left(\phi(\mathbf{x}_i) - \mathbf{m}_c^{\mathcal{H}} \right) \left(\phi(\mathbf{x}_i) - \mathbf{m}_c^{\mathcal{H}} \right)^T \qquad (2)$$

$$\mathbf{S}_t^{\mathcal{H}} \triangleq \sum_{c=1}^{K} \sum_{\mathbf{x}_i} \left(\phi(\mathbf{x}_i) - \mathbf{m}^{\mathcal{H}} \right) \left(\phi(\mathbf{x}_i) - \mathbf{m}^{\mathcal{H}} \right)^T \qquad (3)$$

where $\mathbf{m}_c^{\mathcal{H}} = (1/N_c) \sum_{\mathbf{x}_i \in \mathcal{C}_c} \phi(\mathbf{x}_i)$ is the mean vector of each class.

The three basic optimization problems for finding discriminant projection bases $\mathbf{U}$ are as follows [8], [10], [25], [26]:

$$\mathbf{U}_1 = \max_{\mathbf{U}} \ \mathrm{tr}\left[\mathbf{U}^T \mathbf{S}_b^{\mathcal{H}} \mathbf{U}\right]$$

$$\mathrm{s.t} \qquad \mathbf{U}^T \mathbf{S}_w^{\mathcal{H}} \mathbf{U} = \mathbf{I} \qquad (4)$$

$$\mathbf{U}_2 = \max_{\mathbf{U}} \ \mathrm{tr}\left[\mathbf{U}^T \left(\mathbf{S}_b^{\mathcal{H}} - \mathbf{S}_w^{\mathcal{H}}\right) \mathbf{U}\right]$$

$$\mathrm{s.t} \qquad \mathbf{U}^T \mathbf{U} = \mathbf{I} \qquad (5)$$

and

$$\mathbf{U}_3 = \max_{\mathbf{U}} \ \mathrm{tr}\left[\mathbf{U}^T \left(\mathbf{S}_b^{\mathcal{H}} - \mathbf{S}_w^{\mathcal{H}}\right) \mathbf{U}\right]$$

$$\mathrm{s.t} \qquad \mathbf{U}^T \mathbf{U} = \mathbf{I}, \mathbf{U}^T \mathbf{S}_w^{\mathcal{H}} \mathbf{U} = \mathbf{0}. \qquad (6)$$

The optimization problem (4) was solved in [8] and [25]. The CKFD framework [8] solves optimization problems (4) and (6) simultaneously by projecting $\mathbf{S}_w^{\mathcal{H}}$ onto the non-null space of $\mathbf{S}_t^{\mathcal{H}}$. The solution of optimization problem (5) was recently proposed in [10] and then corrected in [26].

As one might observe, the CKFD framework is complete, but indirect. That is, while the method formulates and solves the optimization problems (4) and (6) corresponding to the range and the null space of the within-class scatter matrix, respectively, the optimal solutions are obtained separately after KPCA has been performed. A by-product of this is that features are extracted from these complementary subspaces separately and only a score level fusion scheme can be used. Additionally, such an approach does not allow the application of eigenspectrum regularization techniques since essentially such methods require the spectrum of the within-class scatter matrix. To alleviate these problems, in the following section we propose a direct complete KDA method based on the complete eigen analysis of the within-class scatter matrix $\mathbf{S}_w^{\mathcal{H}}$.

III. REGULARIZED KDA

A. Eigen Analysis of $\mathbf{S}_w^{\mathcal{H}}$

In this section, we present, for the first time, the eigen analysis of the within-class scatter matrix $\mathbf{S}_w^{\mathcal{H}}$ directly in the feature space. We show that this eigen analysis has the following results.

1) $N - C$ orthonormal eigenvectors $\mathbf{U}_n$, which correspond to the $N - C$ nonzero eigenvalues Λ_n.
2) C orthonormal eigenvectors $\mathbf{U}_l$, which correspond to the null eigenvalues and additionally satisfy $\mathbf{U}_l^T \mathbf{S}_t^{\mathcal{H}} \mathbf{U}_l \neq \mathbf{0}$. $\mathbf{U}_l$ constitutes the useful null-space of $\mathbf{S}_w^{\Phi}$ by projecting every sample to the center of its class.
3) A very large number of eigenvectors $\mathbf{U}_f$, which correspond to null eigenvalues and additionally satisfy $\mathbf{U}_f^T \mathbf{S}_t^{\mathcal{H}} \mathbf{U}_f = \mathbf{0}$. $\mathbf{U}_f$ projects all samples to the same point and therefore cannot be used for classification.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS

3

Let us first define the block $\mathbf{M}_c \triangleq (1/N_c)\mathbf{E}_{N_c}$ and the block diagonal matrix $\mathbf{M}$

$$\mathbf{M} \triangleq \text{diag}[\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_C]. \tag{7}$$

Remarks:

1) $\mathbf{M}$ is idempotent, i.e., $\mathbf{M}^2 = \mathbf{M}$.
2) $\mathbf{I} - \mathbf{M}$ is idempotent, i.e., $(\mathbf{I} - \mathbf{M})^2 = \mathbf{I} - \mathbf{M}$.
3) $\mathbf{M}$ has C eigenvectors corresponding to C nonzero eigenvalues.
4) $\mathbf{I} - \mathbf{M}$ has $N - C$ eigenvectors corresponding to $N - C$ nonzero eigenvalues.
5) Let $\mathbf{A}$ be a matrix with columns linearly independent. Then, $\mathbf{A}\mathbf{M}\mathbf{A}^T$ and $\mathbf{A}(\mathbf{I} - \mathbf{M})\mathbf{A}^T$ have C and $N - C$ eigenvectors corresponding to C and $N - C$ nonzero eigenvalues, respectively.

Using $\mathbf{M}$, we write $\mathbf{S}_b^{\mathcal{H}}$ as

$$\mathbf{S}_b^{\mathcal{H}} = \tilde{\mathbf{X}}^{\mathcal{H}}\mathbf{M}(\tilde{\mathbf{X}}^{\mathcal{H}})^T \tag{8}$$

while using $\mathbf{S}_w^{\mathcal{H}} = \mathbf{S}_t^{\mathcal{H}} - \mathbf{S}_b^{\mathcal{H}}$, $\mathbf{S}_w^{\mathcal{H}}$ takes the form

$$\mathbf{S}_w^{\mathcal{H}} = \tilde{\mathbf{X}}^{\mathcal{H}}(\mathbf{I} - \mathbf{M})(\tilde{\mathbf{X}}^{\mathcal{H}})^T. \tag{9}$$

From Remark 5, $\mathbf{S}_b^{\mathcal{H}}$ and $\mathbf{S}_w^{\mathcal{H}}$ have C and $N - C$ eigenvectors corresponding to nonzero eigenvalues, respectively.

Direct eigen analysis of $\mathbf{S}_w^{\mathcal{H}}$ is intractable since the mapping $\phi(.)$ is implicit. In the following, we show how to compute $\mathbf{U}_n$, Λ_n, and $\mathbf{U}_l$ directly.

To calculate $\mathbf{U}_n$ and Λ_n, we need the following theorem [2]:

Theorem 1: Define matrices $\mathbf{A}$ and $\mathbf{B}$ such that $\mathbf{A} = \boldsymbol{\Phi}\boldsymbol{\Phi}^T$ and $\mathbf{B} = \boldsymbol{\Phi}^T\boldsymbol{\Phi}$. Let $\mathbf{U}_A$ and $\mathbf{U}_B$ be the eigenvectors corresponding to the nonzero eigenvalues Λ_A and Λ_B of $\mathbf{A}$ and $\mathbf{B}$, respectively. Then, $\Lambda_A = \Lambda_B$ and $\mathbf{U}_A = \boldsymbol{\Phi}\mathbf{U}_B\Lambda_A^{-(1/2)}$.

Using Remark 2, we can write $\mathbf{S}_w^{\mathcal{H}}$ as

$$\begin{aligned}
\mathbf{S}_w^{\mathcal{H}} &= \tilde{\mathbf{X}}^{\mathcal{H}}(\mathbf{I} - \mathbf{M})(\tilde{\mathbf{X}}^{\mathcal{H}})^T \\
&= \left(\tilde{\mathbf{X}}^{\mathcal{H}}(\mathbf{I} - \mathbf{M})\right)\left(\tilde{\mathbf{X}}^{\mathcal{H}}(\mathbf{I} - \mathbf{M})\right)^T.
\end{aligned} \tag{10}$$

By Theorem 1, Λ_n corresponds to the $N - C$ largest eigenvalues obtained by the eigen analysis of $\mathbf{K}_w$

$$\mathbf{K}_w = \left(\tilde{\mathbf{X}}^{\mathcal{H}}(\mathbf{I} - \mathbf{M})\right)^T\left(\tilde{\mathbf{X}}^{\mathcal{H}}(\mathbf{I} - \mathbf{M})\right) = (\mathbf{I} - \mathbf{M})\tilde{\mathbf{K}}(\mathbf{I} - \mathbf{M}). \tag{11}$$

Additionally, let $\mathbf{Q}_n$ be the corresponding eigenvectors of $\mathbf{K}_w$. Then, also by Theorem 2

$$\mathbf{U}_n = \tilde{\mathbf{X}}^{\mathcal{H}}(\mathbf{I} - \mathbf{M})\mathbf{Q}_n\Lambda_n^{-\frac{1}{2}}. \tag{12}$$

Unfortunately, we cannot derive $\mathbf{U}_l$ from the Eigen analysis of $\mathbf{K}_w$. To show this, let $\mathbf{Q}_l$ be the complementary subspace and define the projection bases $\mathbf{U}_s = \tilde{\mathbf{X}}^{\mathcal{H}}(\mathbf{I} - \mathbf{M})\mathbf{Q}_l$. We have

$$\mathbf{U}_s^T\mathbf{U}_s = \mathbf{Q}_l^T(\mathbf{I} - \mathbf{M})(\tilde{\mathbf{X}}^{\mathcal{H}})^T\tilde{\mathbf{X}}^{\mathcal{H}}(\mathbf{I} - \mathbf{M})\mathbf{Q}_l = \mathbf{0} \tag{13}$$

which further gives $\mathbf{U}_s = \mathbf{0}$.

To find $\mathbf{U}_l$, we write $\mathbf{U}_l = \tilde{\mathbf{X}}^{\mathcal{H}}\mathbf{V}_l$, $\mathbf{V}_l \in \mathfrak{R}^{N \times C}$. Additionally, $\mathbf{U}_l$ satisfies

$$\mathbf{U}_l^T\mathbf{S}_w^{\mathcal{H}}\mathbf{U}_l = \mathbf{V}_l^T\tilde{\mathbf{K}}(\mathbf{I} - \mathbf{M})\tilde{\mathbf{K}}\mathbf{V}_l = \mathbf{0}. \tag{14}$$

From Remark 5, $\tilde{\mathbf{K}}(\mathbf{I} - \mathbf{M})\tilde{\mathbf{K}}$, $(\tilde{\mathbf{K}}^T = \tilde{\mathbf{K}})$ has $N - C$ nonzero eigenvalues. Moreover, we must impose $\mathbf{U}_l$ to be orthonormal (i.e., $\mathbf{U}_l^T\mathbf{U}_l = \mathbf{I}$). These two constraints [i.e., (14) and orthonormality] can be satisfied by choosing $\mathbf{V}_l = \Xi_l\mathbf{P}^{-1/2}$, where Ξ_l is found by performing eigen analysis of $\tilde{\mathbf{K}}(\mathbf{I} - \mathbf{M})\tilde{\mathbf{K}}$ and keeping the C eigenvectors that correspond to the zero eigenvalues and $\mathbf{P} = \Xi_l^T\tilde{\mathbf{K}}\Xi_l$. It is not difficult to verify that

$$\mathbf{U}_l^T\mathbf{U}_l = \mathbf{P}^{-\frac{1}{2}}\Xi_l^T(\tilde{\mathbf{X}}^{\mathcal{H}})^T\tilde{\mathbf{X}}^{\mathcal{H}}\Xi_l\mathbf{P}^{-\frac{1}{2}} = \mathbf{I} \tag{15}$$

$\mathbf{U}_l^T\mathbf{S}_w^{\mathcal{H}}\mathbf{U}_l = \mathbf{0}$ and $\mathbf{U}_l^T\mathbf{S}_t^{\mathcal{H}}\mathbf{U}_l \neq \mathbf{0}$.

Remark 6) $\mathbf{U}_l^T\mathbf{U}_n = \mathbf{0}$.

The proof of the above remark can be found in Appendix II.

Finally, the remaining null space $\mathbf{U}_f$ coincides with the null space of $\mathbf{S}_t^{\mathcal{H}}$, i.e., $\mathbf{U}_f^T\mathbf{S}_w^{\mathcal{H}}\mathbf{U}_f = \mathbf{U}_f^T\mathbf{S}_t^{\mathcal{H}}\mathbf{U}_f = \mathbf{0}$ according to Lemma B1 of [8].

B. ER-KDA

Based on our results in the previous section, we perform the eigen analysis of $\mathbf{S}_w^{\mathcal{H}}$ and extract discriminant features by applying eigenspectrum regularization and feature level fusion [14], [19]–[21]. The regularization process results in a modified within-class scatter matrix which is now invertible. More specifically, we fit a $1/f$ function to the available spectrum Λ_n [21]. We then define a modified within-class scatter matrix as follows:

$$\mathbf{S}_{w,r}^{\mathcal{H}} = \mathbf{U}_w\Lambda_w\mathbf{U}_w^T \tag{16}$$

where $\mathbf{U}_w = [\mathbf{U}_n \; \mathbf{U}_l]$ and Λ_w is the regularized spectrum consisting of Λ_n ($N - C$ values) and C values all equal to our estimated function evaluated at the $N - C + 1$ sample.

To find discriminant projection bases, we use a feature level fusion approach by diagonalizing $\mathbf{S}_{w,r}^{\mathcal{H}}$. More specifically, we transform the training data using $\Lambda_w^{(-1/2)} \in \mathfrak{R}^N$ for feature weighting

$$\begin{aligned}
\tilde{\mathbf{X}}_w &= \Lambda_w^{-\frac{1}{2}}\mathbf{U}_w^T\tilde{\mathbf{X}}^{\mathcal{H}} \\
&= \Lambda_w^{-\frac{1}{2}}\begin{bmatrix} \Lambda_n^{-\frac{1}{2}}\mathbf{Q}_n^T(\mathbf{I} - \mathbf{M})(\tilde{\mathbf{X}}^{\mathcal{H}})^T \\ \mathbf{P}^{-\frac{1}{2}}\mathbf{V}_l^T(\tilde{\mathbf{X}}^{\mathcal{H}})^T \end{bmatrix}\tilde{\mathbf{X}}^{\mathcal{H}} \\
&= \Lambda_w^{-\frac{1}{2}}\begin{bmatrix} \Lambda_n^{-\frac{1}{2}}\mathbf{Q}_n^T(\mathbf{I} - \mathbf{M}) \\ \mathbf{P}^{-\frac{1}{2}}\mathbf{V}^T \end{bmatrix}\tilde{\mathbf{K}} \\
&= \acute{\mathbf{U}}_w^T\tilde{\mathbf{K}}
\end{aligned} \tag{17}$$

and optimize

$$\begin{aligned}
\acute{\mathbf{U}}_o &= \max_{\acute{\mathbf{U}}} \text{tr}\left(\acute{\mathbf{U}}^T\acute{\mathbf{S}}_b\acute{\mathbf{U}}\right) \\
&\text{s.t.} \quad \acute{\mathbf{U}}^T\acute{\mathbf{U}} = \mathbf{I}
\end{aligned} \tag{18}$$

where $\acute{\mathbf{S}}_b = \tilde{\mathbf{X}}_w\mathbf{M}(\tilde{\mathbf{X}}_w)^T$. Finally, we compute $\acute{\mathbf{X}} = \acute{\mathbf{U}}_p^T\tilde{\mathbf{X}}_w$, where $\acute{\mathbf{U}}_p$ gathers the eigenvectors corresponding to the p largest eigenvalues of $\acute{\mathbf{U}}_o$.

At the recognition stage, given a probe image $\mathbf{y}$, we compute

$$\acute{\mathbf{y}} = (\acute{\mathbf{U}}_w\acute{\mathbf{U}}_p)^T\tilde{\mathbf{k}}(\mathbf{y}) \tag{19}$$

where $[\tilde{\mathbf{k}}(\mathbf{y})]_j = k(\mathbf{x}_j, \mathbf{y}) - (1/N)\sum_{i=1}^N k(\mathbf{x}_j, \mathbf{x}_i) - (1/N)\sum_{i=1}^N k(\mathbf{x}_j, \mathbf{y}) + (1/N^2)\sum_{l=1}^N\sum_{m=1}^N k(\mathbf{x}_l, \mathbf{x}_m)$. Classification is performed using the nearest neighbor rule based on normalized correlation.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

4

IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS
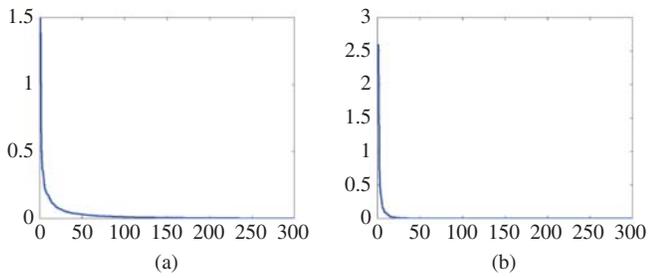
Fig. 1.   Spectrum of (a) $(\mathbf{I} - \mathbf{M})\tilde{\mathbf{K}}(\mathbf{I} - \mathbf{M})$ and (b) $\tilde{\mathbf{K}}(\mathbf{I} - \mathbf{M})\tilde{\mathbf{K}}$.

Fig. 2.   We assume that (a) part of the face region and (b) corresponding "baboon" patch are visually dissimilar. Similarly, (c) face region is "visually unrelated" with (d) image region corresponding to the scarf.

C. Comparison With the Methods in [19] and [27] and the Linear Version of the Method in [14]

Eigenspectrum regularization in the feature domain has been also recently proposed in [19] and [27]. Both schemes express the projection bases as a linear combination of the training samples $\mathbf{U} = \tilde{\mathbf{X}}^{\mathcal{H}}\mathbf{V}$ and project $\mathbf{S}_w^{\mathcal{H}}$ according to

$$\mathbf{U}^T \mathbf{S}_w^{\mathcal{H}} \mathbf{U} = \mathbf{V}^T \tilde{\mathbf{K}}(\mathbf{I} - \mathbf{M})\tilde{\mathbf{K}}\mathbf{V}. \qquad (20)$$

Next, they apply the regularization of [14] to the eigenspectrum of $\tilde{\mathbf{K}}(\mathbf{I} - \mathbf{M})\tilde{\mathbf{K}}$. This procedure is substantially different from ours, since, as we have shown, the positive eigenspectrum of $\mathbf{S}_w^{\mathcal{H}}$ can be found by the eigen analysis of $(\mathbf{I} - \mathbf{M})\tilde{\mathbf{K}}(\mathbf{I} - \mathbf{M})$ and not of $\tilde{\mathbf{K}}(\mathbf{I} - \mathbf{M})\tilde{\mathbf{K}}$ (the matrices do not have the same spectrum). This, in turn, has two effects.

First, the weighting is different, since the regularization is based on estimating and modeling the nonzero part of the spectrum and the above two matrices do not have the same spectrum. More specifically, by Remark 2 and Theorem 2, it is straightforward to show that the spectrum of $\tilde{\mathbf{K}}(\mathbf{I} - \mathbf{M})\tilde{\mathbf{K}}$ can be also obtained from the eigen analysis of $(\mathbf{I} - \mathbf{M})\tilde{\mathbf{K}}^2(\mathbf{I} - \mathbf{M})$. Because of the power of the two, this spectrum decays much faster compared to the decay of the spectrum of $(\mathbf{I} - \mathbf{M})\tilde{\mathbf{K}}(\mathbf{I} - \mathbf{M})$, which is the real spectrum of $\mathbf{S}_w^{\mathcal{H}}$. In contrast to our findings, this is perhaps why the authors in [19] argued that a function of the type $1/f$ cannot be used in kernel-based eigenspectrum regularization. As an example, Fig. 1 illustrates how the spectrum of the two matrices decays. Second, the feature weighting of (17) is applied to different features. This is because the eigenvectors spanning the range space of the two matrices are different.

The proposed methods require the eigen analysis of kernel matrices, thus the complexity is of order $O(N^3)$,[1] similar to that in [5], [8], and [28]. This is in contrast to the linear method of [14], which performs eigenspectrum regularization directly in the input space (thus it does not use inner products) and therefore has complexity $O(K^3)$. Thus our approach can be used to speed up [14] in the case of small-sized problems where $N \ll K$. This can be simply achieved by using the linear kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$.

---

[1]Unfortunately, the order $O(N^3)$ of the complexity would become computationally prohibitive for tasks such as image retrieval in large databases.
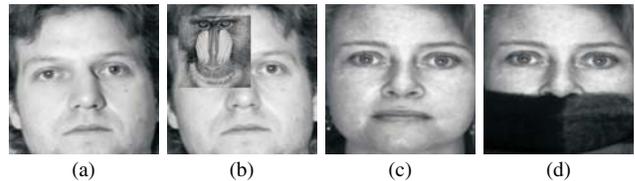
IV. NONLINEAR GRADIENT-BASED KERNEL FOR CONSISTENT FACE SIMILARITY

Given any pair $I_i(\mathbf{m})$, $i = 1, 2$, $\mathbf{m} \in \mathcal{Z}$ of registered facial images, we measure their similarity using a kernel that is equivalent to an inner product in an arbitrary-dimensional Hilbert space. We then combine the proposed ER-KDA described in the previous section with this kernel for robust face recognition and verification. More specifically, we propose to combine the proposed ER-KDA with the nonlinear extension of the robust gradient-based kernel proposed in [22].

Following [22], we replace I_i with the corresponding orientation map $\Theta_i(\mathbf{m}) = \arctan(G_{i,y}(\mathbf{m})/G_{i,x}(\mathbf{m}))$, where $G_{i,x}(\mathbf{m}) = h_x(\mathbf{m}) \star I_i(\mathbf{m})$, $G_{i,y}(\mathbf{m}) = h_y(\mathbf{m}) \star I_i(\mathbf{m})$, and h_x, h_y are filters used to approximate the ideal differentiation operator along the image horizontal and vertical direction, respectively. Given Θ_i, we compute the orientation difference function $\Delta\Theta(\mathbf{m}) = \Theta_1(\mathbf{m}) - \Theta_2(\mathbf{m})$ and define a global similarity criterion as follows:

$$S(I_1, I_2) \triangleq \frac{1}{K} \sum_{\mathbf{m}} \cos \Delta\Theta(\mathbf{m}) \qquad (21)$$

where K is the total number of pixels in I_i.

The direction of image gradients has been occasionally used in image-based object/face recognition as a data representation relatively insensitive to illumination variations [29]–[34], [34], [35]. What is highlighted in [22] and [23] is that this representation, combined with the cosine kernel, can be efficiently used to reject outliers caused by possible global or local mismatches. To show this, let us assume that two images are "visually unrelated" (or dissimilar) so that they locally do not match. As an example, we assume that the part of the face region in Fig. 2(a) and the corresponding "baboon" patch in Fig. 2(b) are visually dissimilar. Similarly, the face region in Fig. 2(c) is "visually unrelated" with the image regions corresponding to the scarf in Fig. 2(d) respectively. Given two "visually unrelated" images, then, it is not difficult to show that, under some rather mild assumptions [23], [22]

$$\sum_{\mathbf{m}} \cos \Delta\Theta(\mathbf{m}) \simeq 0 \qquad (22)$$

holds and therefore $S \to 0$.

Note that the above similarity is nonlinear with respect to gradient orientations but linear with respect to the concatenation of cosines and sines of gradient orientations. Let us denote by $\boldsymbol{\theta}_i$ the $K-$dimensional vector obtained by writing the orientation map Θ_i in lexicographic ordering. By using

TABLE I
AVERAGE RECOGNITION RATES AND STANDARD DEVIATIONS ON THE EXTENDED YALE B DATABASE

5-10-20	Cosine orientation			RBF-cosine orientation			Intensity			RBF-intensity		
KPCA	95.4(1.12)	98.5(0.25)	99.4(0.082)	96.0(1.07)	98.5(0.25)	100(0)	76.1(0.99)	85.4(0.89)	90.6(0.75)	76.6(1.02)	86.8(0.83)	91.1(0.75)
KFisherFaces	98.2(0.69)	99.6(0.07)	99.7(0.07)	98.6(0.7)	99.6(0.07)	100(0)	72.23(1.84)	83.4(1.18)	88.4(0.92)	72.9(1.67)	84.4(1.11)	91.9(0.88)
KPCA+LDA	98.1(0.67)	99.6(0.07)	99.6(0.07)	98.6(0.6)	99.6(0.07)	100(0)	74.7(1.45)	85.7(1.2)	90.4(0.79)	75.4(1.63)	86.1(1.16)	93.5(0.73)
ER-KDA	99.45(0.08)	100(0)	100.0(0)	100(0)	100(0)	100(0)	76.7(1.75)	88.4(1.12)	93.1(1.19)	78.3(1.61)	89.3(1.06)	95.5(0.8)
R-KDA	98.1(0.7)	99.6(0.07)	99.7(0.07)	98.4(0.72)	99.6(0.07)	100(0)	72.8(1.75)	84.3(1.19)	88.9(0.85)	73.1(1.7)	84.4(37)	91.9(0.88)

TABLE II
RECOGNITION RATES ON THE EXT. YALE B DATABASE USING
THE SRC METHOD [40]

Method	Recognition via sparse representation %
5-Train	76.1
10-Train	88.4
20-Train	92.1

TABLE III
RECOGNITION RATES FOR EXPERIMENT 1, VARIABILITY IN
FACIAL EXPRESSION ON THE AR DATABASE

Method	Cos. Or.	RBF-Cos. Or.	Int.	RBF-Int.
KPCA	94.67	94.67	91.67	94.67
KFisherFaces	97.67	99.33	94.33	96.33
KPCA+LDA	98.33	98.33	97.00	98.00
ER-KDA	100.00	100.00	99.00	99.00
R-KDA	98.66	98.66	96.33	96.33

the mapping[2] [22]

$$\boldsymbol{\phi}(\boldsymbol{\theta}_i) = \frac{1}{\sqrt{K}} \left[\cos(\boldsymbol{\theta}_i)^T \ \sin(\boldsymbol{\theta}_i)^T \right]^T \quad (23)$$

where $\cos(\boldsymbol{\theta}_i) = [\cos(\boldsymbol{\theta}_i(1)), \ldots, \cos(\boldsymbol{\theta}_i(K))]^T$ and $\sin(\boldsymbol{\theta}_i) = [\sin(\boldsymbol{\theta}_i(1)), \ldots, \sin(\boldsymbol{\theta}_i(K))]^T$ the above similarity takes the form $S(\boldsymbol{\phi}(\boldsymbol{\theta}_1), \boldsymbol{\phi}(\boldsymbol{\theta}_2)) \triangleq \boldsymbol{\phi}(\boldsymbol{\theta}_1)^T \boldsymbol{\phi}(\boldsymbol{\theta}_2)$, while the distance between $\boldsymbol{\phi}(\boldsymbol{\theta}_1)$ and $\boldsymbol{\phi}(\boldsymbol{\theta}_2)$ is given by $D(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \triangleq (1/2)||\boldsymbol{\phi}(\boldsymbol{\theta}_1) - \boldsymbol{\phi}(\boldsymbol{\theta}_2)||^2 = 1 - S$.

Finally, using S, we define the proposed nonlinear kernel as

$$k(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = \exp \frac{D(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{\sigma^2}. \quad (24)$$

Theorem 2: The proposed kernel is positive semidefinite. See Appendix I for the proof.

V. EXPERIMENTAL RESULTS

We carried out experiments for both face recognition (extended Yale B [36] and AR [37] databases) and verification (XM2VTS [38] database). For all experiments, we manually aligned the face images and used 64×64 image resolution. Finally, all rates reported correspond to the dimension that produces the highest recognition/verification accuracy.

For all databases, we compared the performance of the proposed ER-KDA with that of KPCA, Kernel FisherFaces, KPCA plus LDA, and the R-KDA method proposed in [39]. We used the linear "real" version of the cosine orientation kernel [22], the proposed nonlinear RBF kernel of (24), the linear kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$ operating on pixel intensities, as well as the RBF $k(\mathbf{x}_i, \mathbf{x}_j) = \exp ||\mathbf{x}_i - \mathbf{x}_j||^2 / \sigma^2$. In all cases, σ was found by cross-validation in the training set.

A. Extended Yale B Database

The extended Yale B database [36] contains 16 128 images of 38 subjects under 9 poses and 64 illumination conditions. We used a subset that consists of 64 near frontal images for

---

[2]To be more precise, this is the "real" version of the complex kernel proposed in [22].

each subject. For training, we randomly selected a subset with 5, 10 and 20 images per subject. For testing, we used the remaining images. Finally, we performed 20 different random realizations of the training/test sets.

Table I summarizes the obtained results. As we may observe, our ER-KDA performs the best using all tested kernels (linear and RBF). Additionally, Table II summarizes the best recognition rates achieved by the method proposed in [40] in our data. Overall, compared to previously proposed RBF kernel methods (KPCA, Kernel FisherFaces, KPCA plus LDA, R-KDA) and the state-of-the-art method in [40], the proposed combined approach ER-KDA plus cosine orientation gives a performance improvement which is (in absolute terms) more than 25%, 12%, and 5% when using 5, 10, and 20 training samples, respectively.

B. AR Database

The AR database [37] consists of more than 4000 frontal view face images of 126 subjects. Each subject has up to 26 images taken in two sessions. The first session contains 13 images, numbered from 1 to 13, including different facial expressions (1–4), illumination changes (5–7), and different occlusions under different illumination changes (8–13). The second session duplicates the first session 2 weeks later. We randomly selected a subset with 100 subjects and investigated the robustness of our scheme for the case of facial expressions, illumination variations, and occlusions. More specifically, we carried out the following experiments: In experiment 1, we used images 1–4 of session 1 for training and images 2–4 of session 2 for testing. In experiment 2, we used images 1–4 of session 1 for training and images 5–7 of session 2 for testing. In experiment 3, we used images 1–4 of session 1 for training and images 8–13 of session 2 for testing.

Tables III–V summarize our results. As we can see, the cosine-orientation methods largely outperform the corresponding intensity-based counterparts. This performance improvement is particularly evident for the case of occlusions

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

6

IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS

TABLE IV

RECOGNITION RATES FOR EXPERIMENT 2, ILLUMINATION
CHANGES ON THE AR DATABASE

Method	Cos. Or.	RBF-Cos. Or.	Int.	RBF-Int.
KPCA	100.00	100.00	94.67	91.33
KFisherFaces	100.00	100.00	92.67	94.66
KPCA+LDA	100.00	100.00	82.33	82.33
ER-KDA	100.00	100.00	92.33	92.33
R-KDA	100.00	100.00	93.33	94.00

TABLE V

RECOGNITION RATES FOR EXPERIMENT 3, OCCLUSION-ILLUMINATION
ON THE AR DATABASE

Method	Cos. Or.	RBF-Cos. Or.	Int.	RBF-Int.
KPCA	92.5	93.5	37.72	36.17
KFisherFaces	94.33	95.17	45.58	31.83
KPCA+LDA	87.33	89.67	23.67	24.17
ER-KDA	94.67	95.67	23.00	24.75
R-KDA	94.00	94.67	27.67	29.66

TABLE VI

RECOGNITION RATES ON THE AR DATABASE USING
THE SRC METHOD [40]

Method	Recognition via sparse representation %
Experiment 1	96.0
Experiment 2	92.3
Experiment 3	66.0

TABLE VII

TER ON THE XM2VTS DATABASE

Method	Cos. Or.	RBF-Cos. Or.	Int.	RBF-Int.
KPCA	1.1	1.07	2.6	2.5
KFisherFaces	0.75	0.77	1.74	1.8
KPCA+LDA	0.89	0.86	1.51	1.45
ER-KDA	0.42	0.42	1.1	1.1
R-KDA	0.78	0.78	1.78	1.72

TABLE VIII

BEST TER IN XM2VTS DATABASE

Methods	Best of ICPR2000 [41]	Best of AVBPA2003 [42]	Best of ICB2006 [43]	Our approach
TER%	4.8	1.47	0.96	0.42

TABLE IX

EXECUTION TIME FOR ALL TESTED METHODS IN SECONDS. THE
EXECUTION TIME WAS MEASURED IN A MACHINE RUNNING AN INTEL
CORE I7, 3.4 GHZ, 8 GB RAM, WITH A 64-BIT MATLAB

Method	Ext. YALE B	AR	XM2VTS
KPCA	0.62	0.12	0.23
KFisherFaces	0.83	0.25	0.29
KPCAplusLDA	1.58	0.32	0.56
ER-KDA	1.22	0.29	0.51
R-KDA	0.78	0.23	0.27

(experiment 3), where the gain in recognition accuracy goes up to 50%. However, our ER-KDA does not always perform the best. This is particularly evident for the case of the RBF kernel in experiment 3. This result is somewhat expected, since the testing data in these experiments are significantly different from the ones used for training (having a vast occlusion). Compared to our ER-KDA, this overfitting seems to affect a bit more the KPCA plus LDA method, while Kernel FisherFaces performs the best as it discards the null space information. Note, however, that the use of the cosine-orientation kernels appears to largely circumvent the problem. Overall, the proposed combined approach performs the best in all experiments. Finally, Table VI shows the recognition rates obtained for experiments 1–3 by applying the recently proposed state-of-the-art method of recognition via sparse representation [40]. As we may observe, our combined approach performs significantly better while it is also orders of magnitude faster.

C. XM2VTS Database

We carried out face verification experiments on the test set of Configuration I of the XM2VTS database. The training set contained 200 subjects with three images per subject, which enabled us to apply our kernel combined with the proposed discriminant analysis. The evaluation set contained three images per client for genuine claims and 25 evaluation impostors with eight images per impostor. The testing set contained two images per client and 70 impostors with eight

images per impostor. For additional details on the XM2VTS database and the protocol used, we refer to [38].

As usual, we used the evaluation set to learn the verification decision thresholds for each subject. We obtained three different sets of thresholds corresponding to three different choices for the *false rejection rate* (FRR) achieved at a fixed *false acceptance rate* (FAR): FAE = 0, FRE = 0, and FAE = FRE (*equal error rate*). By fixing these thresholds, we evaluated the performance of our scheme on the testing set by measuring the *total error rate* (TER = FRR + FAE).

Three competitions were conducted with the XM2VTS database in the past decade (2000, 2003, and 2006). Table VII shows the obtained TER for each of the tested methods and Table VIII summarizes the best results of each competition as well as the performance of the proposed scheme. Our method achieved a TER equal to 0.42%, which is the best reported for the XM2VTS database according to the best of our knowledge.

D. Statistical Significance of the Results and Execution Times

Table IX depicts the training time for all the tested methods. Even though the order of complexity of all the tested methods is $O(N^3)$, since both the proposed ER-KDA and KPCA+LDA use both range and null space, their execution time is larger than K FisherFaces, R-KDA, and KPCA.

To evaluate the statistical significance of our results, we used the McNemars test [44]. The McNemars test is a null-hypothesis statistical test based on a Bernoulli model. If the resulting *p*-value is below a desired significance level (for example, 0.02), the null hypothesis is rejected and

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS

7

the difference in performance between two algorithms is considered to be statistically significant. The McNemars test has been widely used to evaluate the statistical significance of the performance improvement between different recognition algorithms [45], [46]. We used the best recognition rates achieved by the original cosine orientation kernel [22] and the proposed nonlinear one and their corresponding pixel-intensity-based counterparts. For all the face recognition experiments (i.e., YALE, AR), with the exception of the experiments corresponding to Table III, we found that $p \ll 0.02$. Thus, we conclude that the performance improvement obtained using the proposed kernels is statistically significant. Furthermore, the performance improvement achieved in the Ext. YALE B database by the proposed ER-KDA method in the case of intensity-based kernels (both linear and RBF) is also statistical significant with $p \ll 0.02$.

VI. CONCLUSION

We have shown, for the first time, how to perform the eigen analysis of the within-class scatter matrix directly in the feature space. The proposed eigen analysis provides, for the first time, the eigenspectrum of its range space and the corresponding eigenvectors as well as the eigenvectors spanning its null space. A by-product of our analysis is that the application of eigenspectrum regularization as well as feature-level schemes becomes now straightforward. Additionally, we combined the proposed ER-KDA with a nonlinear robust gradient-based kernel particularly suitable for face recognition/verification applications which require robustness against outliers caused by occlusions and illumination changes. We applied the proposed framework to several popular databases (extended Yale B, AR, XM2VTS) and achieved state-of-the-art performance for most of our experiments.

APPENDIX I
PROOF OF THEOREM 2

We can write $k(\mathbf{x}_i, \mathbf{x}_j)$ as

$$
\begin{aligned}
k(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) &= \frac{1}{2K} \sum_{m=1}^{K} \cos(\boldsymbol{\theta}_i(m) - \boldsymbol{\theta}_j(m)) \\
&= \frac{1}{2K} \sum_{m=1}^{K} \cos(\boldsymbol{\theta}_i(m)) \cos(\boldsymbol{\theta}_j(m)) \\
&\quad + \frac{1}{2K} \sum_{m=1}^{K} \sin(\boldsymbol{\theta}_i(m)) \sin(\boldsymbol{\theta}_j(m)) \\
&= \boldsymbol{\theta}_i^{\cos T} \boldsymbol{\theta}_j^{\cos} + \boldsymbol{\theta}_i^{\sin T} \boldsymbol{\theta}_j^{\sin}
\end{aligned}
\tag{25}
$$

where $\boldsymbol{\theta}_i^{\cos} = [(1/\sqrt{2K}) \cos(\boldsymbol{\theta}_i(m))]$ and $\boldsymbol{\theta}_i^{\sin} = [(1/\sqrt{2K}) \sin(\boldsymbol{\theta}_i(m))]$. Therefore, we can additionally write

$$
\mathbf{K} = \mathbf{K}_1 + \mathbf{K}_2
\tag{26}
$$

where $\mathbf{K}_1 = [(\boldsymbol{\theta}_i^{\cos})^T \boldsymbol{\theta}_j^{\cos}]$ and $\mathbf{K}_2 = [(\boldsymbol{\theta}_i^{\sin})^T \boldsymbol{\theta}_j^{\sin}]$. $\mathbf{K}_1$ and $\mathbf{K}_2$ are inner-product matrices and therefore are positive (semi)definite. Thus, $\mathbf{K}$ is also positive (semi)definite. Since $\mathbf{K}$ is positive (semi)definite, so is the proposed RBF kernel of (24).

APPENDIX II
PROOF OF REMARK 6

Before we prove Remark 6, we need the following Lemma.
Lemma 1: Let $\mathbf{A} = [\mathbf{a}_1 | \dots | \mathbf{a}_N]$, where $\mathbf{a}_i \in \Re^F$. Then, $\mathbf{A}^T \mathbf{A} = \mathbf{0}$ iff $\mathbf{A} = \mathbf{0}$.

Proof: Let assume $\mathbf{A}^T \mathbf{A} = \mathbf{0}$. Then, the main diagonal $[\mathbf{A}^T \mathbf{A}]_{ii} = \mathbf{a}_i^T \mathbf{a}_i = ||\mathbf{a}_i||_2^2 = 0$. The magnitude of a vector $||\mathbf{a}_i||_2^2 = 0$ iff $\mathbf{a}_i = \mathbf{0}$. Thus, $\mathbf{A} = \mathbf{0}$. The reverse is straightforward.

Proof of Remark 6: By applying the above Lemma to (14) we have

$$
\begin{aligned}
\mathbf{V}_l^T \, \tilde{\mathbf{K}}(\mathbf{I} - \mathbf{M})\tilde{\mathbf{K}}\mathbf{V}_l \\
= \left((\mathbf{I} - \mathbf{M})\tilde{\mathbf{K}}\mathbf{V}_l\right)^T \left((\mathbf{I} - \mathbf{M})\tilde{\mathbf{K}}\mathbf{V}_l\right) = \mathbf{0} \Leftrightarrow \\
(\mathbf{I} - \mathbf{M})\tilde{\mathbf{K}}\mathbf{V}_l = \mathbf{0} \quad \text{and} \quad \mathbf{V}_l^T \tilde{\mathbf{K}}(\mathbf{I} - \mathbf{M}) = \mathbf{0}.
\end{aligned}
\tag{27}
$$

Using $\mathbf{U}_l = \tilde{\mathbf{X}}^{\mathcal{H}} \mathbf{V}_l$, $\mathbf{U}_n = \tilde{\mathbf{X}}^{\mathcal{H}}(\mathbf{I} - \mathbf{M})\mathbf{Q}_n \Lambda_n^{-(1/2)}$, and the above, $\mathbf{U}_l^T \mathbf{U}_n$ can be expanded as

$$
\begin{aligned}
\mathbf{U}_l^T \mathbf{U}_n &= \mathbf{V}_l^T (\tilde{\mathbf{X}}^{\mathcal{H}})^T \tilde{\mathbf{X}}^{\mathcal{H}}(\mathbf{I} - \mathbf{M})\mathbf{Q}_n \Lambda_n^{-\frac{1}{2}} \\
&= \mathbf{V}_l^T \tilde{\mathbf{K}}(\mathbf{I} - \mathbf{M})\mathbf{Q}_n \Lambda_n^{-\frac{1}{2}} = \mathbf{0}.
\end{aligned}
\tag{28}
$$

REFERENCES

[1] W. Zhao, R. Chellappa, P.-J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Survey*, vol. 35, no. 4, pp. 399–458, 2003.

[2] M. Turk and A. P. Pentland, "Eigenfaces for recognition," *J. Cognit. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.

[3] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[4] L. Juwei, K. N. Plataniotis, and A. N. Venetsanopoulos, "Face recognition using kernel direct discriminant analysis algorithms," *IEEE Trans. Neural Netw.*, vol. 14, no. 1, pp. 117–126, Jan. 2003.

[5] M. H. Yang, "Kernel eigenfaces vs. kernel Fisherfaces: Face recognition using kernel methods," in *Proc. 5th IEEE Int. Conf. Automat. Face Gesture Recogn.*, 2002, pp. 215–220.

[6] L. Chengjun, "Gabor-based kernel PCA with fractional power polynomial models for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 5, pp. 572–581, May 2004.

[7] H. Cevikalp, M. Neamtu, M. Wilkes, and A. Barkana, "Discriminative common vectors for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 1, pp. 4–13, Jan. 2005.

[8] J. Yang, A. F. Frangi, J. Yang, D. Zhang, and Z. Jin, "KPCA plus LDA: A complete kernel Fisher discriminant framework for feature extraction and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 230–244, Feb. 2005.

[9] S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas, "Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification," *IEEE Trans. Neural Netw.*, vol. 14, no. 8, pp. 1063–1073, May 2006.

[10] H. Li, T. Jiang, and K. Zhang, "Efficient and robust feature extraction by maximum margin criterion," *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 157–165, Jan. 2006.

[11] H. Cevikalp, M. Neamtu, and M. Wilkes, "Discriminative common vector method with kernels," *IEEE Trans. Neural Netw.*, vol. 17, no. 6, pp. 1550–1565, Nov. 2006.

[12] G. Goudelis, S. Zafeiriou, A. Tefas, and I. Pitas, "Class-specific kernel-discriminant analysis for face verification," *IEEE Trans. Inform. Forensics Security*, vol. 2, no. 3, pp. 570–587, Sep. 2007.

[13] I. Buciu, N. Nikolaidis, and I. Pitas, "Nonnegative matrix factorization in polynomial feature space," *IEEE Trans. Neural Netw.*, vol. 19, no. 6, pp. 1090–1100, Jun. 2008.

[14] X. Jiang, B. Mandal, and A. Kot, "Eigenfeature regularization and extraction in face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 383–394, Mar. 2008.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

8

IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS

[15] S. Zafeiriou and M. Petrou, "Nonlinear nonnegative component analysis," *IEEE Trans. Image Process.*, vol. 19, no. 4, pp. 1050–1066, Apr. 2010.

[16] W. Zheng, L. Zhao, and C. Zou, "Foley-sammon optimal discriminant vectors using kernel approach," *IEEE Neural Netw. Trans.*, vol. 16, no. 1, pp. 1–9, Jan. 2005.

[17] Z. Liang and P. Shi, "Uncorrelated discriminant vectors using a kernel method," *Pattern Recogn.*, vol. 38, no. 2, pp. 307–310, 2005.

[18] T. Xiong, J. Ye, and V. Cherkassky, "Kernel uncorrelated and orthogonal discriminant analysis: A unified approach," in *Proc. IEEE Comput. Vision Pattern Recogn. Comput. Soc. Conf.*, Jun. 2006, pp. 125–131.

[19] X. Jiang, B. Mandal, and A. Kot, "Complete discriminant evaluation and feature extraction in kernel space for face recognition," *Mach. Vision Applicat.*, vol. 20, no. 1, pp. 35–46, 2009.

[20] B. Moghaddam and A. P. Pentland, "Probabilistic visual learning for object representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 696–710, Jul. 1997.

[21] X. Wang and X. Tang, "Dual-space linear discriminant analysis for face recognition," in *Proc. IEEE Comput. Vision Pattern Recogn. Comput. Soc. Conf.*, Jun. 2004, pp. 564–569.

[22] G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "Principal component analysis of image gradient orientations for face recognition," in *Proc. IEEE Automat. Face Gesture Recogn. Workshops Int. Conf.*, Mar. 2011, pp. 553–558.

[23] G. Tzimiropoulos, V. Argyriou, S. Zafeiriou, and T. Stathaki, "Robust FFT-based scale-invariant image registration with image gradients," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 10, pp. 1899–1906, Oct. 2010.

[24] B. Scholkopf and A. Smola, *Learning with Kernels*. Cambridge, MA: MIT Press, 2002.

[25] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Comput.*, vol. 12, no. 10, pp. 2385–2404, 2000.

[26] J. Liu, S. Chen, X. Tan, and D. Zhang, "Comments on efficient and robust feature extraction by maximum margin criterion," *IEEE Trans. Neural Netw.*, vol. 6, no. 18, pp. 1862–1864, Nov. 2007.

[27] S. K. Kim, K. A. Toh, and S. Lee, "Two-fold regularization for kernel Fisher discriminant analysis in face recognition," *IEICE Electron. Exp.*, vol. 6, no. 9, pp. 540–545, 2009.

[28] A. Scholkopf, B. Smola, and K. R. Muller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, 1998.

[29] M. Bichsel, "Strategies of robust object recognition for the automatic identification of human faces," Ph.D. thesis, Swiss Federal Inst. Technol., Zurich, Switzerland, 1991.

[30] R. Brunelli, "Estimation of pose and illuminant direction for face processing," *Image Vision Comput.*, vol. 15, no. 10, pp. 741–748, 1997.

[31] D. Hond and L. Spacek, "Distinctive descriptions for face processing," in *Proc. 8th Brit. Mach. Vision Conf.*, 1997, pp. 320–329.

[32] H. F. Chen, P. N. Belhumeur, and D. W. Jacobs, "In search of illumination invariants," in *Proc. IEEE Comput. Vision Pattern Recogn. Conf.*, 2000, pp. 254–261.

[33] S. Ravela and C. Luo, "Appearance-based global similarity retrieval of images," *Adv. Informat. Retri.*, vol. 7, pp. 267–299, 2000.

[34] M. Osadchy, D. W. Jacobs, and M. Lindenbaum, "Surface dependent representations for illumination insensitive image comparison," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 98–111, Jan. 2007.

[35] T. Zhang, Y. Y. Tang, B. Fang, Z. Shang, and X. Liu, "Face recognition under varying illumination using gradientfaces," *IEEE Trans. Image Process.*, vol. 18, no. 11, pp. 2599–2606, Nov. 2009.

[36] K. C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 684–698, May 2005.

[37] A. M. Martinez and R. Benavente, "The AR face database," CVC Tech. Rep., 1998.

[38] K. Messer, J. Matas, J. V. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *Proc. AVBPA*, Mar. 1999, pp. 72–77.

[39] L. Juwei, K. N. Plataniotis, A. N. Venetsanopoulos, and J. Wang, "An efficient kernel discriminant analysis method," *Pattern Recognition*, vol. 38, no. 10, pp. 1788–1790, 2005.

[40] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[41] J. Matas, M. Hamou, K. Jonsson, J. Kittler, Y. Li, C. Kotropoulos, A. Tefas, I. Pitas, T. Tan, H. Yan, F. Smeraldi, J. Bigun, N. Capdevielle, W. Gerstner, S. Ben-Yacouba, Y. Abdelaoued, and E. Mayoraz, "Comparison of face verification results on the XM2VTS database," in *Proc. Pattern Recogn. 15th Int. Conf.*, Barcelona, Spain, Sep. 2000, pp. 858–863.

[42] K. Messer, J. V. Kittler, M. Sadeghi, S. Marcel, C. Marcel, S. Bengio, F. Cardinaux, C. Sanderson, J. Czyz, L. Vandendorpe, S. Srisuk, M. Petrou, W. Kurutach, A. Kadyrov, R. Paredes, B. Kepenekci, F. B. Tek, G. B. Akar, F. Deravi, and N. Mavity, "Face verification competition on the XM2VTS database," in *Proc. AVBPA*, Jun. 2003, pp. 964–974.

[43] K. Messer, J. Kittler, J. Short, G. Heusch, F. Cardinaux, S. Marcel, Y. Rodriguez, S. Shan, Y. Su, and W. Gao, "Performance characterisation of face recognition algorithms and their sensitivity to severe illumination changes," *Lecture Notes in Computer Science*, vol. 3832, pp. 1–11, 2006.

[44] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

[45] B. A. Draper, K. Baek, M. S. Bartlett, and J. R. Beveridge, "Recognizing faces with PCA and ICA," *Comput. Vision Image Underst.*, vol. 91, nos. 1–2, pp. 115–137, 2003.

[46] S. Zafeiriou, A. Tefas, and I. Pitas, "Minimum class variance support vector machines," *IEEE Trans. Image Process.*, vol. 16, no. 10, pp. 2551–2564, Oct. 2007.