Class-Specific Kernel-Discriminant Analysis for Face Verification

Georgios Goudelis, Stefanos Zafeiriou, Anastasios Tefas, *Member, IEEE*, and Ioannis Pitas, *Fellow, IEEE*

***Abstract*—In this paper, novel nonlinear subspace methods for face verification are proposed. The problem of face verification is considered as a two-class problem (genuine versus impostor class). The typical Fisher's linear discriminant analysis (FLDA) gives only one or two projections in a two-class problem. This is a very strict limitation to the search of discriminant dimensions. As for the FLDA for N class problems (N is greater than two), the transformation is not person specific. In order to remedy these limitations of FLDA, exploit the individuality of human faces and take into consideration the fact that the distribution of facial images, under different viewpoints, illumination variations, and facial expression is highly complex and nonlinear, novel kernel-discriminant algorithms are proposed. The new methods are tested in the face verification problem using the XM2VTS, AR, ORL, Yale, and UMIST databases where it is verified that they outperform other commonly used kernel approaches such as kernel–PCA (KPCA), kernel direct discriminant analysis (KDDA), complete kernel Fisher's discriminant analysis (CKFDA), the two-class KDDA, CKFDA, and other two-class and multiclass variants of kernel-discriminant analysis based on Fisher's criterion.**

***Index Terms*—Face verification, Fisher's linear discriminant analysis (FLDA), kernel techniques, two-class problems.**

I. Introduction

FACE recognition/verification has attracted the attention of the research community for more than two decades and is among the most popular research areas in the field of computer vision and pattern recognition. The two problems of face verification and recognition are conceptually different. On one hand, a recognition system assists a human expert in determining the identity of a test face. On the other hand, person verification systems should decide whether an identity claim is valid or invalid.

The most popular among the techniques used for face recognition/verification are the so-called subspace methods. The subspace algorithms represent the facial image by a feature vector and their aim is to find projections (bases) that optimize some criterion defined over the feature vectors that correspond to different classes. Then, the original high-dimensional image space is projected into a low-dimensional one. The classification is usually performed according to a simple similarity measure in the final multidimensional space.

Various criteria have been employed in order to find the bases of the low-dimensional spaces. Some of them have been defined in order to find projections that best express the population without using the information about the way the data are separated to different classes (e.g., principal component analysis (PCA) [1] and non-negative matrix factorization [2]). Another class of criteria is the one that deals directly with the discrimination between classes (e.g., Fisher's linear discriminant analysis (FLDA) [3]–[5]). Finally, statistical independence in the low-dimensional space can be also used as a criterion in order to find the linear projections (e.g., independent component analysis (ICA) [6], [7].

Face verification and recognition are usually treated differently when discriminant subspace methods are used for feature selection. That is, face verification is treated as a two-class problem while the face recognition is an N class problem (N is the number of different facial classes). This yields a different projection pursuit strategy. On one hand, the strategy for face verification is to find class-specific projections that separate the genuine class from the impostor class optimally (under some criterion). The class-specific-discriminant projection is intuitively motivated by the fact that each face is unique and it should have its own discriminant parts. On the other hand in face recognition systems, the discriminant projections are found by trying to optimally separate all of the genuine classes. This strategy gives a set of discriminant projections that is common for all of the facial classes. For face recognition, the interested reader may refer to [8]–[11] where face recognition is treated as an N class problem.

In this paper, face verification is modelled as a two-class problem. The motivations of such modeling are supported by various methods that take into account the individuality of facial features [12], [13]–[19]. In [13] and [14], two-class problems (genuine versus impostor claims) have been formed for discriminant feature selection in the nodes of elastic graphs. In [12] and [15], two-class problems have been formulated in order to find the class-specific discriminant weights for the facial landmarks that correspond to nodes of elastic graphs and use this information when forming a similarity measure between faces. Recently, it has been shown that the verification performance can be highly improved by using class-specific discriminant functions in every step of elastic graph matching [16]. Moreover, the use of person-specific graphs with nodes placed at discriminant facial landmarks greatly improves the performance of elastic graph matching in frontal face verification [17]. In [18], it has been shown that discriminant non-negative matrix factorization methods with class-specific bases perform better than other approaches with common bases. Additional details about modeling face verification as a two-class problem are given in [19] introducing the class-specific Fisherfaces. The motivations of

Manuscript received November 6, 2006; revised May 31, 2007. This work was supported by the network of excellence BioSecure IST-2002-507634 (Biometrics for Secure Authentication, http://www.biosecure.info) under the Information Society Technologies (IST) priority of the 6th Framework Programme of the European Community. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Rama Chellappa.

G. Goudelis, S. Zafeiriou, A. Tefas, and I. Pitas are with the Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece (e-mail: pitas@aiia.csd.auth.gr; pitas@zeus.csd.auth.gr).

A. Tefas is also with the Department of Information Management, Technological Educational Institute of Kavala, Kavala 65404, Greece.

Digital Object Identifier 10.1109/TIFS.2007.902915

using class-specific transforms are supported by other works as well, which employ class-specific fusion rules for person verification (such as voice, fingerprint, and hand features [20]).

The methods proposed in this paper exploit the individuality of the human face in order to find a nonlinear subspace representation with enhanced discriminant power. In detail in this paper, we propose a novel class-specific discriminant criterion which, when optimized, leads to a discriminant low-dimensional representation of faces. Furthermore, in order to represent the face better in various poses, we combine the proposed criterion with kernel techniques and we present two techniques for optimizing the criterion in arbitrary dimensional Hilbert spaces leading to a novel class-specific kernel-discriminant analysis (CSKDA). However, the main contribution of the proposed CSKDA is that it tries to remedy some of the limitations of the kernel methods based on the Fisher's discriminant criterion that provide a very limited number of features in two-class problems (i.e., the so-called kernel direct discriminant analysis (KDDA) provides only one discriminant projection [9] and the so-called complete kernel Fisher discriminant analysis (CKFDA) [11] has only two discriminant dimensions in two-class problems). These spaces of a very limited number of dimensions may prove to be insufficient for correctly representing facial images. The proposed approach discovers a low-dimensional space with the number of dimensions to be proportional to the number of images available for training. Experiments conducted in the XM2VTS [21], AR [22], [23], the ORL [24], Yale [25], and the UMIST [26] databases using facial images at various poses demonstrate the potential of the proposed methods.

The rest of the paper is organized as follows. The problem of face verification and how kernel subspace methods that can be applied to this problem is discussed in Section II. In Section III, the new criterion is described. In Sections III-B and C, two algorithms for solving the optimization problem and finding the discriminant subspace transform are proposed. A comparison of the proposed method to other commonly used kernel approaches in terms of the number of extracted features and computational complexity is given in Section IV. Experimental results with artificial data and comments on face verification are presented in Section V. Face verification experiments with various databases are shown in Section VI. Finally, conclusions are drawn in Section VII.

II. Face Verification and Kernel Subspace Techniques

In this section, we will briefly outline the problem of face verification and the framework under which a kernel subspace method can be used in order to solve this problem.

The facial image representation, which is the facial image or augmented facial representations (i.e., Gabor features [5]), is scanned row-wise to form a facial vector $\mathbf{z} \in \Re^M$. Let $\mathcal{U}$ be a facial vector database (training set) that contains a total of L facial vectors. Every facial vector $\mathbf{z}$ is supposed to belong to one of the N facial (person) classes $\{\mathcal{U}_1,\mathcal{U}_2,\ldots,\mathcal{U}_N\}$ with $\mathcal{U} = \bigcup_{i=1}^{N}\mathcal{U}_i$, considered as the clients of the verification system. For a face verification system that uses $\mathcal{U}$, a genuine (or client) claim is performed when a person t provides its facial vector $\mathbf{u}$, claiming that $\mathbf{u} \in \mathcal{U}_r$ and $t = r$. When a person t provides its facial vector $\mathbf{u}$ and claims that $\mathbf{u} \in \mathcal{U}_r$, with $t \neq r$, an impostor claim occurs. The scope of a face verification system is to properly handle these claims by accepting the genuine claims and rejecting the impostor ones.

In order to make use of kernel techniques, the original input space is projected to an arbitrary-dimensional space $\mathcal{F}$ (the space $\mathcal{F}$ usually has the structure of a Hilbert space [27], [28]). To do so, let $\phi : \mathbf{z} \in \Re^M \longrightarrow \phi(\mathbf{z}) \in \mathcal{F}$ be a nonlinear mapping from the input space $\Re^M$ to the Hilbert space $\mathcal{F}$. In the Hilbert space, we want to find linear projections to a low-dimensional space with enhanced discriminant power. The discriminant power of the new space is often defined in respect to a discriminant optimization criterion. This discriminant criterion defines an optimization problem which gives a set of linear projections in $\mathcal{F}$ (linear in $\mathcal{F}$ is nonlinear in $\Re^M$).

A linear subspace transformation of $\mathcal{F}$ onto a K-dimensional subspace, which is isomorphic to $\Re^K$, is a matrix $\mathbf{X} = [\mathbf{x}_1,\ldots,\mathbf{x}_K]$ with $\mathbf{x}_i \in \mathcal{F}$. The new vector $\acute{\mathbf{z}} \in \Re^M$ of the facial vector $\mathbf{z}$ is given by

$$\acute{\mathbf{z}} = \mathbf{X}^T\phi(\mathbf{z}) = [\mathbf{x}_1^T\phi(\mathbf{z}),\ldots,\mathbf{x}_K^T\phi(\mathbf{z})]^T. \tag{1}$$

The dimensionality of the new space is usually much smaller than the dimensionality of $\mathcal{F}$ and the dimensionality of the input space $\Re^M$ (i.e., $K \ll M$). The matrix multiplication in (1) is computed indirectly using dot-products in the Hilbert space $\mathcal{F}$ [9], [11], [29] (the so-called kernel trick, see the next section for details). The bases matrix $\mathbf{X}$ can be the same for all facial classes of the database or can be different for each facial class. In the case of class-specific image bases, for the reference person r, the set $\mathcal{I}_r = \mathcal{U} - \mathcal{U}_r$ that corresponds to impostor images is used in order to construct the two-class problem (genuine versus impostor class) and obtain the matrix $\mathbf{X}$ [12], [30].

After the projection, given by (1), a similarity measure is chosen in order to quantify the similarity of a test facial vector to a certain class. This similarity measure can be the L_1 norm, the L_2 norm, the normalized correlation, or the Mahalanobis distance [31].

III. Discriminant Criterion

Before we develop the new optimization problem, we will introduce some notation that is used throughout this paper. Let r be the reference person that will be used for defining the person-specific algorithms. Let L_G and L_I be the numbers of genuine and impostor images in the training set for the person r, respectively. Usually, the number of genuine images is much smaller than the number of impostor images for a reference person r. Thus, in the following analysis, we will work under the assumption that $L_I > L_G$. Let $L = L_G + L_I$ be the total number of images in the training database. The genuine vectors $\mathbf{z}_i$ of the person r will be denoted as $\boldsymbol{\rho}_i = \mathbf{z}_i(\mathbf{z}_i \in \mathcal{U}_r)$ while the impostor images $\mathbf{z}_i$ of the person r will be denoted as $\boldsymbol{\kappa}_i = \mathbf{z}_i(\mathbf{z}_i \in \mathcal{I}_r)$. Also let $\bar{\boldsymbol{\rho}} = (1/L_G)\sum_{i=1}^{L_G}\phi(\boldsymbol{\rho}_i)$, $\bar{\boldsymbol{\kappa}} = (1)/(L_I)\sum_{i=1}^{L_I}\phi(\boldsymbol{\kappa}_i)$ and $\bar{\mathbf{m}} = (1/L)\sum_{i=1}^{L}\phi(\mathbf{z}_i)$ be the mean vectors of the genuine class, the impostor class, and the total mean of the facial vectors in the Hilbert space $\mathcal{F}$. Any function k satisfying the Mercer's condition can be used as a kernel. The dot product of $\phi(\mathbf{z}_i)$ and $\phi(\mathbf{z}_j)$ in the Hilbert space can be calculated without having to explicitly evaluate the mapping $\phi(\cdot)$ as $k(\mathbf{z}_i,\mathbf{z}_j) = \phi(\mathbf{z}_i)^T\phi(\mathbf{z}_j)$ (this is also known as the kernel trick [27], [28]). The typical kernels that have been used in our experiments have been polynomial and radial basis functions (RBF) kernels

$$\begin{aligned} k(\mathbf{x},\mathbf{y}) &= \phi(\mathbf{x})^T\phi(\mathbf{y}) = (\mathbf{x}^T\mathbf{y}+1)^d \\ k(\mathbf{x},\mathbf{y}) &= \phi(\mathbf{x})^T\phi(\mathbf{y}) = e^{-\gamma(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})} \end{aligned} \tag{2}$$

where d is the degree of the polynomial and γ is the spread of the Gaussian cluster. Kernels that do not satisfy the Mercer's condition [27] have also been successfully applied for face recognition [31].

In the experiments described in this paper, we have used fractional power polynomial models as well (i.e., polynomial functions, such as the ones defined in (2)), with $0 < d < 1$. A fractional power polynomial, however, does not necessarily define a kernel function, as it might not define a positive semidefinite Gram matrix [31]. Note that the sigmoid kernels, which are one of the three classes of widely used kernel functions (polynomial kernels, Gaussian kernels, and sigmoid kernels), do not actually define a positive semidefinite Gram matrix either. Nevertheless, the sigmoid kernels have been successfully used in practice, such as in building support vector machines (SVMs) [32]. In the case such models are adopted in the presented methods, additional comments are inserted throughout this paper in order to treat these cases.

A. Class-Specific Criterion

The criterion that is used in this paper in order to build the proposed feature extraction method is the generalization of the criterion used in [13] for discriminant graph node weighting and in [30] and [17] for discriminant feature extraction in graph nodes. Here, we will generalize this criterion for nonlinear feature extraction, with the help of a simple similarity measure in the Hilbert space $\mathcal{F}$. This measure quantifies the similarity of a given feature vector $\mathbf{z}$ to the reference facial class r in the subspace spanned by the columns of the matrix $\boldsymbol{\Psi} = [\boldsymbol{\psi}_1 \dots \boldsymbol{\psi}_K]$, with $\boldsymbol{\psi}_i \in \mathcal{F}$. The L_2 norm in the reduced space spanned by the columns of $\boldsymbol{\Psi}$ is used as a similarity measure

$$\begin{aligned} d_r(\mathbf{z}) &= ||\boldsymbol{\Psi}^T(\phi(\mathbf{z}) - \bar{\boldsymbol{\rho}})||^2 \\ &= \operatorname{tr}[\boldsymbol{\Psi}^T(\phi(\mathbf{z}) - \bar{\boldsymbol{\rho}})(\phi(\mathbf{z}) - \bar{\boldsymbol{\rho}})^T\boldsymbol{\Psi}] \\ &= \sum_{i=1}^{K} \operatorname{tr}[\boldsymbol{\psi}_i^T(\phi(\mathbf{z}) - \bar{\boldsymbol{\rho}})(\phi(\mathbf{z}) - \bar{\boldsymbol{\rho}})^T\boldsymbol{\psi}_i] \end{aligned} \tag{3}$$

which is actually the Euclidean distance of a projected sample to the projected mean of the reference class and is one of most usually employed measures in pattern recognition applications (i.e., the distance from the center of the class). This distance should be low for the samples of the genuine class and should be high for the samples of the impostor class.

Now, in order to find a discriminant linear transformation in $\mathcal{F}$, we demand that the sum of the similarity measures $d_r(\mathbf{z})$ for all $\mathbf{z} \in \mathcal{I}_r$ (impostor similarity measures) are to be maximized while minimizing the sum of the similarity measures $d_r(\mathbf{z})$ for all $\mathbf{z} \in \mathcal{U}_r$ (client similarity measures). Thus, the discriminant projections $\boldsymbol{\psi}_i \in \mathcal{F}$ are found in the training set as the ones that maximize the ratio

$$\begin{aligned} D^{\Phi}(\boldsymbol{\Psi}) &= \frac{\sum_{\mathbf{z}\in\mathcal{I}_r} d_r(\mathbf{z})}{\sum_{\mathbf{z}\in\mathcal{U}_r} d_r(\mathbf{z})} \\ &= \frac{\sum_{\mathbf{z}\in\mathcal{I}_r}\sum_{i=1}^{K} \operatorname{tr}[\boldsymbol{\psi}_i^T(\phi(\mathbf{z}) - \bar{\boldsymbol{\rho}})(\phi(\mathbf{z}) - \bar{\boldsymbol{\rho}})^T\boldsymbol{\psi}_i]}{\sum_{\mathbf{z}\in\mathcal{U}_r}\sum_{i=1}^{K} \operatorname{tr}[\boldsymbol{\psi}_i^T(\phi(\mathbf{z}) - \bar{\boldsymbol{\rho}})(\phi(\mathbf{z}) - \bar{\boldsymbol{\rho}})^T\boldsymbol{\psi}_i]} \\ &= \frac{\sum_{i=1}^{K} \operatorname{tr}[\boldsymbol{\psi}_i^T[\sum_{\mathbf{z}\in\mathcal{I}_r}(\phi(\mathbf{z}) - \bar{\boldsymbol{\rho}})(\phi(\mathbf{z}) - \bar{\boldsymbol{\rho}})^T]\boldsymbol{\psi}_i]}{\sum_{i=1}^{K} \operatorname{tr}[\boldsymbol{\psi}_i^T[\sum_{\mathbf{z}\in\mathcal{U}_r}(\phi(\mathbf{z}) - \bar{\boldsymbol{\rho}})(\phi(\mathbf{z}) - \bar{\boldsymbol{\rho}})^T]\boldsymbol{\psi}_i]} \\ &= \frac{\sum_{i=1}^{K} \operatorname{tr}[\boldsymbol{\psi}_i^T\mathbf{W}^{\Phi}\boldsymbol{\psi}_i]}{\sum_{i=1}^{K} \operatorname{tr}[\boldsymbol{\psi}_i^T\mathbf{B}^{\Phi}\boldsymbol{\psi}_i]} \\ &= \frac{\operatorname{tr}[\boldsymbol{\Psi}^T\mathbf{W}^{\Phi}\boldsymbol{\Psi}]}{\operatorname{tr}[\boldsymbol{\Psi}^T\mathbf{B}^{\Phi}\boldsymbol{\Psi}]} \end{aligned} \tag{4}$$

where $\mathbf{W}^{\Phi} = \sum_{\mathbf{z}\in\mathcal{I}_r}(\phi(\mathbf{z}) - \bar{\boldsymbol{\rho}})(\phi(\mathbf{z}) - \bar{\boldsymbol{\rho}})^T$, $\mathbf{B}^{\Phi} = \sum_{\mathbf{z}\in\mathcal{U}_r}(\phi(\mathbf{z}) - \bar{\boldsymbol{\rho}})(\phi(\mathbf{z}) - \bar{\boldsymbol{\rho}})^T$ and $\operatorname{tr}[\mathbf{M}]$ is the trace of matrix $\mathbf{M}$. The direct optimization of $D^{\Phi}(\boldsymbol{\Psi})$ in $\mathcal{F}$ is an intractable problem due to the fact that both $\mathbf{W}^{\Phi}$ and $\mathbf{B}^{\Phi}$ are matrices with arbitrary dimensions. Before presenting the proposed methods, we will briefly outline research that has been conducted concerning the criterion in (4). A form of the above criterion that produces linear-discriminant transforms has been used in [13] and [33] for discriminant feature extraction in the nodes of elastic graphs. A similar approach has been followed in [14], [30], [34] for the same purpose. The solution of the criterion has been further analyzed in [17], derived in a two-step optimization procedure. Moreover, another form of the above criterion has been considered in [18]. That is, in [18], discriminant costs based on the discriminant criterion (4) have been added in the NMF cost leading to the so-called class-specific discriminant NMF (CSDNMF) algorithm [18], [35]. Independent studies for the benefits of the above criterion have been started by the image retrieval community, leading to the so-called biased discriminant analysis [36] (BDA). Nevertheless, we should note that research concerning BDA has been initiated by the biometric community [13], [34], [30], [33], [14].

A kernelized version of the criterion has been proposed in [36] and it has been solved using a regularization strategy (i.e., adding a scaled version of the identity matrix to the kernel matrices). In the following, two different theoretical sound and numerical stable methods are presented for solving the optimization problem and are similar to the ones used in [9], [11], and [29] for optimizing the PCA or the LDA criterion with kernels.

- In Section III-B, the direct optimization approach will be presented. The direct optimization approach exploits the discriminant information hidden in the null-space of the matrix $\mathbf{B}^{\Phi}$ (i.e., the eigenvectors that correspond to null eigenvalues of $\mathbf{B}^{\Phi}$) and has been motivated by the direct discriminant analysis algorithms that have been proposed in [8], [37], and [38]. The main idea behind the direct optimization algorithm is that the null space of the matrix $\mathbf{B}^{\Phi}$ may contain significant discriminant information in the case where the projection of the matrix $\mathbf{W}^{\Phi}$ is nonzero in that direction. No significant information will be lost if the null space of the matrix $\mathbf{W}^{\Phi}$ is discarded (the geometrical interpretation of this statement is shown in the Appendix).

- In Section III-C, an alternative way for finding projections that maximize the ratio (4) will be presented. This method will be called two-step dimensionality reduction due to the fact that it uses two-dimensionality reduction steps prior to optimization and is inspired from [11]. The main idea of this optimization approach is to find nonlinear mapping of the data that map the arbitrary dimensional Hilbert space to a finite dimensional subspace without losing any information with respect to the optimization problem. Afterwards, the problem is redefined in the finite dimensional subspace and solved there, using typical linear techniques. Finally, two different criteria are defined in finite dimensional subspace in order to take into consideration both the null and the non-null spaces of the matrices, producing two types of discriminant information: the regular and the irregular discriminant information. The two criteria that give the discriminant features will be defined below in their general form.

B. Direct Optimization of the Discriminant Criterion

1) Eigenanalysis of $\mathbf{W}^{\Phi}$ in the Hilbert Space $\mathcal{F}$: In order to solve the optimization problem (4), we start by solving the eigenvalue problem of $\mathbf{W}^{\Phi}$, which can be rewritten here as follows:

$$\begin{aligned}\mathbf{W}^{\Phi} &= \sum_{\mathbf{z}\in\mathcal{I}_k}(\phi(\mathbf{z})-\bar{\boldsymbol{\rho}})(\phi(\mathbf{z})-\bar{\boldsymbol{\rho}})^T\\ &= \sum_{i=1}^{L_I}(\phi(\boldsymbol{\kappa}_i)-\bar{\boldsymbol{\rho}})(\phi(\boldsymbol{\kappa}_i)-\bar{\boldsymbol{\rho}})^T\\ &= \sum_{i=1}^{L_I}\tilde{\boldsymbol{\kappa}}_i\tilde{\boldsymbol{\kappa}}_i^T = \boldsymbol{\Phi}_w\boldsymbol{\Phi}_w^T \end{aligned} \tag{5}$$

where L_I is the number of impostor facial vectors in the training set $\tilde{\boldsymbol{\kappa}}_i = \phi(\boldsymbol{\kappa}_i) - \bar{\boldsymbol{\rho}}$ and $\boldsymbol{\Phi}_w = [\tilde{\boldsymbol{\kappa}}_1 \ldots \tilde{\boldsymbol{\kappa}}_{L_I}]$. The first m (with $m \leq L_I - 1$) most significant eigenvectors of $\mathbf{W}^{\Phi}$, which correspond to its nonzero eigenvalues, can be indirectly derived from the eigenvectors of the matrix $\boldsymbol{\Phi}_w^T\boldsymbol{\Phi}_w$ $(L_I \times L_I)$. The computation of $\boldsymbol{\Phi}_w^T\boldsymbol{\Phi}_w$ can be performed by only using dot products in the Hilbert space $\mathcal{F}$

Using the kernel function, for the genuine and impostor class the four dot product matrices can be defined as:

$$\begin{aligned}[\mathbf{K}_1]_{i,j} &= \phi(\boldsymbol{\rho}_i)^T\phi(\boldsymbol{\rho}_j) = k(\boldsymbol{\rho}_i,\boldsymbol{\rho}_j)\\ &\qquad i = 1\ldots L_G \quad \text{and} \quad j = 1\ldots L_G\\ [\mathbf{K}_2]_{i,j} &= \phi(\boldsymbol{\kappa}_i)^T\phi(\boldsymbol{\rho}_j) = k(\boldsymbol{\kappa}_i,\boldsymbol{\rho}_j)\\ &\qquad i = 1\ldots L_I \quad \text{and} \quad j = 1\ldots L_G\\ [\mathbf{K}_3]_{i,j} &= \phi(\boldsymbol{\rho}_i)^T\phi(\boldsymbol{\kappa}_j) = k(\boldsymbol{\rho}_i,\boldsymbol{\kappa}_j) = \mathbf{K}_2^T\\ [\mathbf{K}_4]_{i,j} &= \phi(\boldsymbol{\kappa}_i)^T\phi(\boldsymbol{\kappa}_j) = k(\boldsymbol{\kappa}_i,\boldsymbol{\kappa}_j)\\ &\qquad i = 1\ldots L_I \quad \text{and} \quad j = 1\ldots L_I. \end{aligned} \tag{6}$$

Using the previously defined matrices, $\boldsymbol{\Phi}_w^T\boldsymbol{\Phi}_w$ can be expressed as

$$\begin{aligned}\boldsymbol{\Phi}_w^T\boldsymbol{\Phi}_w &= \mathbf{K}_4 - \frac{1}{L_G}\mathbf{K}_2\mathbf{1}_{L_GL_I}\\ &\quad - \frac{1}{L_G}\mathbf{1}_{L_IL_G}\mathbf{K}_3 + \frac{1}{N_G^2}\mathbf{1}_{L_IL_G}\mathbf{K}_1\mathbf{1}_{L_GL_I} \end{aligned} \tag{7}$$

where $\mathbf{1}_{L_GL_I}$ is a $L_G \times L_I$ matrix with terms all equal to one. The detailed derivation of (7) can be found in the Appendix.

Let λ_i^w and $\mathbf{r}_i(i = 1\ldots L_I)$ be the ith eigenvalue and the corresponding eigenvector of $\boldsymbol{\Phi}_w^T\boldsymbol{\Phi}_w$, sorted in ascending order of eigenvalues. It is true that $(\boldsymbol{\Phi}_w\boldsymbol{\Phi}_w^T)(\boldsymbol{\Phi}_w\mathbf{r}_i) = \lambda_i^w(\boldsymbol{\Phi}_w\mathbf{r}_i)$. Thus, $\mathbf{v}_i = \boldsymbol{\Phi}_w\mathbf{r}_i$ are the eigenvectors of $\mathbf{W}^{\Phi}$. In order to remove the null space of $\mathbf{W}^{\Phi}$, the first $m \leq L_I - 1$ eigenvectors $\mathbf{V} = [\mathbf{v}_1 \ldots \mathbf{v}_m] = \boldsymbol{\Phi}_w\mathbf{R}$, where $\mathbf{R} = [\mathbf{r}_1 \ldots \mathbf{r}_m]$, whose corresponding eigenvalues are greater than 0 should be calculated. Thus $\mathbf{V}^T\mathbf{W}^{\Phi}\mathbf{V} = \boldsymbol{\Lambda}_w$, with $\boldsymbol{\Lambda}_w = \operatorname{diag}[\lambda_1^{w2} \ldots \lambda_m^{w\,2}]$ is a $m \times m$ diagonal matrix. In the case that strictly positive kernels are employed, such as the polynomial kernel with $d \geq 1$ and $d \in \mathcal{Z}$ and RBF kernels, the matrix $\mathbf{W}^{\Phi}$ is positive semidefinite. In case that fractional power polynomial models are adopted (i.e., $0 < d < 1$), it is possible that negative eigenvalues may occur. In this case, there are two different alternatives:

- to remove the eigenvectors that correspond to negative eigenvalues. This step is preferred when the negative eigenvalues are few and their magnitude is very small compared to the magnitude of the positive eigenvalues. This method has been successfully used for face recognition when using KPCA with fractional polynomial models [31].
- to use only the magnitude of the negative eigenvalues. This step is preferred when the magnitude of the negative eigenvalues is not small, or when there are a lot of dimensions that correspond to negative eigenvalues in the embedding [39], [40].

We have used both alternatives approaches when using the fractional polynomial models and both have lead to approximately similar verification results.

2) Eigenanalysis of $\boldsymbol{\Phi}_w^T\mathbf{B}^{\Phi}\boldsymbol{\Phi}_w$ in the Hilbert Space: Let $\mathbf{U} = \mathbf{V}\boldsymbol{\Lambda}_w^{-1/2}$. Using the matrix $\mathbf{U}$, it can be easily seen that $\mathbf{U}^T\mathbf{W}^{\Phi}\mathbf{U} = \mathbf{I}$ while $\mathbf{U}^T\mathbf{B}^{\Phi}\mathbf{U}$ can be expanded as

$$\begin{aligned}\mathbf{U}^T\mathbf{B}^{\Phi}\mathbf{U} &= (\boldsymbol{\Phi}_w\mathbf{R}\boldsymbol{\Lambda}_w^{-1/2})^T\mathbf{B}^{\Phi}(\boldsymbol{\Phi}_w\mathbf{R}\boldsymbol{\Lambda}_w^{-1/2})\\ &= (\mathbf{R}\boldsymbol{\Lambda}_w^{-1/2})^T(\boldsymbol{\Phi}_w^T\mathbf{B}^{\Phi}\boldsymbol{\Phi}_w)(\mathbf{R}\boldsymbol{\Lambda}_w^{-1/2}). \end{aligned} \tag{8}$$

Using the kernel matrices $\mathbf{K}_1, \mathbf{K}_3, \mathbf{K}_2$ and $\mathbf{K}_4$, a closed-form expression of $\boldsymbol{\Phi}_w^T\mathbf{B}\boldsymbol{\Phi}_w$ can be formed as

$$\boldsymbol{\Phi}_w^T\mathbf{B}^{\Phi}\boldsymbol{\Phi}_w = \mathbf{A}_1 - L_G\mathbf{A}_2 \tag{9}$$

where $\mathbf{A}_1$ and $\mathbf{A}_2$ are defined in Appendix B along with the detailed derivation of the expression (9).

The matrix $\mathbf{U}^T\mathbf{B}^{\Phi}\mathbf{U}$ has size $m \times m$. Thus, the eigenanalysis of $\mathbf{U}^T\mathbf{B}^{\Phi}\mathbf{U}$ is computationally feasible. Let $\mathbf{p}_i$ be the ith-ordered eigenvector of matrix $\mathbf{U}^T\mathbf{B}^{\Phi}\mathbf{U}$ with size $m \times m$, with $i = 1\ldots m$ sorted in descending order of the corresponding eigenvalue λ_i^b. In the set of the ordered eigenvectors, those that correspond to the smaller eigenvalues maximize the discriminant ratio in (4). Discarding the vectors with the highest eigenvalues, the $l \leq m$ remaining eigenvectors are represented in the form of the matrix $\mathbf{P} = [\mathbf{p}_1 \ldots \mathbf{p}_l]$. Defining a matrix $\mathbf{Q} = \mathbf{U}\mathbf{P}$, we can obtain $\mathbf{Q}^T\mathbf{B}^{\Phi}\mathbf{Q} = \boldsymbol{\Lambda}_b$, with $\boldsymbol{\Lambda}_b = \operatorname{diag}[\lambda_i^b \ldots \lambda_l^b]$, a $l \times l$ diagonal matrix. When adopting fractional power polynomial models, a similar approach to Section III-B1 should be followed in order to take care of the negative eigenvalues that may occur during the eigenanalysis of $\boldsymbol{\Phi}_w^T\mathbf{B}^{\Phi}\boldsymbol{\Phi}_w$.

Based on the previously presented analysis, a set of optimal discriminant features can be derived from $\mathbf{\Gamma} = \mathbf{Q}\mathbf{\Lambda}_b^{-1/2}$. The features form a low-dimensional subspace in $\mathcal{F}$, where the discriminant ratio in (4) is maximized. However, it is highly possible that eigenvalues exist with $\lambda_i^b = 0$ in $\mathbf{\Lambda}_b$. This may occur in many real-world applications where the impostor claims are more than the genuine claims in the training set. In order to solidify the procedure and prevent the existence of zero eigenvalues in the final transform, a regularized alternative criterion is used as [41] and [42]

$$D(\mathbf{\Psi}) = \frac{\mathrm{tr}[\mathbf{\Psi}^T\mathbf{W}^{\Phi}\mathbf{\Psi}]}{\mathrm{tr}[(1-\vartheta)\mathbf{\Psi}^T\mathbf{B}^{\Phi}\mathbf{\Psi} + \vartheta\mathbf{\Psi}^T\mathbf{W}^{\Phi}\mathbf{\Psi}]} \tag{10}$$

where $0 < \vartheta < 1$ is a regularization parameter. The alternative criterion (10) can be easily proven to be equivalent to (4) by using an analysis similar to the one used for the conventional LDA criterion in [8], [9], [41], and [42]. Thus, the matrix $\tilde{\mathbf{\Lambda}}_b = \mathbf{Q}^T(\vartheta\mathbf{W}^{\Phi} + (1-\vartheta)\mathbf{B}^{\Phi})\mathbf{Q} = \vartheta\mathbf{I} + (1-\vartheta)\mathbf{\Lambda}_b$ is nonsingular.

3) Feature Extraction: The matrix $\tilde{\mathbf{\Gamma}}^T = \mathbf{Q}\tilde{\mathbf{\Lambda}}_b^{-1/2}$ is used for discriminant dimensionality reduction in the Hilbert space $\mathcal{F}$. Let a test facial image be scanned row-wise to form a facial vector $\mathbf{y}$ in order to extract the low-dimensional vector $\acute{\mathbf{y}}$ from the facial vector $\mathbf{y}$ using the proposed method. This procedure is detailed in Appendix D. The number of dimensions of the vector $\acute{\mathbf{y}}$ is $l \leq L_I - 1$. The measure that quantifies the similarity of a facial vector to the reference facial class r is given by

$$d_r(\mathbf{y}) = \|\tilde{\mathbf{\Gamma}}^T(\phi(\mathbf{y}) - \bar{\boldsymbol{\rho}})\|^2. \tag{11}$$

In order to accept the claim (i.e., verify that the facial vector $\mathbf{y}$ belongs to the reference person r), the measure $d_r(\mathbf{y})$ should be compared to a threshold T_r (i.e., $d_r(\mathbf{y}) < T_r$).

C. Two-Step Optimization Method for the Discriminant Criterion

In this section, an alternative way for finding projections that maximize the ratio (4) will be presented.

In the Hilbert space $\mathcal{F}$, it is almost impossible to make $\mathbf{B}^{\Phi}$ invertible (the matrix $\mathbf{B}^{\Phi}$ is invertible if the dimension of the feature vectors is smaller than the number of client images). Thus, vectors $\boldsymbol{\psi}_i$ such that $\boldsymbol{\psi}_i^T\mathbf{B}^{\Phi}\boldsymbol{\psi}_i = 0$ always exist. These vectors are very effective for discrimination if they satisfy $\boldsymbol{\psi}_i^T\mathbf{W}^{\Phi}\boldsymbol{\psi}_i > 0$ at the same time since, for these vectors, it is valid that $D^{\Phi}(\mathbf{\Psi}) \rightarrow +\infty$. A geometrical interpretation of the effect of the vectors $\boldsymbol{\psi}_i$ that satisfy $\boldsymbol{\psi}_i^T\mathbf{B}^{\Phi}\boldsymbol{\psi}_i = 0$ and $\boldsymbol{\psi}_i^T\mathbf{W}^{\Phi}\boldsymbol{\psi}_i = 0$ on the training genuine and impostor vectors is given in Appendix A. In such a case, the criterion (4) degenerates into the following:

$$D_b^{\Phi}(\mathbf{\Psi}) = \mathrm{tr}[\mathbf{\Psi}^T\mathbf{W}^{\Phi}\mathbf{\Psi}] \times (\mathbf{\Psi} = [\ldots\boldsymbol{\psi}_i\ldots], \|\boldsymbol{\psi}_i\| = 1). \tag{12}$$

Using the criteria D_b^{Φ} and D^{Φ}, two kind of discriminant features can be calculated. We will call the discriminant projections of the criterion D^{Φ} as regular while the ones of the criterion D_b^{Φ} will be called irregular.

1) Reducing $\mathcal{F}$: The first step is to reduce the Hilbert space $\mathcal{F}$ by using a linear mapping without discarding any discriminant information. This mapping is comprised of the non-null eigenvectors of $\mathbf{S}^{\Phi} = \mathbf{W}^{\Phi} + \mathbf{B}^{\Phi}$. The non-null eigenvectors of $\mathbf{S}^{\Phi}$ can be calculated using the kernel matrices defined in Section III-B. First $\mathbf{S}^{\Phi}$ can be written as

$$\begin{aligned}\mathbf{S}^{\Phi} &= \mathbf{W}^{\Phi} + \mathbf{B}^{\Phi}\\ &= \sum_{\mathbf{z}\in\mathcal{I}_r}(\phi(\mathbf{z}) - \bar{\boldsymbol{\rho}})(\phi(\mathbf{z}) - \bar{\boldsymbol{\rho}})^T\\ &\quad + \sum_{\mathbf{z}\in\mathcal{U}_r}(\phi(\mathbf{z}) - \bar{\boldsymbol{\rho}})(\phi(\mathbf{z}) - \bar{\boldsymbol{\rho}})^T\\ &= \sum_{\mathbf{z}\in\mathcal{U}}(\phi(\mathbf{z}) - \bar{\boldsymbol{\rho}})(\phi(\mathbf{z}) - \bar{\boldsymbol{\rho}})^T\\ &= \sum_{i=1}^{L}\tilde{\boldsymbol{\mu}}_i\tilde{\boldsymbol{\mu}}_j^T = \mathbf{\Phi}_s\mathbf{\Phi}_s^T\end{aligned} \tag{13}$$

where $\tilde{\boldsymbol{\mu}}_i = \phi(\mathbf{z}_i) - \bar{\boldsymbol{\rho}}$ and $\mathbf{\Phi}_s = [\tilde{\boldsymbol{\mu}}_1\ldots\tilde{\boldsymbol{\mu}}_L]$. Only the first n (with $n \leq L-1$) positive eigenvalues of $\mathbf{S}^{\Phi}$ are of interest to us. These eigenvectors can be indirectly derived from the eigenvectors of the matrix $\mathbf{\Phi}_s^T\mathbf{\Phi}_s(L \times L)$.

The $\mathbf{\Phi}_s^T\mathbf{\Phi}_s$ can be expanded as

$$\begin{aligned}\mathbf{\Phi}_s^T\mathbf{\Phi}_s = \mathbf{K} - \frac{1}{L_G}\mathbf{E}\mathbf{1}_{L_GL} - \frac{1}{L_G}\mathbf{1}_{LL_G}\mathbf{E}\\ + \frac{1}{N_G^2}\mathbf{1}_{LL_G}\mathbf{K}_1\mathbf{1}_{L_GL}.\end{aligned} \tag{14}$$

The detailed derivation of (14), along with the definition of the matrices $\mathbf{K}$ and $\mathbf{E}$, can be found in Appendix E. Let λ_i^s and $\mathbf{c}_i(i = 1\ldots L_I)$ be the ith eigenvalue and the corresponding eigenvector of $\mathbf{\Phi}_s^T\mathbf{\Phi}_s$, sorted in ascending order of eigenvalues. It is true that $(\mathbf{\Phi}_s\mathbf{\Phi}_s^T)(\mathbf{\Phi}_s\boldsymbol{\varpi}_i) = \lambda_i^s(\mathbf{\Phi}_s\mathbf{c}_i)$. Thus, $\boldsymbol{\varpi}_i = \mathbf{\Phi}_s\mathbf{c}_i$ are the eigenvectors of $\mathbf{S}^{\Phi}$. In order to remove the null space of $\mathbf{S}^{\Phi}$, the first $n \leq L_I - 1$ eigenvectors (given in the matrix $\mathbf{\Pi} = [\boldsymbol{\varpi}_1\ldots\boldsymbol{\varpi}_n] = \mathbf{\Phi}_s\mathbf{C}$, where $\mathbf{C} = [\mathbf{c}_1\ldots\mathbf{c}_n]$), whose corresponding eigenvalues are nonzero should be calculated. Thus, $\mathbf{\Pi}^T\mathbf{S}^{\Phi}\mathbf{\Pi} = \mathbf{\Lambda}_s$, with $\mathbf{\Lambda}_s = \mathrm{diag}[\lambda_1^{s\,2}\ldots\lambda_n^{s\,2}]$, a $n \times n$ diagonal matrix. The orthonormal eigenvectors of $\mathbf{S}^{\Phi}$ are the columns of the matrix

$$\mathbf{\Pi}_1 = \mathbf{\Phi}_s\mathbf{\Pi}\mathbf{\Lambda_s}^{-1/2}. \tag{15}$$

When adopting fractional power polynomial models, a similar approach to Section III-B1 should be followed in order to take care of the negative eigenvalues that may occur during the eigenanalysis of $\mathbf{S}^{\Phi}$.

It can be easily proven that $\mathbf{S}^{\Phi}$ is compact and self-adjoint and, thus, the columns of the matrix $\mathbf{\Pi}_1$ form an orthonormal basis in $\mathcal{F}$. In [11], for the Fisher's discriminant ratio, this space has been the KPCA space spanned by the orthogonal eigenvectors that correspond to the non-null eigenvalues of the total scatter matrix. We define the two orthogonal complementary subspaces $\mathcal{O}$ and $\mathcal{O}^{\perp}$ of $\mathcal{F}$ ($\mathcal{F} = \mathcal{O} \oplus \mathcal{O}^{\perp}$). $\mathcal{O}$ is spanned by the column vectors of $\mathbf{\Pi}_1$. Its orthogonal $\mathcal{O}^{\perp}$ is the one that corresponds to the null space of $\mathbf{S}^{\Phi}$. We can now easily prove that there is no discriminant information in $\mathcal{O}^{\perp}$ in respect to the criterions $\mathbf{D}^{\Phi}$ and $\mathbf{D}_b^{\Phi}$, since for the vectors $\boldsymbol{\zeta} \in \mathcal{O}^{\perp}$, it is valid that $\boldsymbol{\zeta}^T\mathbf{B}^{\Phi}\boldsymbol{\zeta} = 0$ and $\boldsymbol{\zeta}^T\mathbf{W}^{\Phi}\boldsymbol{\zeta} = 0$ at the same time (please refer to Appendix A). Thus, all of the discriminant information lies inside $\mathcal{O}$.

Now, based on the previous remarks, the two alternative discriminant criteria can be defined as

$$D(\mathbf{H}) = \frac{\text{tr}[\mathbf{H}^T\mathbf{W}\mathbf{H}]}{\text{tr}[\mathbf{H}^T\mathbf{B}\mathbf{H}]} \tag{16}$$

and

$$D_b(\mathbf{H}) = \text{tr}[\mathbf{H}^T\mathbf{W}\mathbf{H}] \left(\|\boldsymbol{\eta}_i\| = 1 \quad \text{and} \quad \boldsymbol{\eta}_i^T\mathbf{B}\boldsymbol{\eta}_i = 0\right) \tag{17}$$

where $\mathbf{W} = \mathbf{\Pi}_1^T\mathbf{W}^{\Phi}\mathbf{\Pi}_1$, $\mathbf{B} = \mathbf{\Pi}_1^T\mathbf{B}^{\Phi}\mathbf{\Pi}_1$, and $\mathbf{H} = [\dots\ \boldsymbol{\eta}_i \dots]$ with $\boldsymbol{\eta}_i \in \Re^n$.

2) Feature Extraction: Let $\mathbf{\Xi} = [\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_n]$ be all of the eigenvectors of $\mathbf{B}$. The first $q = L_G - 1$ eigenvectors correspond to the nonzero eigenvalues (range space). The two orthogonal complementary subspaces of $\mathbf{B}$ are defined as $\mathcal{O}_B = \text{span}\ \{\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_q\}$ and $\mathcal{O}_B^{\perp} = \text{span}\ \{\boldsymbol{\xi}_{q+1}, \dots, \boldsymbol{\xi}_n\}$. Thus, $\Re^n = \mathcal{O}_B^{\perp} \oplus \mathcal{O}_B$. In the space $\mathcal{O}_B$, we seek for the regular discriminant projections, while in the space $\mathcal{O}_B^{\perp}$, we seek the irregular discriminant projections. It is easy to prove that for any nonzero vector $\boldsymbol{\tau}_i \in \mathcal{O}_B^{\perp}$, the inequality $\boldsymbol{\tau}_i^T\mathbf{W}\boldsymbol{\tau}_i > 0$ holds (i.e., the matrix $\mathbf{W}$ is strictly positive definite). The previous statement shows that discriminant projections exist that maximize D_b in $\mathcal{O}_B^{\perp}$. For testing a facial vector $\mathbf{y}$, two discriminant vectors $\acute{\mathbf{y}}_1$ and $\acute{\mathbf{y}}_2$ can be derived. The above procedure is summarized in Appendix F. Thus, two distinct similarity measures can be defined. The first corresponds to the regular discriminant information

$$d_r(\acute{\mathbf{y}}_1) = \|\acute{\mathbf{y}}_1 - \bar{\boldsymbol{\rho}}_1\|^2 \tag{18}$$

where $\bar{\boldsymbol{\rho}}_1$ is the regular discriminant vector of $\bar{\boldsymbol{\rho}}$. The second similarity measure corresponds to the irregular discriminant information

$$d_r(\mathbf{y}) = \|\acute{\mathbf{y}}_2 - \bar{\boldsymbol{\rho}}_2\|^2 \tag{19}$$

where $\bar{\boldsymbol{\rho}}_2$ is the irregular discriminant vector of $\bar{\boldsymbol{\rho}}$. The two similarity measures can be used in an independent fashion or can be fused using empirical or discriminant fusion rules [11], [18].

IV. Comparison With KDDA and KPCA Plus LDA

In this section, we compare the proposed techniques with the multiclass and two-class KDDA and CKFDA in terms of the number of extracted features and in terms of computational complexity.

A. Direct CSKDA Versus Multiclass and Two-Class KDDA Approach

In [9], kernel direct LDA (KDDA) has been proposed as the nonlinear extension of direct LDA (DLDA). KDDA has been proven to be effective for face recognition in [9]. For a multiclass problem, (having N classes), the method begins with the eigenanalysis of the between-class scatter matrix in a Hilbert space. The between-class scatter matrix has, at most, $N - 1$ eigenvectors that correspond to nonzero eigenvalues. Then, the within-class scatter matrix is projected to the non-null space of the between-class scatter matrix. Finally, eigenanalysis is performed to the projected within-class scatter matrix in order to produce the final discriminant transform that gives no more than $N - 1$ discriminant projections. For a two-class problem (i.e., $N = 2$), KDDA returns only one discriminant vector which is given by the projection of the samples into the difference of the two-class mean vectors.

When applying the direct optimization approach in order to find the discriminant projections of the proposed CSKDA method then, eigenanalysis is initially performed to the impostor matrix $\mathbf{W}^{\Phi}$. The matrix $\mathbf{W}^{\Phi}$ has, at most, $L_I - 1$ eigenvectors that correspond to nonzero eigenvalues. Then, the matrix $\mathbf{B}^{\Phi}$ is projected to the non-null space of $\mathbf{W}^{\Phi}$. Finally, eigenanalysis is performed on the projected matrix $\mathbf{B}^{\Phi}$. This method gives, at most, $L_I - 1$ discriminant features which, in practice, is much bigger than the $N - 1$ features given by the multiclass KDDA method and the one-dimensional space that is given by the two-class KDDA variant.

The computational complexity for finding the discriminant transform of the multiclass KDDA method is $O(L^3)$, where L is the number of training samples (the complexity of the eigenanalysis of an $L \times L$ matrix) [9], [11]. The class-specific KDDA should be calculated for every client; thus, its complexity is $O(NL^3)$. It can be easily proven that the complexity of the direct optimization of the proposed CSKDA is $O(NL^3)$ as well.

B. Two-Step CSKDA Versus Multiclass and Two-Class CKFDA Approach

The most recent method for optimizing Fisher's criterion with kernels is a combination of KPCA with LDA, the so-called CKFDA [11]. The CKFDA starts with the eigenanalysis of the total scatter matrix that has, at most, $L-1$ eigenevectors that correspond to non-null eigenvalues. The between-class and within-class scatter matrices are projected into the non-null space of the total scatter matrix. Two discriminant criteria are then formulated, one that corresponds to the null and non-null space of the within-class scatter matrix, respectively. When optimizing the two criteria, the regular and irregular discriminant transforms are produced. Both the regular and the irregular transforms give, at most, $N-1$ features. Thus, a total of $2 \times (N-1)$ features can be derived from the multiclass CKFDA approach. In two-class problems, the CKFDA procedure results in only two discriminant projections.

The proposed two-step optimization procedure begins with the eigenanalysis of the matrix $\mathbf{S}^{\Phi}$ which contains, at most, $L-1$ eigenvectors that correspond to non-null eigenvalues. Then, the impostor matrix $\mathbf{W}^{\Phi}$ and the client matrix $\mathbf{B}^{\Phi}$ are projected to the non-null space of the matrix $\mathbf{S}^{\Phi}$. Then, two optimization criteria are formulated which give, at most, $L_G - 1$ regular plus $L_I - 1$ irregular discriminant features.

The computational complexity for finding the discriminant transform of the multiclass CKFDA is $O(L^3)$ and $O(NL^3)$ for the two-class approach [11]. $O(NL^3)$ is the computational complexity of the CSKDA using the two-step optimization method.

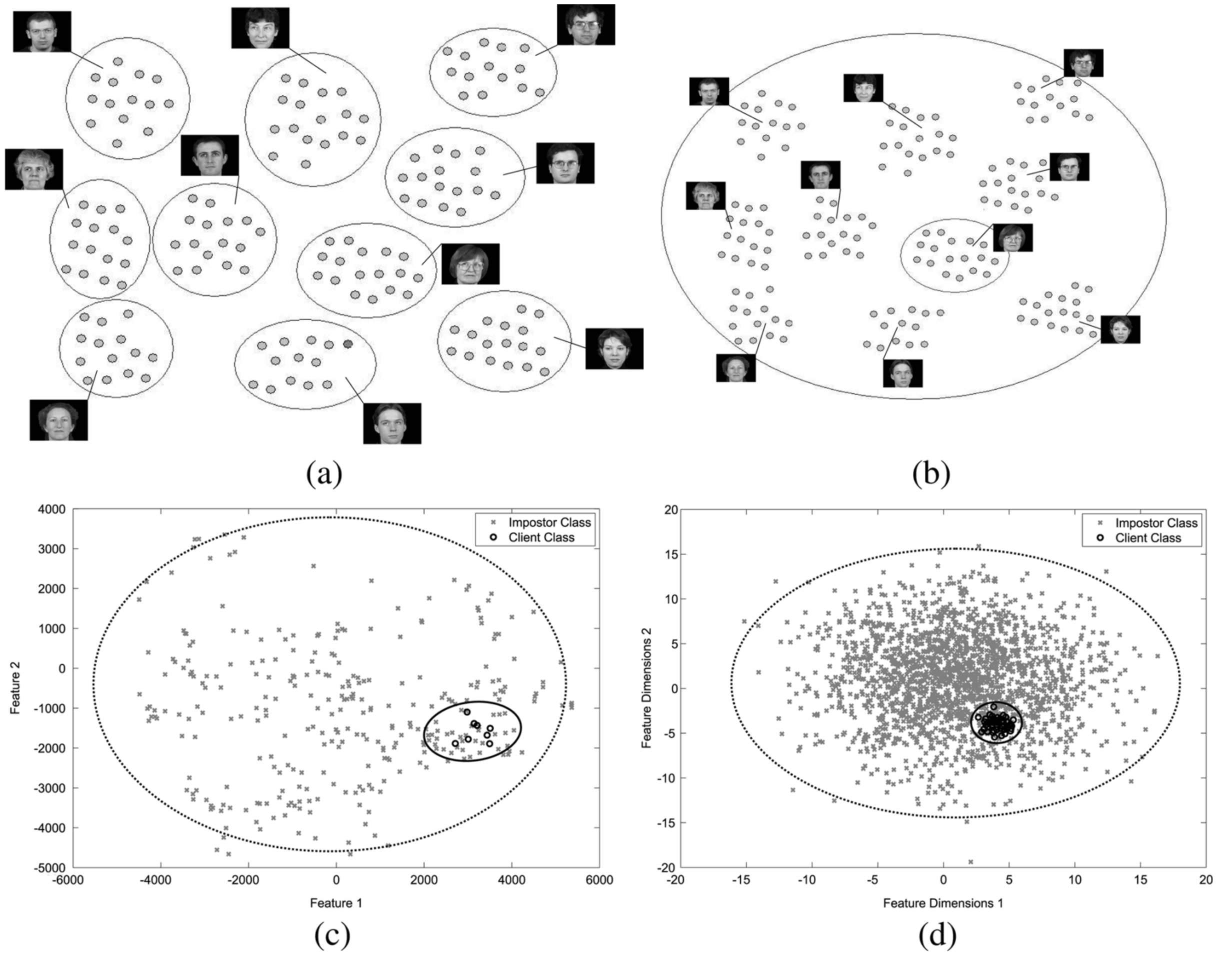


Fig. 1. (a) Multiclass face recognition modeling. (b) Two-class face verification modeling. (c) Distribution of the first two features projected to the first two principal components. (d) A simulation distribution derived from two bivariate normal distributions for impostors and clients.

V. Experiments With Artificial Data and Face Verification Modeling

It is widely accepted that the distribution of facial images, under different viewpoints, illumination variations, and facial expression is highly complex and nonlinear [9], [11], [31], [43]–[46], [10], [47]. Thus, a variety of nonlinear techniques has been developed in order to successfully capture the underlying nonlinearity of data and the most popular have been the so-called kernel techniques [9], [11], [31], [45], [10], [47]. The success of kernel techniques is mostly attributed to the fact that linear techniques can be easily modified to their nonlinear counterparts, such as FLDA to the various kernel Fisher discriminant alternatives [9], [11], [10], [47], [48].

As has already been mentioned for discovering discriminant features, there are modeling differences between face recognition and face verification. On one hand, face recognition is treated as a multiclass problem, where the space is separated to various face classes. On the other hand, the strategy for face verification is to find class-specific projections that separate the genuine (client) class from the impostor class. In Fig. 1(a) and (b), the two different modellings (i.e., face recognition and face verification) can be seen. An example of the two-class face verification problem for 39 people from the XM2VTS database is illustrated in Fig. 1(c). For every person, the first two features, derived from the projection to the two dominant eigenvectors of PCA, are depicted.

A simulation example can be found in Fig. 1(d) where two classes have been created using bivariate normal distributions. The first class represents the client class, having 50 samples, while the second one models the impostor class, containing 2000 samples. It is obvious that nonlinear methods should be applied in order to capture the distribution of the data.

In order to provide some first insights of the benefits of CSKDA, we have applied nonlinear modeling using RBF kernels in the artificial data of Fig. 1(d). The kernel Fisher discriminant alternatives give a very limited subspace of one dimension [9], [11], [49], [50]. On the other hand, KPCA [29] provides a set of features, but has the disadvantage that does not consider class distribution characteristics. For the simulation example in Fig. 1(d), the KPCA resulted in a 100-D space. The proposed approach has resulted in an 100-D space as well.

Let the similarity between a data sample in the new space and the genuine class be measured by using the Euclidean distance to the center of the genuine class. The distribution of the client and impostor similarities, after applying KFDA and CSKDA, with the client class, can be found in Fig. 2. The zoomed area represents the distribution of the client distances. As can be seen in Fig. 2(a), when using the 1-D space of KFDA, the data are somewhat well separable. When more dimensions are kept by

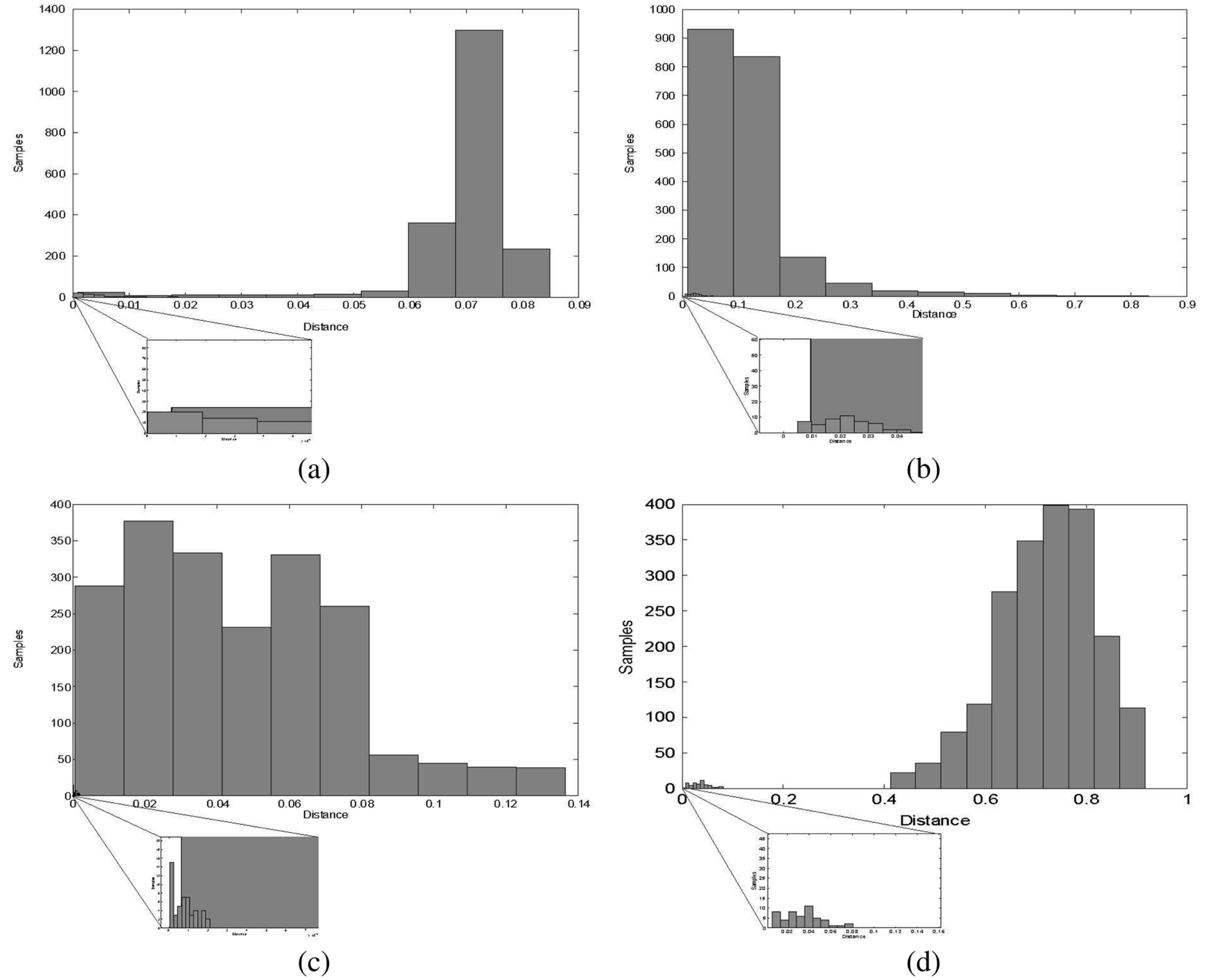


Fig. 2. Histograms of sample distances with: (a) kernel Fisher's discriminant analysis; (b) kernel Fisher's discriminant analysis with more than one dimensions by adding the a small noisy diagonal matrix to the between class scatter matrix; (c) proposed kernel-discriminant analysis with only the first dimension; and (d) proposed kernel-discriminant analysis with 100 dimensions.

adding an diagonal matrix with small noisy elements to the between-class scatter matrix, the two classes are heavily confused [see Fig. 2(b)].

In many cases, in approximation and regularization theory [48], [51], [10], a scaled version of the identity matrix is added to a matrix in order to become invertible [48], [51], [10]. The scaled version of the identity matrix is a simplified version of the noisy diagonal matrix that we have used in the experiments. Using this fact, we provide a theoretical indication (Appendix D) concerning why the use of additional dimensions of between-class scatter matrix deteriorates the performance. On the other hand, the samples of the two classes are not well separated using only the first dimension of the proposed method, but they become fully separated when using 100 dimensions.

Let the maximum distance of the client samples be considered as a threshold for accepting or rejecting a claim (this means that false rejection is equal to zero). Using this threshold in Fig. 3, a comparison of false acceptances introduced from

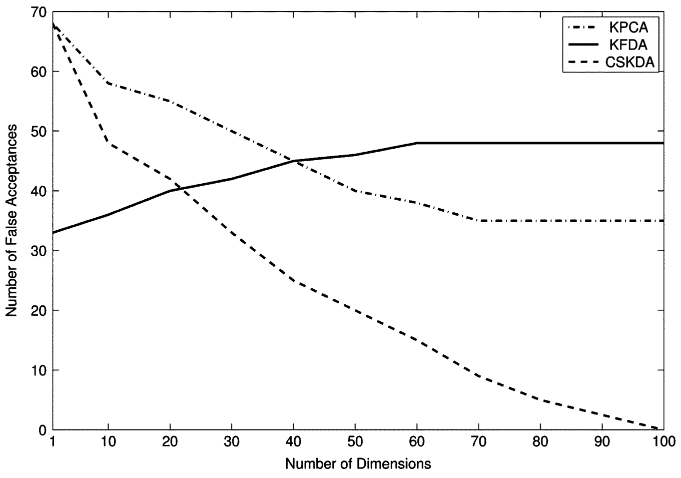


Fig. 3. Number of false acceptances versus dimensionality.

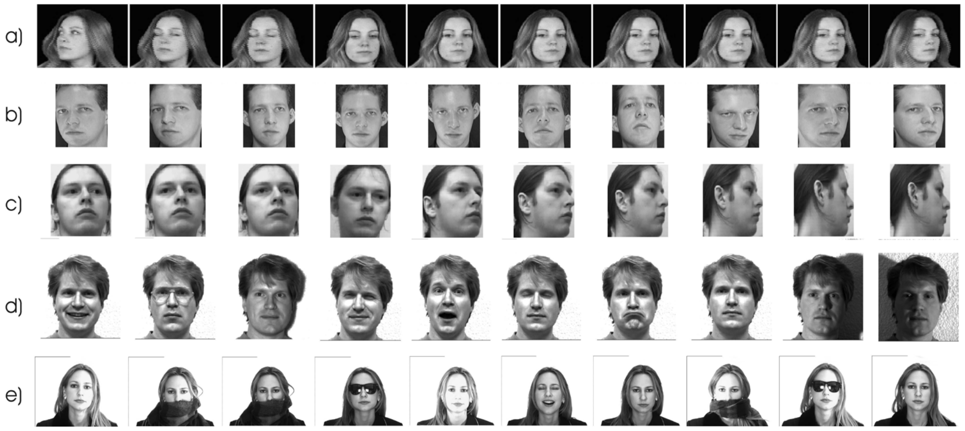


Fig. 4. Data samples used for the experimental procedure. Each row represents images taken from one session to consist of one person's class for (a) XM2VTS, (b) ORL, (c) UMIST, (d) Yale, and (e) AR.

KFDA, KPCA, and the proposed techniques for various kept dimensions is shown. As can be seen when more than one dimension is kept for KFDA, by adding the identity matrix to the between-class scatter matrix, the performance deteriorates and more false acceptances are introduced. On the other hand, the performance of KPCA and the proposed kernel technique increases with the number of kept dimensions.

The above example indicates that:

- the one-dimensional space of class-specific kernel Fisher discriminant analysis may be insufficient for correctly representing data in two class cases;
- simple tricks, such as adding noisy diagonal matrices to the between-class scatter matrix, in order to have larger KFDA spaces, deteriorates the performance;
- the proposed criterion provides a multidimensional space where the data can be well represented.

In the following section, we will show the superiority of the proposed technique against many other kernel-based approaches in face verification.

VI. Experiments With Real Data

A. Measures

In many cases for evaluating the performance of a face recognition system, only the percentage of correctly identified faces within a number of matches is adequate (recognition rate) [52], [31]. By varying the number of matches, the curve of the cumulative match score versus the number of matches is obtained [53], [54]. On the other hand, the performance of face verification systems is measured in terms of the false rejection rate (FRR) achieved at a fixed false acceptance rate (FAR) [12]. There is a tradeoff between FAR and FRR. That is, it is possible to reduce either of them with the risk of increasing the other one. This tradeoff between the FAR and FRR can create a curve where FRR is plotted as a function of FAR. This curve is called the receiver operating characteristic (ROC) curve [55], [12]. The performance of a verification system is often quoted by a particular operating point of the ROC curve where $\mathrm{FAR} = \mathrm{FRR}$. This operating point is called the equal error rate (EER). The EER will be used to quantify the performance of the tested methods in the next section.

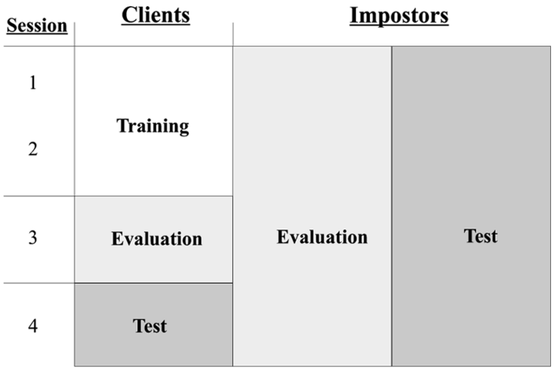
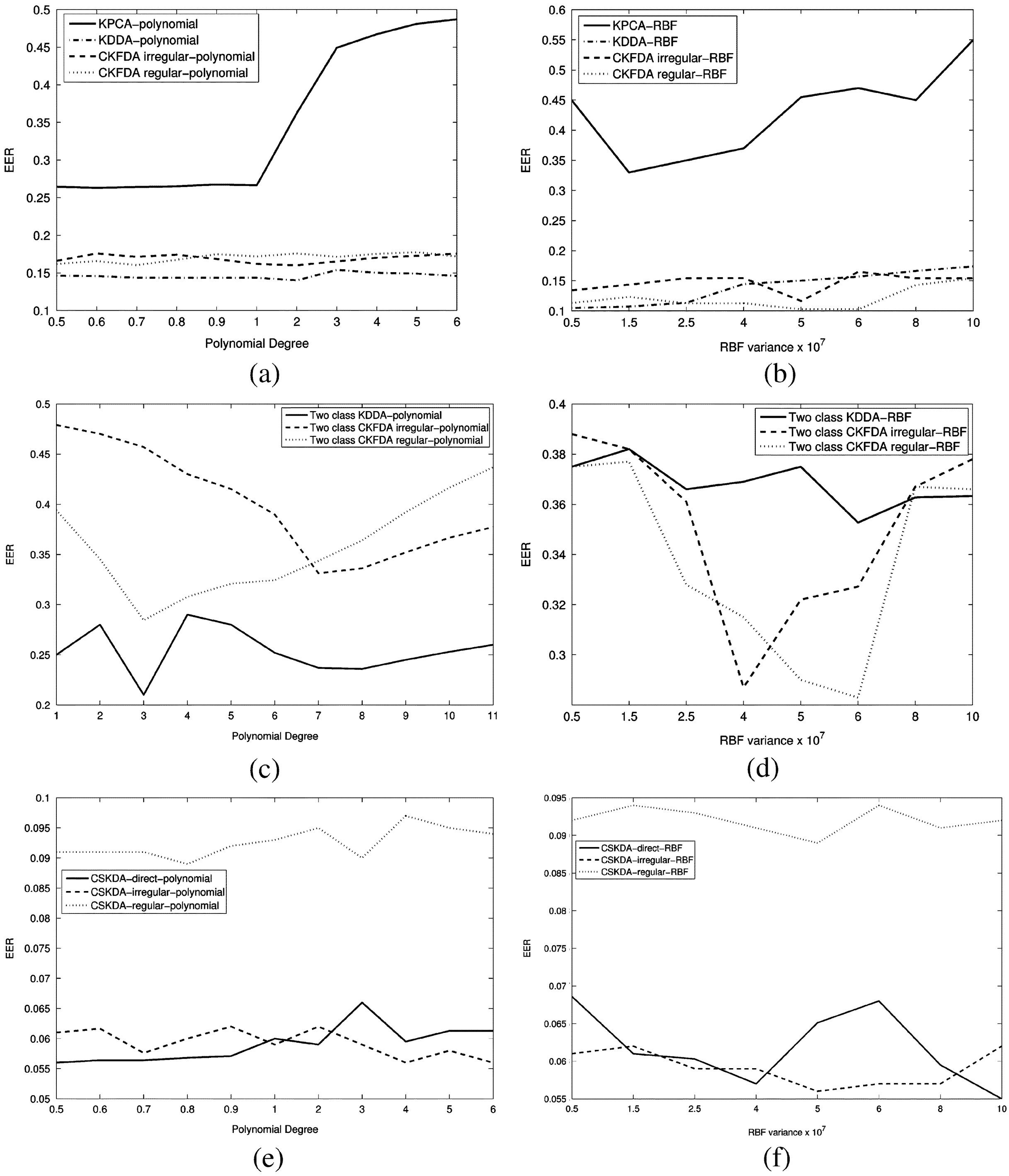


Fig. 5. Diagram showing the partitioning of the database used according to the protocol.

B. Databases

There are a number of databases and protocols used for face verification experiments in the literature. The most popular are the M2VTS [56] and XM2VTS [21], FERET [57]. Our purpose in this work is to test the proposed discriminant analysis to nonfrontal images. However, the protocols of these data bases are strict to frontal face. Thus, we used the video XM2VTS database, from which we have extracted the required frames. More specifically, we have extracted frames that represent the frontal images of the people as well as right and left profiles until approximately 60° divergence from the frontal pose and random images between [Fig. 4(a)]. The specific database contains four recordings of 295 subjects taken over a period of four months. Each recording contains a speaking head shot and a rotating head shot. In the specific procedure, only the rotation shots have been used. In order to reinforce our experimental results, we have also run experiments on the AR [22], [23], the ORL [24], Yale [25], and the UMIST [26] databases. Some of the faces in ORL, UMIST, Yale, and AR are shown in Fig. 4(b)–(e), respectively.

In the ORL database, there are each ten different images of 40 distinct subjects. For some subjects, the images were taken at different times, varying the lighting, facial expressions (open/closed eyes, smiling/not smiling), and facial details (glasses/no



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 6. (a) ERR for multiclass KPCA, KDDA, and CKFDA methods (regular and irregular space) using polynomial kernels. (b) ERR for multiclass KPCA, KDDA, and CKFDA methods (regular and irregular space) using RBF kernels. (c) EER for class-specific KDDA and CKFDA (regular and irregular space) using polynomial kernels. (d) EER for class-specific KDDA and CKFDA (regular and irregular space) using polynomial kernels. (e) EER for the proposed CSKDA method for polynomial kernels. (f) EER for the proposed CSKDA method for RBF kernels.

glasses). All of the images were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement).

The Yale face database contains 165 grayscale images in GIF format of 15 individuals. There are 11 images per subject, one per different facial expression or configuration: center-light, with glasses, happy, left-light, without glasses, normal, right-light, sad, sleepy, surprised, and wink.

The UMIST face database consists of 564 images of 20 people. Each covers a range of poses from profile to frontal views. Subjects cover a range of race, sex, and appearance. The files are all in PGM format, approximately 220 × 220 pixels in 256 shades of gray.

The AR face database contains more than 3 000 color images corresponding to 130 of people's faces (70 men and 60 women). The images feature frontal view faces with different facial expressions, illumination conditions, and occlusions (sun glasses and scarf). Each person participated in two sessions, separated by two weeks (14 days) time. The same pictures were taken in both sessions.

C. Experiments in XM2VTS

In our framework, every video of the XM2VTS database is processed frame by frame and is transformed to grayscale while it gets resized to a smaller resolution to reduce the processing time (from 720×576 to 97×68). Afterwards, the uniform background is detected using thresholding. The pixel values that correspond to the background obtain a zero value to reduce the external noise. To achieve a better and more accurate verification rate, the algorithm resizes each video, according to a factor produced by a given standard distance between the right and left eye, when the subject is in frontal position, as it keeps the frame size stable. This way, the scaling problem occurring in different sessions of the same person is resolved, while the head is aligned for the frame representing the frontal pose and, consequently, for the rest of the movement.

From the 295 people in the database, 120 randomly chosen people, including all sessions, were processed and used for the experiments. The protocol used to evaluate the method is an XM2VTS-like protocol, which has been designed similar to the protocol described in [21]. The protocol is defined for the task of person verification where an individual claims an identity. The verification system compares the features of that person with stored features corresponding to the claim identity and computes his or her similarity, which is referred to as score. Depending on the score, the system decides whether the identity claim is true. This verification task corresponds to an open test set scenario where people, unknown to the system, might claim access.

The testing database is comprised of 120 subjects, four recording sessions, and one shot of moving head per recording session. We should note here that each session in the XM2VTS as well as in the video XM2VTS database has been captured with one-month time intervals between each other. The database was randomly divided into 60 clients and 60 impostors. Two sessions out of four of the clients' class were used for training the system, while one session was used for evaluation and one for testing. For the impostors, two sessions were used for evaluation and two for testing. The number of images taken from each session for one person was 10. So for the training set, 1200 images (60 clients $\times$ 10 images $\times$ 2 sessions) were used. The number of images that were used for the evaluation set has been (60 clients$\times$ 10 images$\times$ 1 session $+$ 30 impostors$\times$ 10 images$\times$ 2 session $=$ 1200) and the test set (60 clients $\times$ 10 images$\times$ 1 session $+$ 30 impostors$\times$ 10 images$\times$ 2 session $=$ 1200), respectively. Thus, we have a total of $60 \times 10 = 600$ client claims and $60 \times 30 \times 10 \times 2 = 36000$ impostor claims for both—the evaluation and the test sets. A scheme of the experimental protocol is illustrated in Fig. 5.

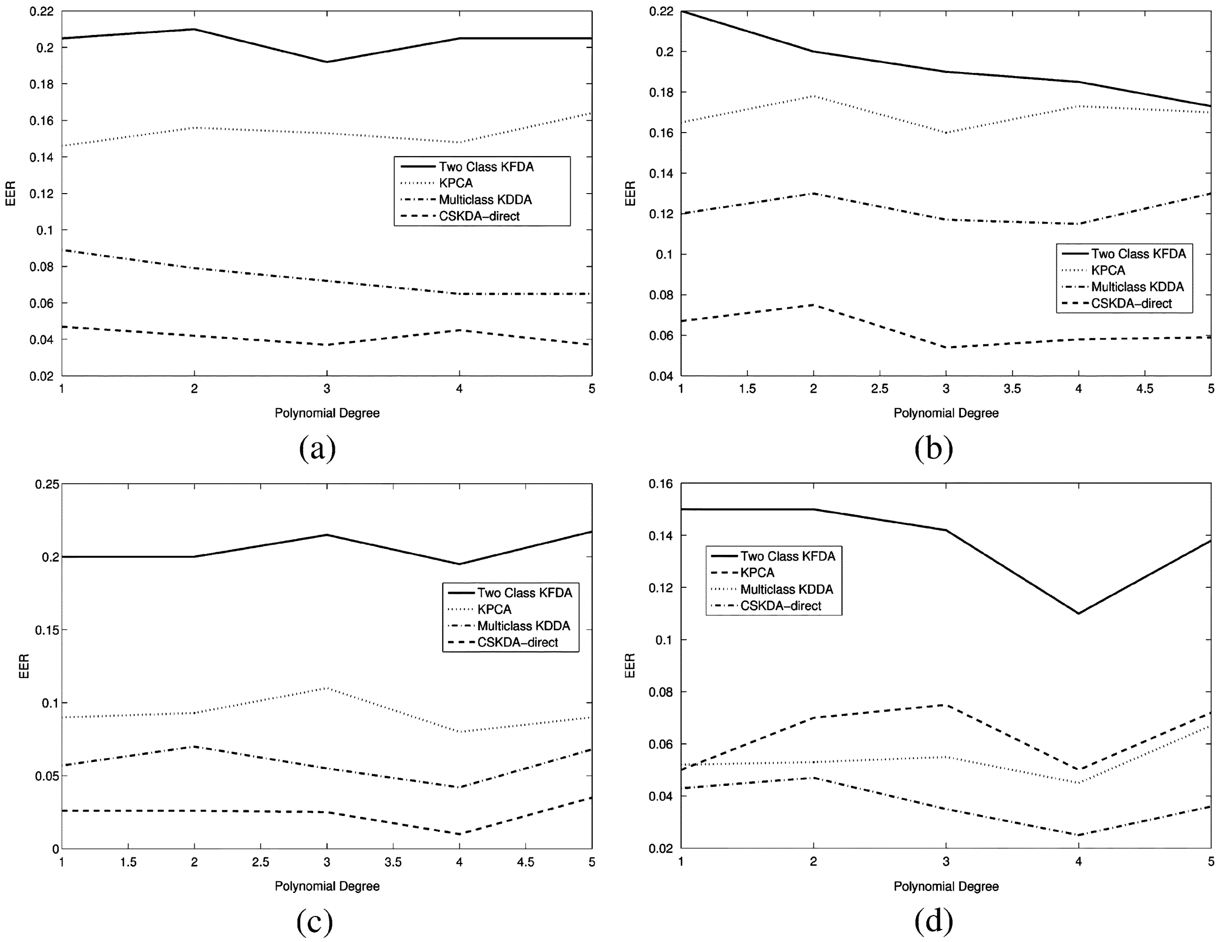
The evaluation set is used in order to produce client and impostor access scores, which are used to find a person-specific threshold that determines whether a person is accepted or rejected while it tunes the algorithm in order to find the proper kernel and the dimensionality of the new space. The threshold can be set to satisfy certain performance levels on the evaluation set. In the case of multimodal classifiers, the evaluation set might also be used to optimally combine the outputs of several classifiers. The test set is selected to simulate real authentication tests using the thresholds and the operating points calculated in the evaluation set.

TABLE I
COMPARISON OF THE BEST EERS MEASURED AT THE TEST SET

Type	Algorithm	Kernel	Parameter	EER %
Multiclass	KDDA	Polynomial	2	14
Multiclass	CSKFDA	RBF	0.5×10^7	10
Two-Class	KDDA	Polynomial	0.7	21
Two-Class	CKFDA-regular	RBF	6×10^7	28
Proposed Two-Class	CSKDA - direct	Polynomial	0.5	**5.6**
Proposed Two-Class	CKFDA - direct	RBF	$10{\times}10^7$	**5.5**

A similarity measure $d_r(\mathbf{y})$ between faces is found in all of the tested methods. In the proposed approaches, the similarity measure was the one defined in (3). For the other tested methods, we have used various similarity measures such as L_2 and the L_1 norm, the Mahalanobis distance, and the cosine similarity measure, but we will report the best results that have been achieved for all of the methods (actually the cosine similarity measure was the one that had the best performance for the other tested approaches). In order to reject or accept an identity claim, a threshold should be used on this similarity measure. For choosing the thresholds, the method proposed in [30] has been used. In detail, the similarity measures for every person are calculated in the training set and form the distance vector $\mathbf{o}(r)$. The elements of the vector $\mathbf{o}(r)$ are sorted in ascending order and are used for the person-specific thresholds on the distance measure. Let $T_Q(r)$ denote the Qth order statistic of the vector of distances $\mathbf{o}(r)$. The threshold of the person r is chosen to be equal to $T_Q(r)$. A claim of a person t is considered valid if $d_r(\mathbf{y}) < T_Q(r)$. Obviously, by varying Q, different pairs of FAR and FRR can be created to produce the ROC curve.

In the experimental procedure described in this paper, the training set has been used to train the KPCA, KDDA, CKFDA, two-class KDDA, two-class CKFDA, and the proposed CSKDA. The maximum number of features for the KPCA has been 1199; for multiclass KDDA, it has been 59 features; for multiclass CKFDA, it has been 59 for the regular discriminant information; and 59 for the irregular-discriminant information. For the two-class (genuine versus impostors) modeling of face verification, the KDDA gives only one feature and the CKFDA only two features one that corresponds to the regular and one that corresponds to the irregular-discriminant information. The proposed CSKDA gives, at maximum, 1179 features using the direct optimization approach (Section III-B). While using the two-step optimization procedure, in Section III-C, we have 19 features for the regular-discriminant information and 1179 for the irregular. For all of the methods, we have experimented



(a)

(b)

(c)

(d)

Fig. 7. ERR for KPCA, multiclass KDDA, two-class KFDA, and CKFDA-direct methods using polynomial kernels in (a) AR database, (b) UMIST database, (c) Yale database, and (d) ORL database.

using various feature dimensionality but due to space limitations, we report only the best results for all of the tested methods.

In Fig. 6(a), the EERs for the test set are plotted for various polynomial kernel parameters for the multiclass KPCA, KDDA, and CKFDA approaches (regular and irregular information). The kernels we have applied are fractional power polynomial kernels with powers from 0.5 to 0.9 and polynomial kernels with powers from 1 to 6. The EERs for RBF kernels are plotted in Fig. 6(b). The RBF kernels had similar variances as in [9] and take values from 0.5×10^7 to 10^8. The best EER achieved for these methods has been measured at about 14% for the KDDA using polynomial kernels and 10% for the KDDA using RBF kernels.

The EERs for the class-specific KDDA and CKFDA (regular and irregular) using various polynomial kernel parameters are plotted in Fig. 6(c). The corresponding best EER using RBF kernels is plotted in Fig. 6(d). The best EER using RBF kernels have been measured at about 28% for the regular CKFDA. The corresponding best EER using polynomial kernels has been measured at about 21% for the KDDA. As can be seen, the performance of the two-class variants of KDDA and CKFDA is worse than the multiclass KDDA and CKFDA. This is attributed to the very limited feature space that is provided by the two-class KDDA and CKFDA.

TABLE II
COMPARISON OF THE BEST EERS MEASURED USING SVMS AT VARIOUS FEATURE EXTRACTION METHODS

Algorithm	EER%-XM2VTS	EER%-ORL
GrayScale+SVMs	14.2	5.3
KPCA+SVMs	11.1	3.5
Multiclass KDDA + SVMs	7	3.5
CSKDA-direct + SVMs	3.8	2.5

The EERs for the proposed CSKDA methods using the direct approach and the CSKDA irregular and regular information are plotted in Fig. 6(e) for various polynomial kernel parameters. The corresponding EERs for the RBF kernels are plotted in Fig. 6(f). The best EER for the proposed methods has been measured at about 5.6% for the polynomial kernels using the direct optimization approach. The best EER using RBF kernels has been measured at about 5.5% for the direct optimization approach as well.

A comparison of the best EERs for the tested methods can be found in Table I. The best EER that has been achieved by our method is measured at about 5.5%, which is a very good

TABLE III
COMPARISON OF THE BEST EERs MEASURED USING GABOR FEATURE VECTORS AND FRACTIONAL POLYNOMIAL MODELS AT VARIOUS FEATURE EXTRACTION METHODS

Algorithm	EER%-XM2VTS	EER%-ORL	EER%-Yale
Gabor + KPCA with Fractional Polynomial Models	10.2	5.3	7
Gabor + Multiclass KFDA with Fractional Polynomial Models	6.8	4.2	3.4
Gabor + CSKDA with Fractional Polynomial Models	3.3	3.2	1.6

performance if we consider that the database contains faces at various poses. We have also compared our method to the kernel Fisher discriminant variant, proposed in [48], but it resulted in an EER of no less than 20% (such as class-specific KDDA and CSKDA). Thus, we do not include detailed experiments with this variant.

In order to understand whether the proposed CSKDA approach is statistically significantly better than the other tested approaches, the McNemar's [58]–[61], [11] has been used. McNemar's 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 the performance difference between two algorithms is considered to be statistically significant. Using this test, it has been verified that the proposed methods CSKDA-direct and CSKDA-irregular outperform the other tested classifiers in the demonstrated experiments at a significant level that is less than $p = 10^{-5}$. Moreover, we have measured that the difference between multiclass KDDA and KPCA is statistically significant (this also holds for multiclass CKFDA-regular and irregular). The difference between multiclass KDDA and two-class KDDA is also statistically significant (this also holds for multiclass CKFDA-regular and irregular).

D. Experiments in AR

A similar experimental protocol to the one applied in the XM2VTS database has been used for the AR database. That is, we have used 120 people of the AR database and they have been distributed to clients and impostors as described before. The grayscale information at a resolution of 97×68 has been considered in the experiments.

In Fig. 7(a), the EER is plotted versus the degree of the polynomial kernel for the AR face database. The two-class KFDA is an implementation of the algorithm in [48]. As can be seen, the proposed method outperforms all other tested methods.

E. Experiments in UMIST–Yale–ORL

For the UMIST, Yale, and ORL database, in order to make maximal use of the data, we have considered a circular protocol. In order to implement this protocol, we have combined principles of the leave one out strategy and the rotation estimates (i.e., a variant of the jack-knife method [30], [14], [13], [12]). In each circle of the protocol, one person becomes the impostor and the images are used for impostor claims (not seen in the training phase). Then, 80% of the data of the remaining people are used for training and the remaining 20% serve for client claims.

In Fig. 7(b)–(d), the EER is plotted versus the degree of the polynomial kernel for the UMIST, Yale, and ORL face databases, respectively. The proposed method CSKDA-direct outperforms all of the other tested methods.

F. Experiments With SVMs

Among the most popular methods used in pattern classification applications are SVMs [62]. Motivated by the successful application of SVMs with KPCA in the ORL database for face recognition [63], we have combined SVMs with different feature extract methods tested in this paper. That is, we have combined SVMs with KPCA, multiclass–KDDA, the proposed CSKDA–direct and, for completeness, we have applied SVMs directly to the grayscale facial images. We have tested these in XM2VTS and ORL databases. The best EERs achieved are summarized in Table II.

G. Experiments With Gabor Features and Fractional Polynomial Models

Gabor-based facial features, combined with kernel methods (e.g., KPCA [31] and variants of multiclass kernel Fisher's discriminant analysis [10]) and with fractional polynomial models, are among the state-of-the-art face verification and recognition systems in the literature. In these methods, an information pyramid is created by applying a Gabor filter bank to the original facial image. Afterwards, an augmented feature vector is created by concatenating the various output magnitude images from the Gabor multiscale analysis. Finally, a subspace is discovered by applying KPCA [31] or kernel Fisher discriminant analysis variants with fractional polynomial models [10]. We have conducted experiments using the augmented Gabor features proposed in [10] and [31]. Moreover, we have applied the proposed method using these Gabor features and we have verified that it has superior performance and outperforms Gabor–KPCA and multiclass Gabor–KFDA with fractional polynomial models. The tests have been conducted in the XM2VTS, ORL, and Yale database. The best EERs in the various tested databases are summarized in Table III. The proposed CSKDA-direct has the best performance when combined with Gabor features.

VII. CONCLUSION

Face verification has been modelled as a nonlinear two-class problem (clients versus impostors). The majority of discriminant feature extraction methods that are used for face recognition are based on Fisher's discriminant analysis. The problem with Fisher's discriminant analysis is that in two class problems, it provides only one or two discriminant dimensions. This limited subspace may be proved insufficient for feature extraction for face verification. We have demonstrated this by extensive experiments using both artificial and real data. We have also demonstrated that other kind of tricks, such as adding small noisy diagonal matrices to the between-class scatter matrix in order to extract additional features, have not led to performance improvement. In order to extract additional features in two class cases, novel kernel-based methods for discriminant feature extractions have been defined. The proposed method minimizes the variance of the client class around the client class center, while maximizing the distance of the impostor samples from the client center. It has been proven that the proposed criterion produces a set of discriminant projections with their number being proportional to the number of training samples. The proposed approaches have been tested in face verification using various face databases, where they show how to outperform many other popular kernel methods.

Appendix A
Proof of Proposition 1

Proposition 1: Let that for a vector $\boldsymbol{\zeta}$, it is valid that $\boldsymbol{\zeta}^T\mathbf{B}^\Phi\boldsymbol{\zeta} = 0$ then if it satisfies $\boldsymbol{\zeta}^T\mathbf{W}^\Phi\boldsymbol{\zeta} > 0$ at the same time, all of the training client vectors are mapped on the same point $\delta = \boldsymbol{\zeta}^T\phi(\boldsymbol{\rho}_i)$ and not all of the training impostor vectors do not fall on δ.

Proof: Let the matrix $\mathbf{X}^\Phi = [\phi(\boldsymbol{\rho}_1)\ldots\phi(\boldsymbol{\rho}_{N_G})]$ that has the training client vectors in $\mathcal{F}$ as columns, then the matrix $\mathbf{B}^\Phi$ can be written as

$$\begin{aligned}\mathbf{B}^\Phi &= \sum_{i=1}^{N_G}(\phi(\boldsymbol{\rho}_i)-\bar{\boldsymbol{\rho}})(\phi(\boldsymbol{\rho}_i)-\bar{\boldsymbol{\rho}})^T\\ &= \left(\mathbf{X}^\Phi - \mathbf{X}^\Phi\mathbf{G}_{N_G}\right)\left(\mathbf{X}^\Phi - \mathbf{X}^\Phi\mathbf{G}_{N_G}\right)^T \qquad (20)\end{aligned}$$

where $\mathbf{G}_{N_G}$ is a matrix with elements that are equal to N_G^{-1}. Then, by letting $\mathbf{I}_{N_G}$ be the identity $N_G \times N_G$ matrix, we have

$$\begin{aligned}\boldsymbol{\zeta}^T\mathbf{B}^\Phi\boldsymbol{\zeta} = 0 \Leftrightarrow \boldsymbol{\zeta}^T&\left(\mathbf{X}^\Phi - \mathbf{X}^\Phi\mathbf{G}_{N_G}\right)\\ \times&\left(\mathbf{X}^\Phi - \mathbf{X}^\Phi\mathbf{G}_{N_G}\right)^T\boldsymbol{\zeta} = 0\\ \left\|\left(\mathbf{I}_N - \mathbf{G}_{N_G}\right)\mathbf{X}^{\Phi^T}\boldsymbol{\zeta}\right\|^2&\\ = 0 \Leftrightarrow \boldsymbol{\zeta}^T&\phi(\boldsymbol{\rho}_i) = \boldsymbol{\zeta}^T\bar{\boldsymbol{\rho}}\\ &\forall i = 1,\ldots,N_G. \qquad (21)\end{aligned}$$

Thus, all of the training client vectors fall on the same point $\boldsymbol{\zeta}^T\bar{\boldsymbol{\rho}}$. In the same manner, when $\boldsymbol{\zeta}^T\mathbf{W}^\Phi\boldsymbol{\zeta} > 0$ the impostor vectors do not fall on the same point, while if $\boldsymbol{\zeta}^T\mathbf{W}^\Phi\boldsymbol{\zeta} = 0$, then the training impostors fall to $\boldsymbol{\zeta}^T\bar{\boldsymbol{\rho}}$ as well. Thus, when for a vector $\boldsymbol{\zeta}$, it is valid that $\boldsymbol{\zeta}^T\mathbf{B}^\Phi\boldsymbol{\zeta} = 0$ and $\boldsymbol{\zeta}^T\mathbf{W}^\Phi\boldsymbol{\zeta} = 0$ at the same time, then no discrimination can be performed in a space spanned by vectors such as $\boldsymbol{\zeta}$ since all of the vectors are mapped to a specific point.

Appendix B
Computation of $\boldsymbol{\Phi}_w^T\boldsymbol{\Phi}_w$

The $\boldsymbol{\Phi}_w^T\boldsymbol{\Phi}_w$ is expanded as

$$\boldsymbol{\Phi}_w^T\boldsymbol{\Phi}_w = \left[\tilde{\boldsymbol{\kappa}}_1\ldots\tilde{\boldsymbol{\kappa}}_{L_I}\right]^T\left[\tilde{\boldsymbol{\kappa}}_1\ldots\tilde{\boldsymbol{\kappa}}_{L_I}\right] = \left[\tilde{\boldsymbol{\kappa}}_i^T\tilde{\boldsymbol{\kappa}}_j\right] \qquad (22)$$

where

$$\begin{aligned}\tilde{\boldsymbol{\kappa}}_i^T\tilde{\boldsymbol{\kappa}}_j &= \phi(\boldsymbol{\kappa}_i)^T\phi(\boldsymbol{\kappa}_j) - \frac{1}{L_G}\sum_{m=1}^{L_G}\phi(\boldsymbol{\kappa}_i)^T\phi(\boldsymbol{\rho}_m)\\ &\quad - \frac{1}{L_G}\sum_{m=1}^{L_G}\phi(\boldsymbol{\rho}_m)^T\phi(\boldsymbol{\kappa}_j)\\ &\quad + \frac{1}{N_G^2}\sum_{m=1}^{L_G}\phi(\boldsymbol{\rho}_m)^T\phi(\boldsymbol{\rho}_m)\\ &= [\mathbf{K}_4]_{i,j} - \left[\frac{1}{L_G}\mathbf{K}_2\mathbf{1}_{L_GL_I}\right]_{i,j}\\ &\quad - \left[\frac{1}{L_G}\mathbf{1}_{L_IL_G}\mathbf{K}_3\right]_{i,j}\\ &\quad + \left[\frac{1}{N_G^2}\mathbf{1}_{L_IL_G}\mathbf{K}_1\mathbf{1}_{L_GL_I}\right]_{i,j}\\ &= \left[\mathbf{K}_4 - \frac{1}{L_G}\mathbf{K}_2\mathbf{1}_{L_GL_I}\right.\\ &\quad \left. - \frac{1}{L_G}\mathbf{1}_{L_IL_G}\mathbf{K}_3 + \frac{1}{N_G^2}\mathbf{1}_{L_IL_G}\mathbf{K}_1\mathbf{1}_{L_GL_I}\right]_{i,j}.\end{aligned} \qquad (23)$$

Thus

$$\begin{aligned}\boldsymbol{\Phi}_w^T\boldsymbol{\Phi}_w = \mathbf{K}_4 - \frac{1}{L_G}\mathbf{K}_2\mathbf{1}_{L_GL_I} - \frac{1}{L_G}\mathbf{1}_{L_IL_G}\mathbf{K}_3&\\ + \frac{1}{N_G^2}\mathbf{1}_{L_IL_G}\mathbf{K}_1\mathbf{1}_{L_GL_I}.& \qquad (24)\end{aligned}$$

Appendix C
Computation of $\boldsymbol{\Phi}_w^T\mathbf{B}^\Phi\boldsymbol{\Phi}_w$

When expanding $\boldsymbol{\Phi}_w^T\mathbf{B}^\Phi\boldsymbol{\Phi}_w$, we have

$$\begin{aligned}\boldsymbol{\Phi}_w^T\mathbf{B}^\Phi\boldsymbol{\Phi}_w &= \left[\tilde{\boldsymbol{\kappa}}_1\ldots\tilde{\boldsymbol{\kappa}}_{L_I}\right]^T\mathbf{B}\left[\tilde{\boldsymbol{\kappa}}_1\ldots\tilde{\boldsymbol{\kappa}}_{L_I}\right]\\ &= \left[\tilde{\boldsymbol{\kappa}}_i^T\mathbf{B}^\Phi\tilde{\boldsymbol{\kappa}}_j\right]\end{aligned} \qquad (25)$$

where

$$\begin{aligned}\tilde{\boldsymbol{\kappa}}_i^T\mathbf{B}^\Phi\tilde{\boldsymbol{\kappa}}_j &= \tilde{\boldsymbol{\kappa}}_i^T\sum_{m=1}^{L_G}(\phi(\boldsymbol{\rho}_m)-\bar{\boldsymbol{\rho}})(\phi(\boldsymbol{\rho}_m)-\bar{\boldsymbol{\rho}})^T\tilde{\boldsymbol{\kappa}}_j\\ &= \tilde{\boldsymbol{\kappa}}_i^T\left(\sum_{m=1}^{L_G}\phi(\boldsymbol{\rho}_m)\phi(\boldsymbol{\rho}_m)^T\right.\\ &\quad - \sum_{m=1}^{L_G}\phi(\boldsymbol{\rho}_m)\bar{\boldsymbol{\rho}}^T\\ &\quad \left. - \sum_{m=1}^{L_G}\bar{\boldsymbol{\rho}}\phi(\boldsymbol{\rho}_m)^T + \sum_{m=1}^{L_G}\bar{\boldsymbol{\rho}}\bar{\boldsymbol{\rho}}^T\right)\tilde{\boldsymbol{\kappa}}_j\\ &= \tilde{\boldsymbol{\kappa}}_i^T\left(\sum_{m=1}^{L_G}\phi(\boldsymbol{\rho}_m)\phi(\boldsymbol{\rho}_m)^T - L_G\bar{\boldsymbol{\rho}}\bar{\boldsymbol{\rho}}^T\right)\tilde{\boldsymbol{\kappa}}_j\\ &= \sum_{m=1}^{L_G}\tilde{\boldsymbol{\kappa}}_i^T\phi(\boldsymbol{\rho}_m)\phi(\boldsymbol{\rho}_m)^T\tilde{\boldsymbol{\kappa}}_j - L_G\tilde{\boldsymbol{\kappa}}_i^T\bar{\boldsymbol{\rho}}\bar{\boldsymbol{\rho}}^T\tilde{\boldsymbol{\kappa}}_j.\end{aligned} \qquad (26)$$

We expand the first term $[\mathbf{A}_1]_{i,j} = \sum_{m=1}^{L_G} \tilde{\boldsymbol{\kappa}}_i^T \phi(\boldsymbol{\rho}_m)\phi(\boldsymbol{\rho}_m)^T \tilde{\boldsymbol{\kappa}}_j$ of (26) as

$$\begin{aligned}
\mathbf{A}_1 &= \sum_{m=1}^{L_G} (\phi(\boldsymbol{\kappa}_i) - \bar{\boldsymbol{\rho}})^T \phi(\boldsymbol{\rho}_m)\phi(\boldsymbol{\rho}_m)^T (\phi(\boldsymbol{\kappa}_j) - \bar{\boldsymbol{\rho}}) \\
&= \sum_{m=1}^{L_G} (\phi(\boldsymbol{\kappa}_i)^T \phi(\boldsymbol{\rho}_m)\phi(\boldsymbol{\rho}_m)^T \phi(\boldsymbol{\kappa}_j) \\
&\quad - \phi(\boldsymbol{\kappa}_i)^T \phi(\boldsymbol{\rho}_m)\phi(\boldsymbol{\rho}_m)^T \bar{\boldsymbol{\rho}} \\
&\quad - \bar{\boldsymbol{\rho}}^T \phi(\boldsymbol{\rho}_m)\phi(\boldsymbol{\rho}_m)^T \phi(\boldsymbol{\kappa}_j) \\
&\quad + \bar{\boldsymbol{\rho}}^T \phi(\boldsymbol{\rho}_m)\phi(\boldsymbol{\rho}_m)^T \bar{\boldsymbol{\rho}}) \\
&= \sum_{m=1}^{L_G} \phi(\boldsymbol{\kappa}_i)^T \phi(\boldsymbol{\rho}_m)\phi(\boldsymbol{\rho}_m)^T \phi(\boldsymbol{\kappa}_j) \\
&\quad - \sum_{m=1}^{L_G} \phi(\boldsymbol{\kappa}_i)^T \phi(\boldsymbol{\rho}_m)\phi(\boldsymbol{\rho}_m)^T \bar{\boldsymbol{\rho}} \\
&\quad - \sum_{m=1}^{L_G} \bar{\boldsymbol{\rho}}^T \phi(\boldsymbol{\rho}_m)\phi(\boldsymbol{\rho}_m)^T \phi(\boldsymbol{\kappa}_j) \\
&\quad + \sum_{m=1}^{L_G} \bar{\boldsymbol{\rho}}^T \phi(\boldsymbol{\rho}_m)\phi(\boldsymbol{\rho}_m)^T \bar{\boldsymbol{\rho}} \\
&= \Bigg[\mathbf{K}_2\mathbf{K}_3 - \frac{1}{L_G}\mathbf{K}_2\mathbf{K}_1\mathbf{1}_{L_G L_I} \\
&\quad - \frac{1}{L_G}\mathbf{1}_{L_I L_G}\mathbf{K}_1\mathbf{K}_3 + \frac{1}{N_G^2}\mathbf{1}_{L_I L_G}\mathbf{K}_1^2\mathbf{1}_{L_G L_I}\Bigg]_{i,j}. \qquad (27)
\end{aligned}$$

Then, the second term $[\mathbf{A}_2]_{i,j} = \tilde{\boldsymbol{\kappa}}_i^T \bar{\boldsymbol{\rho}}\bar{\boldsymbol{\rho}}^T \tilde{\boldsymbol{\kappa}}_j$ of (26) can be expanded as

$$\begin{aligned}
\mathbf{A}_2 &= (\phi(\boldsymbol{\kappa}_i) - \bar{\boldsymbol{\rho}})^T \bar{\boldsymbol{\rho}}\bar{\boldsymbol{\rho}}^T (\phi(\boldsymbol{\kappa}_j) - \bar{\boldsymbol{\rho}}) \\
&= \phi(\boldsymbol{\kappa}_i)^T \bar{\boldsymbol{\rho}}\bar{\boldsymbol{\rho}}^T \phi(\boldsymbol{\kappa}_j) - \phi(\boldsymbol{\kappa}_i)^T \bar{\boldsymbol{\rho}}\bar{\boldsymbol{\rho}}^T \\
&\quad \times \bar{\boldsymbol{\rho}} - \bar{\boldsymbol{\rho}}^T \bar{\boldsymbol{\rho}}\bar{\boldsymbol{\rho}}^T \phi(\boldsymbol{\kappa}_j) + \bar{\boldsymbol{\rho}}^T \bar{\boldsymbol{\rho}}\bar{\boldsymbol{\rho}}^T \bar{\boldsymbol{\rho}} \\
&= \frac{1}{N_G^2}\phi(\boldsymbol{\kappa}_i)^T \sum_{m=1}^{L_G} \phi(\boldsymbol{\rho}_m) \sum_{m=1}^{L_G} \phi(\boldsymbol{\rho}_m)^T \phi(\boldsymbol{\kappa}_j) \\
&\quad - \frac{1}{N_G^3}\phi(\boldsymbol{\kappa}_i)^T \sum_{m=1}^{L_G} \phi(\boldsymbol{\rho}_m) \\
&\quad \times \sum_{m=1}^{L_G} \phi(\boldsymbol{\rho}_m)^T \sum_{m=1}^{L_G} \phi(\boldsymbol{\rho}_m) \\
&\quad - \frac{1}{N_G^3} \sum_{m=1}^{L_G} \phi(\boldsymbol{\rho}_m)^T \\
&\quad \times \sum_{m=1}^{L_G} \phi(\boldsymbol{\rho}_m) \sum_{m=1}^{L_G} \phi(\boldsymbol{\rho}_m)^T \phi(\boldsymbol{\kappa}_i) \\
&\quad + \frac{1}{N_G^4} \sum_{m=1}^{L_G} \phi(\boldsymbol{\rho}_m)^T \sum_{m=1}^{L_G} \phi(\boldsymbol{\rho}_m) \\
&\quad \times \sum_{m=1}^{L_G} \phi(\boldsymbol{\rho}_m)^T \sum_{m=1}^{L_G} \phi(\boldsymbol{\rho}_m) \\
&= \Bigg[\frac{1}{N_G^2}\mathbf{K}_2\mathbf{1}_{L_G L_G}\mathbf{K}_3 - \frac{1}{N_G^3}\mathbf{K}_2\mathbf{1}_{L_G L_G}\mathbf{K}_1\mathbf{1}_{L_G L_I} \\
&\quad - \frac{1}{N_G^3}\mathbf{1}_{L_I L_G}\mathbf{K}_1\mathbf{1}_{L_G L_G}\mathbf{K}_3 \\
&\quad + \frac{1}{N_G^4}\mathbf{1}_{L_I L_G}\mathbf{K}_1\mathbf{1}_{L_G L_G}\mathbf{K}_1\mathbf{1}_{L_G L_I}\Bigg]_{i,j}. \qquad (28)
\end{aligned}$$

Thus

$$\boldsymbol{\Phi}_w^T \mathbf{B}^{\Phi} \boldsymbol{\Phi}_w = \left[\tilde{\boldsymbol{\kappa}}_i^T \mathbf{B}^{\Phi} \tilde{\boldsymbol{\kappa}}_j\right] = \mathbf{A}_1 - L_G\mathbf{A}_2. \qquad (29)$$

Appendix D
Feature Extraction for CSKDA Using Direct Approach

$$\begin{aligned}
\acute{\mathbf{y}} &= \tilde{\boldsymbol{\Gamma}}^T \phi(\mathbf{y}) = \left(\boldsymbol{\Phi}_w \mathbf{R}\boldsymbol{\Lambda}_w^{-1/2}\mathbf{P}\tilde{\boldsymbol{\Lambda}}_b^{-1/2}\right)^T \phi(\mathbf{y}) \\
&= \left(\mathbf{R}\boldsymbol{\Lambda}_w^{-1/2}\mathbf{P}\tilde{\boldsymbol{\Lambda}}_b^{-1/2}\right)^T \left(\boldsymbol{\Phi}_w^T \phi(\mathbf{y})\right). \qquad (30)
\end{aligned}$$

The vector $\boldsymbol{\Phi}_w^T \phi(\mathbf{y})$ can be expanded as

$$\begin{aligned}
\boldsymbol{\Phi}_w^T \phi(\mathbf{y}) &= \left[\tilde{\boldsymbol{\kappa}}_1 \ldots \tilde{\boldsymbol{\kappa}}_{L_I}\right]^T \phi(\mathbf{y}) \\
&= \left[\phi(\boldsymbol{\kappa}_1) \ldots \phi\left(\boldsymbol{\kappa}_{L_I}\right)\right]^T \phi(\mathbf{y}) \\
&\quad - [\bar{\boldsymbol{\rho}} \ldots \bar{\boldsymbol{\rho}}]^T \phi(\mathbf{y}) \\
&= \left[\phi(\boldsymbol{\kappa}_1) \ldots \phi(\boldsymbol{\kappa}_{L_I})\right]^T \phi(\mathbf{y}) \\
&\quad - \frac{1}{L_G}\mathbf{1}_{L_I L_G} \left[\phi(\boldsymbol{\rho}_1) \ldots \phi(\boldsymbol{\rho}_{L_G})\right]^T \phi(\mathbf{y}) \\
&= \left[k(\boldsymbol{\kappa}_1, \mathbf{y}) \ldots k\left(\boldsymbol{\kappa}_{L_I}, \mathbf{y}\right)\right]^T \\
&\quad - \frac{1}{L_G}\mathbf{1}_{L_I L_G} \left[k(\boldsymbol{\rho}_1, \mathbf{y}) \ldots k\left(\boldsymbol{\rho}_{L_G}, \mathbf{y}\right)\right] \\
&= \boldsymbol{\eta}(\phi(\mathbf{y})) - \frac{1}{L_G}\mathbf{1}_{L_I L_G}\boldsymbol{\omega}(\phi(\mathbf{y})) \qquad (31)
\end{aligned}$$

where $\boldsymbol{\eta}(\phi(\mathbf{y})) = [k(\boldsymbol{\kappa}_1, \mathbf{y}) \ldots k(\boldsymbol{\kappa}_{L_I}, \mathbf{y})]^T$ and $\boldsymbol{\omega}(\phi(\mathbf{y})) = [k(\boldsymbol{\rho}_1, \mathbf{y}) \ldots k(\boldsymbol{\rho}_{L_G}, \mathbf{y})]^T$ are $L_I \times 1$ and $L_G \times 1$ kernel vectors, respectively. By combining (30) and (31), we obtain

$$\begin{aligned}
\acute{\mathbf{y}} &= \tilde{\boldsymbol{\Gamma}}^T \phi(\mathbf{y}) \\
&= \left(\mathbf{R}\boldsymbol{\Lambda}_w^{-1/2}\mathbf{P}\tilde{\boldsymbol{\Lambda}}_b^{-1/2}\right)^T (\boldsymbol{\eta}(\phi(\mathbf{y})) \\
&\quad - \frac{1}{L_G}\mathbf{1}_{L_I L_G}\boldsymbol{\omega}(\phi(\mathbf{y}))). \qquad (32)
\end{aligned}$$

Appendix E
Computation of $\boldsymbol{\Phi}_s^T \boldsymbol{\Phi}_s$

The $\boldsymbol{\Phi}_s^T \boldsymbol{\Phi}_s$ is expanded as

$$\boldsymbol{\Phi}_s^T \boldsymbol{\Phi}_s = \left[\tilde{\boldsymbol{\mu}}_1 \ldots \tilde{\boldsymbol{\mu}}_{L_I}\right]^T \left[\tilde{\boldsymbol{\mu}}_1 \ldots \tilde{\boldsymbol{\mu}}_{L_I}\right] = \left[\tilde{\boldsymbol{\mu}}_i^T \tilde{\boldsymbol{\mu}}_j\right] \qquad (33)$$

where

$$\begin{aligned}
\tilde{\boldsymbol{\mu}}_i^T \tilde{\boldsymbol{\mu}}_j &= \phi(\mathbf{z}_i)^T \phi(\mathbf{z}_j) - \phi(\mathbf{z}_i)^T \bar{\boldsymbol{\rho}} - \bar{\boldsymbol{\rho}}^T \phi(\mathbf{z}_j) + \bar{\boldsymbol{\rho}}^T \bar{\boldsymbol{\rho}} \\
&= [\mathbf{K}]_{i,j} - \frac{1}{L_G} \sum_{m=1}^{L_G} \phi(\mathbf{z}_i)^T \phi(\boldsymbol{\rho}_m) \\
&\quad - \frac{1}{L_G} \sum_{m=1}^{L_G} \phi(\boldsymbol{\rho}_m)^T \phi(\mathbf{z}_j)
\end{aligned}$$

$$\begin{aligned}
&+\frac{1}{N_G^2}\sum_{m=1}^{L_G}\phi(\boldsymbol{\rho}_m)^T\sum_{m=1}^{L_G}\phi(\boldsymbol{\rho}_m)\\
&=[\mathbf{K}]_{i,j}-\left[\frac{1}{L_G}\mathbf{E}\mathbf{1}_{L_GL}\right]_{i,j}\\
&-\left[\frac{1}{L_G}\mathbf{1}_{LL_G}\mathbf{E}\right]_{i,j}+\left[\frac{1}{N_G^2}\mathbf{1}_{LL_G}\mathbf{K}_1\mathbf{1}_{L_GL}\right]_{i,j}\\
&=\left[\mathbf{K}-\frac{1}{L_G}\mathbf{E}\mathbf{1}_{L_GL}-\frac{1}{L_G}\mathbf{1}_{LL_G}\mathbf{E}\right.\\
&\left.+\frac{1}{N_G^2}\mathbf{1}_{LL_G}\mathbf{K}_1\mathbf{1}_{L_GL}\right]_{i,j}
\end{aligned}\tag{34}$$

where $\mathbf{K}$ is the total kernel function defined as

$$\mathbf{K}=\begin{bmatrix}\mathbf{K}_4 & \mathbf{K}_2\\ \mathbf{K}_3 & \mathbf{K}_1\end{bmatrix}\tag{35}$$

and $\mathbf{E}$ is defined as

$$\mathbf{E}=\begin{bmatrix}\mathbf{K}_2\\ \mathbf{K}_3\end{bmatrix}.\tag{36}$$

Appendix F
Feature Extraction for CSKDA Using Two-Step Approach

Step 1) Calculate the eigenvalues and the eigenvectors of $\boldsymbol{\Phi}_s^T\boldsymbol{\Phi}_s$ and project each facial vector $\mathbf{z}_i\in\mathcal{U}$ as

$$\begin{aligned}
\boldsymbol{\Pi}_1^T\phi(\mathbf{z}_i)&=\left(\boldsymbol{\Pi}\boldsymbol{\Lambda}_{\mathbf{s}}{}^{-1/2}\right)^T\boldsymbol{\Phi}_s^T\phi(\mathbf{z}_i)\\
&=\left(\boldsymbol{\Pi}\boldsymbol{\Lambda}_{\mathbf{s}}{}^{-1/2}\right)^T[\tilde{\boldsymbol{\mu}}_1\ldots\tilde{\boldsymbol{\mu}}_{L_I}]^T\phi(\mathbf{z}_i)\\
&=(\boldsymbol{\Pi}\boldsymbol{\Lambda}_{\mathbf{s}}{}^{-1/2})^T([\phi(\mathbf{z}_1)\ldots\phi(\mathbf{z}_L)]^T\\
&\quad\times\phi(\mathbf{z}_i)-[\bar{\boldsymbol{\rho}}\ldots\bar{\boldsymbol{\rho}}]^T\phi(\mathbf{z}_i))\\
&=(\boldsymbol{\Pi}\boldsymbol{\Lambda}_{\mathbf{s}}{}^{-1/2})^T\left([\phi(\mathbf{z}_1)\ldots\phi(\mathbf{z}_L)]^T\right.\\
&\quad\times\phi(\mathbf{z}_i)-\frac{1}{L_G}\mathbf{1}_{LL_G}\\
&\quad\left.\times\left[\phi(\boldsymbol{\rho}_1)\ldots\phi\left(\boldsymbol{\rho}_{L_G}\right)\right]^T\phi(\mathbf{z}_i)\right)\\
&=\left(\boldsymbol{\Pi}\boldsymbol{\Lambda}_{\mathbf{s}}{}^{-1/2}\right)^T([k(\mathbf{z}_1,\mathbf{z}_i)\ldots k(\mathbf{z}_L,\mathbf{z}_i)]^T\\
&\quad-\frac{1}{L_G}\mathbf{1}_{LL_G}[k(\boldsymbol{\rho}_1,\mathbf{z}_i)\ldots k(\boldsymbol{\rho}_{L_G},\mathbf{z}_i)]).
\end{aligned}\tag{37}$$

Step 2) In the new space, calculate $\mathbf{W}$ and $\mathbf{B}$. Perform eigenanalysis to $\mathbf{B}$ and obtain a set of orthonormal eigenvectors. Create the two matrices $\boldsymbol{\Xi}_1=[\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_q]$ and $\boldsymbol{\Xi}_2=[\boldsymbol{\xi}_{q+1},\ldots,\boldsymbol{\xi}_n]$, where $q=\operatorname{rank}(\mathbf{B})$ that correspond to nonzero and zero eigenvalues, respectively.

Step 3) Calculate $\tilde{\mathbf{W}}=\boldsymbol{\Xi}_1^T\mathbf{W}\boldsymbol{\Xi}_1$, $\tilde{\mathbf{B}}=\boldsymbol{\Xi}_1^T\mathbf{B}\boldsymbol{\Xi}_1$ and find the regular discriminant features using the matrix $\tilde{\mathbf{J}}=[\tilde{\boldsymbol{\zeta}}_1\ldots\tilde{\boldsymbol{\zeta}}_q]$ whose columns are the eigenvectors of $\tilde{\mathbf{B}}^{-1}\tilde{\mathbf{W}}$ in descending order of the eigenvalues.

Step 4) Calculate $\hat{\mathbf{W}}=\boldsymbol{\Xi}_2^T\mathbf{W}\boldsymbol{\Xi}_2$ and find the irregular-discriminant projections using the matrix $\tilde{\mathbf{T}}=[\tilde{\boldsymbol{\tau}}_1\ldots\tilde{\boldsymbol{\tau}}_q]$ whose columns are the orthonormal eigenvectors of $\hat{\mathbf{W}}$.

After following these steps, the regular-discriminant projection for a test facial vector $\mathbf{y}$ is given by:

$$\begin{aligned}
\acute{\mathbf{y}}_1=\left(\boldsymbol{\Pi}\boldsymbol{\Lambda}_{\mathbf{s}}{}^{-1/2}\tilde{\mathbf{J}}\boldsymbol{\Xi}_1\right)^T&([k(\mathbf{z}_1,\mathbf{y})\ldots k(\mathbf{z}_L,\mathbf{y})]^T\\
&-\frac{1}{L_G}\mathbf{1}_{LL_G}[k(\boldsymbol{\rho}_1,\mathbf{y})\ldots k(\boldsymbol{\rho}_{L_G},\mathbf{y})])
\end{aligned}\tag{38}$$

the number of dimensions of the regular discriminant vectors is less than or equal to L_G-1. The irregular discriminant projection for the facial vector $\mathbf{y}$ is given by

$$\begin{aligned}
\acute{\mathbf{y}}_2=(\boldsymbol{\Pi}\boldsymbol{\Lambda}_{\mathbf{s}}{}^{-1/2}\tilde{\mathbf{T}}\boldsymbol{\Xi}_2)^T&([k(\mathbf{z}_1,\mathbf{y})\ldots k(\mathbf{z}_L,\mathbf{y})]^T\\
&-\frac{1}{L_G}\mathbf{1}_{LL_G}[k(\boldsymbol{\rho}_1,\mathbf{y})\ldots k(\boldsymbol{\rho}_{L_G}\mathbf{y})])
\end{aligned}\tag{39}$$

the number of dimensions of the feature vector $\acute{\mathbf{y}}_2$ is less than or equal to L_I-1.

Appendix G
Null Space of the Between-Class Scatter Matrix

As can be proven, the matrix $\mathbf{S}_b^\Phi$ (in the two-class case) [11] has only one eigenvector that corresponds to the non-null eigenvector. Diminish the null eigenvalues of $\mathbf{S}_b^\Phi$ by adding the scaled version of the identity matrix as

$$\mathbf{S}_b^\Phi\boldsymbol{\zeta}=0\Leftrightarrow\mathbf{S}_b^\Phi\boldsymbol{\zeta}+\sigma\boldsymbol{\zeta}=\boldsymbol{\zeta}\Leftrightarrow\left(\mathbf{S}_b^\Phi+\sigma\mathbf{I}\right)\boldsymbol{\zeta}=\sigma\boldsymbol{\zeta}\tag{40}$$

where $\sigma>0$. Thus, the eigenvectors of $\mathbf{S}_b^\Phi$ that correspond to null eigenvalues are the same ones that correspond to eigenvalues that are equal to σ for the matrix $\mathbf{S}_b^\Phi+\sigma\mathbf{I}$. The property of the projection to the null eigenvectors of $\mathbf{S}_b^\Phi$ that may indicate poor classification performance is given in the following proposition.

Proposition 2: If for some $\boldsymbol{\zeta}\in\mathcal{H}$, $\boldsymbol{\zeta}^T\mathbf{S}_b^\Phi\boldsymbol{\zeta}=0$, then under the projection $\boldsymbol{\zeta}$, for the two training mean vectors (genuine and impostor), $\mathbf{m}_G^\Phi,\mathbf{m}_I^\Phi$, it is valid that $\boldsymbol{\zeta}^T\mathbf{m}_G^\Phi=\boldsymbol{\zeta}^T\mathbf{m}_I^\Phi$. In other words, under the projection $\boldsymbol{\zeta}$, the two centers $\mathbf{m}_G^\Phi,\mathbf{m}_I^\Phi$ fall in the same point, which means that this projection does not help to separate the two classes (is not optimal in the sense of FLDA, where this projection makes the criterion equal to zero).

Proof: The between-class scatter matrix $\mathbf{S}_b^\Phi$ can be written as

$$\mathbf{S}_b^\Phi=N_GN_I\left(\mathbf{m}_G^\Phi-\mathbf{m}_I^\Phi\right)\left(\mathbf{m}_G^\Phi-\mathbf{m}_I^\Phi\right)^T.\tag{41}$$

Then, we have

$$\begin{aligned}
&\boldsymbol{\zeta}^T\mathbf{S}_b^\Phi\boldsymbol{\zeta}=0\Leftrightarrow\boldsymbol{\zeta}^T\left(\mathbf{m}_G^\Phi-\mathbf{m}_I^\Phi\right)\left(\mathbf{m}_G^\Phi-\mathbf{m}_I^\Phi\right)^T\boldsymbol{\zeta}=0\\
&\Leftrightarrow\left\|\left(\mathbf{m}_G^\Phi-\mathbf{m}_I^\Phi\right)^T\boldsymbol{\zeta}\right\|^2=0\Leftrightarrow\boldsymbol{\zeta}^T\mathbf{m}_G^\Phi=\boldsymbol{\zeta}^T\mathbf{m}_I^\Phi
\end{aligned}\tag{42}$$

and for the Fisher criterion $F(\zeta)=(\boldsymbol{\zeta}^T\mathbf{S}_b^\Phi\zeta)/(\boldsymbol{\zeta}^T\mathbf{S}_w^\Phi\zeta)=0$, where $\mathbf{S}_w^\Phi$ is the within-class scatter matrix. ■

References

[1] M. Kirby and L. Sirovich, "Application of the Karhunen-Loeve procedure for the characterization of human faces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 1, pp. 103–108, Jan. 1990.

[2] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[3] D. L. Swets and J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 831–836, Aug. 1996.

[4] 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.

[5] L. Chengjun and H. Wechsler, "Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, Apr. 2002.

[6] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, "Face recognition by independent component analysis," *IEEE Trans. Neural Netw.*, vol. 13, no. 6, pp. 1450–1464, Nov. 2002.

[7] L. Chengjun and H. Wechsler, "Independent component analysis of Gabor features for face recognition," *IEEE Trans. Neural Netw.*, vol. 14, no. 4, pp. 919–928, Jul. 2003.

[8] L. Juwei, K. N. Plataniotis, and A. N. Venetsanopoulos, "Face recognition using LDA-based algorithms," *IEEE Trans. Neural Netw.*, vol. 14, no. 1, pp. 195–200, Jan. 2003.

[9] 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.

[10] L. Chengjun, "Capitalize on dimensionality increasing techniques for improving face recognition grand challenge performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 725–737, May 2006.

[11] 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.

[12] A. Tefas, C. Kotropoulos, and I. Pitas, "Using support vector machines to enhance the performance of elastic graph matching for frontal face authentication," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 7, pp. 735–746, Jul. 2001.

[13] B. Due, S. Fischer, and J. Bigün, "Face authentication with Gabor information on deformable graphs," *IEEE Trans. Image Process.*, vol. 8, no. 4, pp. 504–516, Apr. 1999.

[14] C. Kotropoulos, A. Tefas, and I. Pitas, "Frontal face authentication using morphological elastic graph matching," *IEEE Trans. Image Process.*, vol. 9, no. 4, pp. 555–560, Apr. 2000.

[15] A. Tefas, C. Kotropoulos, and I. Pitas, "Face verification using elastic graph matching based on morphological signal decomposition," *Signal Process.*, vol. 82, no. 6, pp. 833–851, 2002.

[16] S. Zafeiriou, A. Tefas, and I. Pitas, "Exploiting discriminant information in elastic graph matching," in *Proc. IEEE Int. Conf. Image Processing*, Genova, Italy, Sep. 11–14, 2005, vol. 3, pp. 768–771.

[17] S. Zafeiriou, A. Tefas, and I. Pitas, "Learning discriminant person-specific facial models using expandable graphs," *IEEE Trans. Inf. Forensics Security*, vol. 2, no. 1, pp. 55–68, Mar. 2007.

[18] 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. 17, no. 3, pp. 683–695, May 2006.

[19] Y. P. Kittler, J. Li, and J. Matas, J. T. Kent and R. G. Aykroyd, Eds., "Face verification using client specific Fisher faces," *Stat. Directions, Shapes Images*, pp. 63–66, 2000.

[20] K.-A. Toh, J. Xudong, and W.-Y. Yau, "Exploiting global and local decisions for multimodal biometrics verification," *IEEE Trans. Signal Process.*, vol. 52, no. 10, pt. 2, pp. 3059–3072, Oct. 2004.

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

[22] A. M. Martinez and R. Benavente, The AR face database Available information in [Online]. Available: http://rvll.ecn.purdue.edU//aleixaleix_face_DB.html.

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

[24] ORL database of faces, AT&T Laboratories Cambridge [Online]. Available: http://www.camorl.co.uk/facedatabase.html.

[25] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.

[26] UMIST Face Database [Online]. Available: http://images.ee.umist.ac.uk/danny/database.html.

[27] K.-R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, "An introduction to kernel-based learning algorithms," *IEEE Trans. Neural Netw.*, vol. 12, no. 2, pp. 181–201, Mar. 2001.

[28] B. Scholkopf, S. Mika, C. J. C. Burges, P. Knirsch, K.-R. Muller, G. Ratsch, and A. J. Smola, "Input space vs. feature space in kernel-based methods," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 1000–1017, Sep. 1999.

[29] A. Scholkopf, B. Smola, and K. R. Muller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, pp. 1299–1319, 1998.

[30] C. Kotropoulos, A. Tefas, and I. Pitas, "Frontal face authentication using discriminating grids with morphological feature vectors," *IEEE Trans. Multimedia*, vol. 2, no. 1, pp. 14–26, Mar. 2000.

[31] 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.

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

[33] B. Due, S. Fischer, and J. Bigun, "Face authentication with sparse grid gabor information," in *Proc. ICASSP*, Munich, Germany, 1997, pp. 3053–3056.

[34] A. Tefas, C. Kotropoulos, and I. Pitas, "Variants of dynamic link architecture based on mathematical morphology for frontal face authentication," in *Proc. CVPR*, Santa Barbara, CA, 1998, vol. 1, pp. 814–819.

[35] S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas, "Class-specific discriminant non-negative matrix factorization for frontal face verification," presented at the ICAPR, Bath, U.K., Aug. 22–25, 2005.

[36] X. S. Zhou and T. S. Huang, "Small sample learning during multimedia retrieval using biasmap," in *Proc. CVPR*, 2001, vol. 1, pp. 11–17.

[37] L. F. Chen, H. Y. M. Liao, M. T. Ko, J. C. Lin, and G. J. Yu, "A new LDA-based face recognitions system which can solve the small sample size problem," *Pattern Recognit.*, vol. 33, pp. 1713–1726, 2000.

[38] H. Yu and J. Yang, "A direct LDA algorithm for high-dimensional data with application to face recognition," *Pattern Recognit.*, vol. 34, pp. 2067–2070, 2001.

[39] E. Pekalska, P. Paclik, and R. P. W. Duin, "A generalized kernel approach to dissimilarity-based classification," *J. Mach. Learning Res.*, vol. 2, pp. 175–211, 2001.

[40] B. Haasdonk, "Feature space interpretation of SVMs with indefinite kernels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 4, pp. 482–492, Apr. 2005.

[41] L. Juwei, K. N. Plataniotis, and A. N. Venetsanopoulos, "Regularization studies of linear discriminant analysis in small sample size scenarios with application! to face recognition," *Pattern Recognit. Lett.*, vol. 26, no. 2, pp. 181–191, 2005.

[42] L. Juwei, K. N. Plataniotis, A. N. Venetsanopoulos, and J. Wang, "An efficient kernel discriminant analysis method," *Pattern Recognit.*, vol. 38, pp. 1788–1790, Oct. 2005.

[43] M. J. Er, S. Wu, J. Lu, and H. L. Toh, "Face recognition with radial basis function (RBF) neural networks," *IEEE Trans. Neural Netw.*, vol. 13, no. 3, pp. 697–710, May 2002.

[44] M. J. Er, W. Chen, and S. Wu, "High-speed face recognition based on discrete cosine transform and RBF neural networks," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 679–691, May 2005.

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

[46] Q. Liu, X. Tang, H. Lu, and S. Ma, "Face recognition using kernel scatter-difference-based discriminant analysis," *IEEE Trans. Neural Netw.*, vol. 17, no. 4, pp. 1081–1085, Jul. 2006.

[47] H. Cevikalp, M. Neamtu, and M. Wilkes, "Discriminative common vector method with kernels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 6, pp. 1550–1565, Nov. 2006.

[48] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, A. Smola, and K.-R. Muller, "Constructing descriptive and discriminative nonlinear features: Rayleigh coefficients in kernel feature spaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 5, pp. 623–628, May 2003.

[49] S. Mika, G. Ratsch, B. Scholkopf, A. Smola, J. Weston, and K.-R. Muller, "Invariant feature extraction and classification in kernel spaces," *Adv. Neural Inf. Process. Syst.*, vol. 12, pp. 1299–1319, 1999.

[50] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Muller, "Fisher discriminant analysis with kernels," in *Proc. IEEE Int. Workshop Neural Networks for Signal Processing IX*, 1999, pp. 41–48.

[51] S. K. Zhou and R. Chellappa, "From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel hilbert space," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 6, pp. 917–929, Jun. 2006.

[52] L. Chengjun and H. Wechsler, "Evolutionary pursuit and its application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 6, pp. 570–582, Jun. 2000.

[53] P. J. Phillips, "Matching pursuit filters applied to face identification," *IEEE Trans. Image Process.*, vol. 7, no. 8, pp. 1150–1164, Aug. 1998.

[54] B.-L. Zhang, H. Zhang, and S. Sam Ge, "Face recognition by applying wavelet subband representation and kernel associative memory," *IEEE Trans. Neural Netw.*, vol. 15, no. 1, pp. 166–177, Jan. 2004.

[55] C. Kotropoulos, A. Tefas, and I. Pitas, "Morphological elastic graph matching applied to frontal face authentication under well-controlled and real conditions," *Pattern Recognit.*, vol. 33, no. 12, pp. 31–43, Oct. 2000.

[56] S. Pigeon and L. Vandendorpe, J. Bigun, G. Chollet, and G. Borgefors, Eds., "The M2VTS multimodal face database," *Lecture Notes in Computer Science: Audioand Video—Based Biometric Person Authentication*, vol. 1206, pp. 403–409, 1997.

[57] P. J. Phillips, H. Wechsler, J. S. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image Vis. Comput.*, vol. 16, no. 5, pp. 295–306, 1998.

[58] I. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, pp. 153–157, 1947.

[59] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. ICASSP*, Glasgow, U.K., 1989, pp. 532–535.

[60] W. Yambor, B. Draper, and R. Beveridge, "Analyzing pea-based face recognition algorithms: Eigenvector selection and distance measures," in *Empirical Evaluation Methods in Computer Vision*, H. Christensen and J. Phillips, Eds. Singapore: World Scientific, 2002.

[61] B. A. Draper, K. Baek, M. S. Bartlett, and J. R. Beveridge, "Recognizing faces with pea and ica," *Comput. Vis. Image Understanding*, vol. 91, no. 1–2, pp. 115–137, 2003.

[62] V. Vapnik*, The Nature of Statistical Learning Theory*. New York: Springer Verlag, 1995.

[63] K. I. Kim, K. Jung, and H. J. Kim, "Face recognition using kernel principal component analysis," *IEEE Signal Process. Lett.*, vol. 9, no. 2, pp. 40–42, Feb. 2002.

Georgios Goudelis received the B.Eng. degree in electronic engineering with medical electronics and the M.Sc. degree in electronic engineering from the University of Kent, Canterbury (UKC), U.K., in 2003 and 2004, respectively. He is currently pursing the Ph.D. degree at the Artificial Intelligence and Information Analysis Lab of the Department of Informatics at the Aristotle University of Thessaloniki, Thessaloniki, Greece.

From 2003 to 2004, he was a Part-Time Teacher at UKC. His research interests include digital signal processing, computer vision, pattern recognition, and object tracking.

Stefanos Zafeiriou was born in Thessaloniki, Greece, in 1981. He received the B.Sc. (Hons.) and Ph.D. degrees in informatics from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2003 and 2007, respectively.

Currently, he is a Researcher and Teaching Assistant in the Department of Informatics at the Aristotle University of Thessaloniki. His current research interests are in the areas of signal and image processing, computational intelligence, pattern recognition, and computer vision, as well as in the area of watermarking for copyright protection and authentication of digital media. He has coauthored many journal and conference publications.

Dr. Zafeiriou has received various scholarships and awards during his undergraduate and Ph.D. studies.

Anastasios Tefas (M'04) received the B.Sc. degree in informatics and the Ph.D. degree in informatics from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 1997 and 2002, respectively.

Currently, he is an Assistant Professor in the Department of Information Management, Technological Educational Institute of Kavala, Kavala, Greece. From 1997 to 2002, he was a Researcher and Teaching Assistant in the Department of Informatics, University of Thessaloniki. From 2003 to 2004, he was a Temporary Lecturer in the Department of Informatics, University of Thessaloniki, where he is a Senior Researcher. He has coauthored many journal and conference papers. His research interests include computational intelligence, pattern recognition, digital signal and image processing, detection and estimation theory, and computer vision.

Ioannis Pitas (SM'94–F'07) received the D.Eng. and Ph.D. degrees in electrical engineering from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 1980 and 1985, respectively.

Currently, he is a Professor at the Department of Informatics, Aristotle University of Thessaloniki. From 1980 to 1993, he was Scientific Assistant, Lecturer, Assistant Professor, and Associate Professor in the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki. He was a Visiting Research Associate at the University of Toronto, Toronto, ON, Canada; University of Erlangen, Nuernberg, Germany; Tampere University of Technology, Tampere, Finland; and Visiting Assistant Professor at the University of Toronto and as Visiting Professor at the University of British Columbia, Vancouver, BC, Canada. He was a Lecturer in short courses for continuing education. He has published many journal and conference papers and contributed to many books in his areas of interest. He is the co-author of the books "*Nonlinear Digital Filters: Principles and Applications*" (Kluwer, 1990), "*3-D Image Processing Algorithms*" (Wiley, 2000), *Nonlinear Model-Based Image/Video Processing and Analysis* (Wiley, 2001), and author of "*Digital Image Processing Algorithms and Applications*" (Wiley, 2000). He is the editor of the book "*Parallel Algorithms and Architectures for Digital Image Processing, Computer Vision and Neural Networks*" (Wiley, 1993). His current interests are in the areas of digital image and video processing and analysis, multidimensional signal processing, watermarking, and computer vision.

Dr. Pitas has also been an invited speaker and/or member of the program committee of several scientific conferences and workshops. In the past, he was Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS, IEEE TRANSACTIONS ON NEURAL NETWORKS, IEEE TRANSACTIONS ON IMAGE PROCESSING, *EURASIP Journal on Applied Signal Processing*, and co-editor of *Multidimensional Systems and Signal Processing*. He was General Chair of the 1995 IEEE Workshop on Nonlinear Signal and Image Processing (NSIP95), Technical Chair of the 1998 European Signal Processing Conference, and General Chair of IEEE ICIP 2001.