DISCRIMINATIVE COMMON VECTOR METHOD WITH KERNELS

Hakan Cevikalp[1], Marian Neamtu[2], and Mitch Wilkes[1]

[1]Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, Tennessee, USA.

[2]Center for Constructive Approximation, Department of Mathematics, Vanderbilt University, Nashville, Tennessee, USA.

CORRESPONDENCE ADDRESS:

Prof. Mitch Wilkes

Department of Electrical Engineering and Computer Science,

Vanderbilt University, Nashville, Tennessee, USA

Tel: (615) 343-6016

Fax: (615) 322-7062

e-mail: mitch.wilkes@vanderbilt.edu

Abstract

In some pattern recognition tasks, the dimension of the sample space is larger than the number of the samples in the training set. This is known as the "small sample size problem". The Linear Discriminant Analysis (LDA) techniques cannot be applied directly to the small sample size case. The small sample size problem is also encountered when kernel approaches are used for recognition. In this paper we try to answer the question of "How should we choose the optimal projection vectors for feature extraction for the small sample size case?" Then, we propose a new method called the Kernel Discriminative Common Vector (Kernel DCV) method, based on our findings. In this method, we first nonlinearly map the original input space to an implicit higher-dimensional feature space through a kernel mapping, where the data are hoped to be linearly separable. Then, the optimal projection vectors are computed in the transformed space. The proposed method yields an optimal solution for maximizing the modified Fisher's Linear Discriminant criterion given in the paper. Thus, a 100% recognition rate is always guaranteed for the training set samples. Experiments on test data sets also show that the generalization ability of the proposed method outperforms other kernel approaches in many situations.

Index Terms: Discriminative common vectors, feature extraction, Fisher's linear discriminant analysis, kernel methods, principal component analysis, small sample size, subspace methods.

I. INTRODUCTION

Feature extraction has been one of the most important issues in pattern recognition. In the problem of feature extraction, the aim is to select the variables that contain the most discriminatory information. Most of the feature extraction methods have centered on finding linear transformations that map the original high-dimensional sample space into a lower-dimensional space, which hopefully contains all the necessary discriminatory information. The principal motivation behind dimensionality reduction is that it may reduce the worst effects of the curse of dimensionality [1]. Often, improved performance is achieved over the application of the selected classifier in the original sample space. Also, linear feature extraction techniques are often used as pre-processors for more complex nonlinear classifiers. However, sometimes linear methods may not provide sufficient nonlinear discriminant power for classification of linearly non-separable classes (e.g., exclusive-or problem). Thus, kernel approaches have been proposed to overcome this limitation. The main idea is to transform the input data into a higher-dimensional space by a nonlinear kernel mapping and then apply the linear discriminant techniques in this space. The motivation behind this is to transform the linearly non-separable data into a higher-dimensional space where the data are linearly separable. Therefore, it turns out that a nonlinear discriminant method is applied in the original sample space.

One of the most popular feature extraction methods is the Principal Component Analysis (PCA) method. In this method we find the best set of projection directions in the sample space that will maximize the total scatter across all samples such that the criterion $J_{PCA}(W_{opt}) = \arg\max |W^T S_T W|$ is maximized. Here W is the matrix whose columns are the projection vectors and S_T is the total scatter matrix of the training set samples. This criterion is maximized when the most significant eigenvectors (the eigenvectors corresponding to the largest

eigenvalues of S_T) are chosen as the projection vectors. The PCA method is an unsupervised method since it does not consider the classes of the training set data. Although it is useful for reconstruction, it is not necessarily optimal from a discrimination point of view. Thus, the projection vectors chosen for optimal reconstruction may obscure the existence of separate classes [1], [2]. The Fisher's Linear Discriminant Analysis (FLDA) method was proposed to overcome the limitations of the PCA method [3]. It has been successfully applied in many classification problems such as image recognition, multimedia information retrieval, and medical applications. The method uses the FLDA criterion, which tries to maximize the ratio, $J_{FLDA}(W_{opt}) = \arg\max \frac{|W^T S_B W|}{|W^T S_W W|}$, where S_W is the within-class scatter matrix, and S_B is the between-class scatter matrix. The above criterion is maximized when the eigenvectors of $S_W{}^{-1} S_B$ are employed as projection vectors. Since the matrix $S_W{}^{-1} S_B$ is typically not symmetric, the eigen-decomposition may be unstable. To avoid this problem, the simultaneous diagonalization algorithm is often employed to obtain a stable eigen-decomposition [4], [5]. The major drawback of the FLDA method is that it cannot be applied directly if the rank of the within-class scatter matrix S_W is smaller than the dimension of the sample space since S_W is singular in this case. This problem is also known as the "*small sample size problem*" [4]. The Perturbation method has been used in [6] and [7], where S_W is perturbed so as to become nonsingular. Swets and Weng [5] proposed a two-stage PCA+FLDA method, also known as the Fisherface method, in which PCA is first used for dimension reduction so as to make S_W nonsingular before the application of the FLDA.

Recently, Yu and Yang have proposed the Direct-LDA method to overcome the small sample size problem [8]. This method also uses simultaneous diagonalization for finding the projection

vectors in the range space of S_B. First, the null space of S_B is discarded, and then the projection vectors that minimize the within-class scatter in the transformed space are selected from the range space of S_B. However, the range space of S_B does not necessarily include the optimal projection vectors [9], [10], [11]. This fact can be clearly seen in the example that is given in Fig. 1. In Fig. 1, we plotted two linearly separable classes with similar covariance matrices. The class means are shown as stars. Since the class distributions are similar, we expect the decision regions produced by the linear discriminant analysis techniques to be optimal in this case. As can be seen in the figure, although the FLDA method correctly discriminates all samples, the Direct-LDA method fails for this example. Note that, the FLDA and the Direct-LDA methods produce the same results if the ranks of both the between-class scatter and the within-class scatter matrices are larger than or equal to the dimensionality of the sample space, or the within-class scatter of the samples is isotropic. These conditions are not typically satisfied for the small sample size case. Therefore, the Direct-LDA method fails to extract optimal projection vectors for feature extraction in most cases.

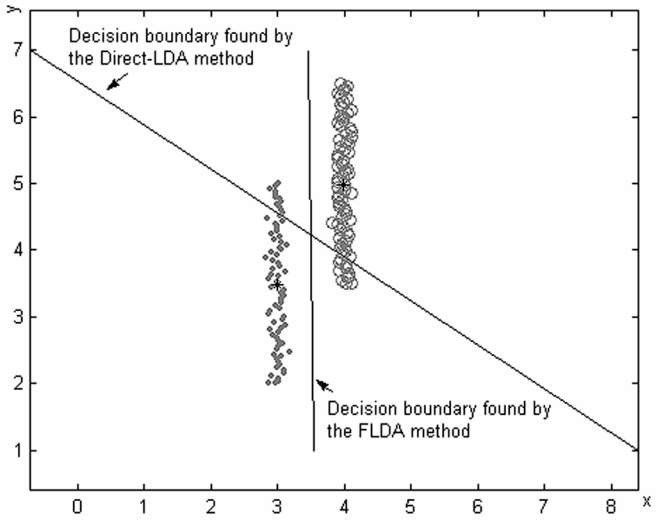

Fig. 1. Two linearly separable classes with the similar covariance matrices are plotted. Stars represent class means and lines represent the decision boundaries found by the Direct-LDA and the FLDA methods.

Chen *et al*. proposed the Null Space method for the small sample size case based on the modified FLDA criterion, $J_{MFLDA}(W_{opt}) = \arg\max_{W} \frac{\left|W^T S_B W\right|}{\left|W^T S_T W\right|}$ [12]. In this method, all training samples are first projected onto the null space of S_W, resulting in a new within-class scatter that is a zero matrix. Then, PCA is applied to the transformed samples to obtain the final projection vectors. Chen *et al*. also proved that by applying this method, the modified FLDA criterion attains its maximum of 1; therefore the Null Space method extracts features, which are optimal from a discrimination point of view. It turns out that the orthonormal projection vectors obtained by the Null Space method span the space, which is the intersection of the null space of S_W and the range space of S_T. We call it the *optimal discriminant subspace* since it is spanned by vectors that extract the optimal features for discrimination. However, Chen *et al*. did not give an efficient algorithm for applying this method in the original sample space. Instead, a pixel grouping method was applied to extract geometric features and reduce the dimension of the sample space. Then, they applied the Null Space method in this new reduced space. Vapnik suggests that when solving a given problem, one should avoid solving a more general problem as an intermediate step [13]. Following this suggestion we showed that any pre-processing step, such as a pixel grouping method that reduces the original dimension of the null space, is likely to reduce the performance and therefore should be avoided [14]. In [14], we proposed the Discriminant Common Vector (DCV) method to find optimal orthonormal projection vectors in the optimal discriminant subspace. This method is equivalent to the Null Space method, with the exception that the pixel reducing step is omitted and therefore the method exploits the original high-dimensional space. Two efficient algorithms were given to compute the optimal projection vectors. One algorithm uses the range space of S_W, while the other uses subspace methods and

the Gram-Schmidt orthogonalization procedure. Another novel method, the PCA+Null Space method was proposed by Huang *et al*. in [10] to find optimal projection vectors that span the optimal discriminant subspace. In this method, PCA is first applied to remove the null space of S_T. Then, optimal projection vectors are found in the remaining lower-dimensional space by using the Null Space method. However, this method is computationally expensive compared to the DCV method (see [14], for a comparison of these methods).

The Kernel PCA method was proposed as a nonlinear extension of PCA [15]. The basic idea is first to transform the data into a higher-dimensional space via a nonlinear mapping and then apply the linear PCA method in this space. The projection onto this higher-dimensional space and the application of PCA in the transformed space are performed by using a set of kernel functions without explicitly working in the transformed space, making this process computationally feasible. However, the Kernel PCA method is also an unsupervised technique in that it extracts features that may not be optimal from the discrimination point of view. Therefore, discriminant analysis techniques that use kernels have been recently proposed [16], [17]. Similarly to the Kernel PCA, these methods also use kernel functions to project data into a higher-dimensional space via a nonlinear kernel mapping, and then the Linear Discriminant Analysis (LDA) is performed in this higher-dimensional space. However, the singularity problem of the matrices is encountered in these techniques. Two different approaches are adopted to solve this problem. Mika *et al*. use a perturbation method in which a small perturbation matrix is added to make singular matrices nonsingular [16]. Baudat and Anouar use the modified FLDA criterion instead of the original FLDA criterion [17]. They first project the data onto the range space of the total scatter matrix of mapped samples through the Kernel PCA, and then they apply the LDA method that maximizes the modified FLDA criterion in the reduced

space [18]. The first approach is called the Kernel Fisher's Discriminant Analysis (Kernel FDA) method and latter approach is called the Kernel Generalized Discriminant Analysis (Kernel GDA) method.

In this paper we propose a new method called the Kernel DCV method, which applies the DCV method in the nonlinearly transformed higher-dimensional space. Since the modified FLDA criterion is guaranteed to attain its maximum value when using the Kenel DCV method, just as in the DCV method, the optimal features for discrimination are extracted from the nonlinearly transformed higher-dimensional space.

The remainder of this paper is organized as follows. In Section II, we describe the optimal discriminant subspace concept in detail, and then we show how to extract the optimal projection directions from this subspace. In Section III, the Kernel DCV method is introduced. In Section IV, we describe the data sets and experimental results. Finally, our conclusions are presented in Section V.

II. OPTIMAL PROJECTION VECTORS

The modified FLDA criterion tries to maximize the ratio, $J_{MFLDA}(W_{opt}) = \arg\max_{W} \frac{\left|W^T S_B W\right|}{\left|W^T S_T W\right|}$. However, this criterion is not appropriate since its maximum is not unique for the small sample size case. In particular, every projection vector matrix W such that $W^T S_W W = 0$ and $W^T S_B W \neq 0$ maximizes the modified FLDA criterion. Note that if S_W is singular, which is always the case for the small sample size problem, there are many such projection vector matrices *W*. However, it is not reasonable to use matrices *W* with a small number of projection vectors since they are not sufficient for an optimal feature extraction. On the other hand, the

following criterion given in [19] has a unique maximum and it also maximizes the modified FLDA criterion

$$J(W_{opt}) = \arg\max_{|W^T S_W W|=0} |W^T S_B W| = \arg\max_{|W^T S_W W|=0} |W^T S_T W|. \quad (1)$$

Therefore, to find the optimal projection vectors *w* in the null space $N(S_W)$ of S_W, we project the training set samples onto $N(S_W)$ and then obtain the projection vectors by performing PCA. From this operation we get a set of orthonormal vectors that is a basis for a space, which we call the optimal discriminant subspace. The optimal discriminant subspace is the intersection of $N(S_W)$ and the range space $R(S_T)$ of the total scatter matrix S_T. The criterion given in (1) attains its maximum for orthonormal vectors that form a basis for the optimal discriminant subspace. There are numerous algorithms to find this optimal subspace and an orthonormal basis for it. Some efficient algorithms are given in [14].

A. The Optimal Discriminant Subspace Concept

Let the training set be composed of *C* classes, where the *i*-th class contains N_i samples, and let x_m^i be a *d*-dimensional column vector which denotes the *m*-th sample from the *i*-th class. There will be a total of $M = \sum_{i=1}^{C} N_i$ samples in the training set. Suppose that *d>M-C.* In this case, the within-class scatter matrix S_W, the between-class scatter matrix S_B, and the total scatter matrix S_T are defined as

$$S_W = \sum_{i=1}^{C} \sum_{m=1}^{N_i} (x_m^i - \mu_i)(x_m^i - \mu_i)^T, \quad (2)$$

$$S_B = \sum_{i=1}^{C} N_i (\mu_i - \mu)(\mu_i - \mu)^T, \quad (3)$$

and

$$S_T = \sum_{i=1}^{C} \sum_{m=1}^{N_i} (x_m^i - \mu)(x_m^i - \mu)^T = S_W + S_B, \quad (4)$$

where μ is the mean of all samples, and μ_i is the mean of samples in the *i*-th class.

If the dimensionality *d* of the sample space is larger than *M*-1, all scatter matrices will be rank deficient. Thus, if we apply eigen-decomposition to the scatter matrices, we will obtain some eigenvectors corresponding to the zero eigenvalues that span the null spaces of the scatter matrices. As explained previously, if the projection directions are chosen from $N(S_W)$, the modified FLDA criterion attains its maximum, 1. Therefore, we have to project the training set data onto $N(S_W)$. Then, optimal projection vectors can be obtained by applying PCA to the samples, which are projected onto $N(S_W)$. The fact that the optimal projection vectors span the optimal discriminant subspace follows from the following lemma.

Lemma 1: Suppose $\overline{U}$ is a matrix whose column vectors $u_k (k = r_T + 1, ..., d$, where r_T is the rank of S_T) are orthonormal vectors that span the null space $N(S_T)$ of S_T. If all samples in the training set are projected onto $N(S_T)$, they produce a unique common vector such that

$$x = \overline{U}\,\overline{U}^T x_m^i, \quad i = 1, ..., C, \quad m = 1, ..., N_i, \quad (5)$$

where *x* is independent of indices *i* and *m*.

Proof: By definition, a vector $u \in R^d$ is in $N(S_T)$ if $S_T u = 0$. Let μ be the mean vector of the samples in the training set, $1_M \in R^{MxM}$ be the matrix with all elements equal to M^{-1}, and $X \in R^{dxM}$ be the matrix whose columns are the training set samples. Thus, by multiplying both sides of identity $S_T u = 0$ by u^T, we get

$$0 = \sum_{i=1}^{C} \sum_{m=1}^{N_i} u^T (x_m^i - \mu)(x_m^i - \mu)^T u = u^T X (I - 1_M)(I - 1_M)^T X^T u = \| (I - 1_M) X^T u \|^2, \quad (6)$$

where ||.|| denotes the Euclidean norm. Thus, (6) holds if $(I - 1_M)X^T u_k = 0$, or $X^T u_k = 1_M X^T u_k$. From this relation it can be seen that

$$(x_m^i)^T u_k = \mu^T u_k, \qquad i = 1,...,C, \quad m = 1,...,N_i, \quad k = r_T + 1,...,d. \tag{7}$$

Thus the projection of any x_m^i onto $N(S_T)$,

$$x = \sum_{k=r_T+1}^{d} \langle x_m^i, u_k \rangle u_k = \sum_{k=r_T+1}^{d} \langle \mu, u_k \rangle u_k, \quad i = 1,...,C, \quad m = 1,...,N_i, \tag{8}$$

is independent of *m* and *i*, which proves the lemma. □

This lemma shows that, $N(S_T)$ does not contain any discriminative information, which can be used in the course of obtaining the optimal projection vectors. Therefore this null space can be removed. Then, the remaining subspace for extracting the features of discrimination will be the intersection of $N(S_W)$ and $R(S_T)$.

There are numerous algorithms to find the optimal discriminant subspace and optimal projection vectors that span it. The following observation proposed by Therrien [20] can be used to find optimal projection vectors and the optimal discriminant subspace.

Observation 1: Let $H^{(i)}$ be a subspace of R^d. A vector *e* is contained in $\bigcap_{i=1}^{n} H^{(i)}$ if and only if it is an eigenvector of Ψ corresponding to an eigenvalue of 1, where

$$\Psi = \sum_{i=1}^{n} a_i P^{(i)} \tag{9}$$

with $P^{(i)}$ being the projection matrix (also called the orthogonal projection operator) of the *i*-th subspace and $0 < a_i < 1, \sum_{i=1}^{n} a_i = 1$.

In our case we can choose $H^{(1)}$ and $H^{(2)}$ as $R(S_T)$ and $N(S_W)$, respectively, to find orthonormal vectors that span the optimal discriminant space. However, this approach is not always practical for real applications since the size of projection matrices of subspaces may be too large (e.g., images of size 256 by 256 yield projection matrices of size 65,536 by 65,536). We will use this observation for the numerical example given at the end of this section.

There are better ways to find the optimal projection vectors. This is a result of the fact that the projection matrices of $N(S_W)$ and $R(S_T)$ commute, as shown in Theorem 1 below, namely $P^{(1)}P^{(2)} = P^{(2)}P^{(1)}$, where $P^{(1)}$ and $P^{(2)}$ represent the projection matrices of $R(S_T)$ and $N(S_W)$, respectively. In this case, the projection matrix of the intersection $N(S_W) \cap R(S_T)$ is found by the equation

$$P_{opt} = P^{(1)}P^{(2)} = P^{(2)}P^{(1)}, \tag{10}$$

where P_{opt} is the projection matrix of the optimal discriminant subspace [21]. A consequence of Theorem 1 is that to obtain the optimal projection vectors we can first project the training set samples onto $N(S_W)$ and then apply PCA or, alternatively, we can first project the training set samples onto $R(S_T)$ through PCA, and then find the null space in the transformed space. The DCV method uses the first approach, whereas the PCA+Null Space method uses the second approach.

Before we prove Theorem 1, we need the following auxiliary lemmas.

Lemma 2: The following holds:

i) $N(S_T) \subset N(S_W)$,

ii) $(N(S_T) + R(S_W)) \cap N(S_W) = N(S_T) \cap N(S_W)$,

where "+" denotes the usual sum of sets.

Proof: i) Let $v \in N(S_T)$. Then

$$S_T v = (S_B + S_W)v = 0. \tag{11}$$

Since the scatter matrices are positive semi-definite, this implies [14]

$$S_B v = S_W v = 0, \tag{12}$$

and hence $v \in N(S_W)$.

ii) To show that the equation $(N(S_T) + R(S_W)) \cap N(S_W) = N(S_T) \cap N(S_W)$ holds, we have to show that $(N(S_T) + R(S_W)) \cap N(S_W) \supset N(S_T) \cap N(S_W)$ and $(N(S_T) + R(S_W)) \cap N(S_W) \subset N(S_T) \cap N(S_W)$. It is trivial to see that the first inclusion holds. As for the second inclusion, let $v \in (N(S_T) + R(S_W)) \cap N(S_W)$. The vector v can be decomposed as $v = a + b$, where $a \in N(S_T)$ and $b \in R(S_W)$. Then since $v \in N(S_W)$,

$$S_W (a + b) = 0. \tag{13}$$

Moreover, as noted above, $S_T a = 0$ implies $S_W a = 0$. These facts imply $S_W b = 0$. On the other hand, since $b \in R(S_W)$,

$$S_W b = 0 \Leftrightarrow b = 0. \tag{14}$$

Thus, $v = a \in N(S_T)$ and this verifies the considered inclusion. □

Lemma 3: Let $H^{(1)}$ and $H^{(2)}$ be subspaces of R^d. Let $P^{(1)}$ and $P^{(2)}$ be the orthogonal projections onto $H^{(1)}$ and $H^{(2)}$, respectively. Let $H = H^{(1)} \cap H^{(2)}$ and $H'^{(2)}$ be defined by

$$H'^{(2)} \oplus H = H^{(2)}, \tag{15}$$

i.e., $H'^{(2)}$ is the orthogonal complement of H in $H^{(2)}$. Then $P^{(1)}$ and $P^{(2)}$ commute, that is $P^{(1)}P^{(2)} = P^{(2)}P^{(1)}$, if and only if $H'^{(2)} \perp H^{(1)}$.

Proof: Let $H'^{(1)}$ be the orthogonal complement of H in $H^{(1)}$, i.e., $H'^{(1)} \oplus H = H^{(1)}$. Suppose first that $H'^{(2)} \perp H^{(1)}$. Thus, $H^{(1)} + H^{(2)} = H \oplus H'^{(1)} \oplus H'^{(2)}$ and each $v \in R^d$ has a unique decomposition $v = a + b + c + d$, where $a \in H$, $b \in H'^{(1)}$, $c \in H'^{(2)}$, and $d \in (H^{(1)} + H^{(2)})^{\perp}$. Clearly,

$$P^{(2)}P^{(1)}v = P^{(2)}(a+b) = a = P^{(1)}(a+c) = P^{(1)}P^{(2)}v. \tag{16}$$

Conversely, if $P^{(1)}$ and $P^{(2)}$ commute, we have

$$P^{(1)}H'^{(2)} = P^{(1)}P^{(2)}H'^{(2)} = P^{(2)}P^{(1)}H'^{(2)}, \tag{17}$$

and hence, $P^{(1)}H'^{(2)} \subset H^{(2)}$ since $P^{(2)}$ is the identity operator on $H^{(2)}$. On the other hand, by definition we also have $P^{(1)}H'^{(2)} \subset H^{(1)}$. As a result, $P^{(1)}H'^{(2)} \subset H^{(1)} \cap H^{(2)} = H$. Finally, one can have $P^{(1)}H'^{(2)} \subset H$ only if $P^{(1)}H'^{(2)} = \{0\}$ since $H'^{(2)} \perp H$. However this implies $H'^{(2)} \perp H^{(1)}$. □

We are now ready to prove the following theorem:

Theorem 1: Let $H^{(1)} = R(S_T)$, $H^{(2)} = N(S_W)$, and $H = H^{(1)} \cap H^{(2)} = R(S_T) \cap N(S_W)$, the optimal discriminant subspace. Then, the projection matrices $P^{(1)}$ and $P^{(2)}$ of the subspaces $H^{(1)}$ and $H^{(2)}$ commute.

Proof: Using the notation from Lemma 3, we know that $H'^{(2)} \perp H^{(1)} \Rightarrow P^{(1)}P^{(2)} = P^{(2)}P^{(1)}$. The orthogonal complement of $H = R(S_T) \cap N(S_W)$ in R^d is

$$H^{\perp} = (R(S_T) \cap N(S_W))^{\perp} = N(S_T) + R(S_W). \tag{18}$$

Thus, $H'^{(2)}$ equals,

$$H'^{(2)} = (N(S_T) + R(S_W)) \cap N(S_W). \tag{19}$$

From Lemma 2 we know that,

$$H'^{(2)} = (N(S_T) + R(S_W)) \cap N(S_W) = N(S_T) \cap N(S_W) = N(S_T). \quad (20)$$

Thus, it is clear that $H_2^{'} \perp H_1$ since this is equivalent to $N(S_T) \perp R(S_T)$. Invoking Lemma 3 now finishes the proof. □

In [22], [23], and [24], the authors claim that the Direct-LDA method finds the projection vectors in the intersection space of $N(S_W)$ and $R(S_B)$, thus the projection vectors found by this method should be optimal and equivalent to the ones found by the Null Space method (equivalently the DCV method and the PCA+Null Space method). However this statement is incorrect. In fact, neither the Direct-LDA method nor the Null Space method finds the projection vectors in the intersection space of $R(S_B)$ and $N(S_W)$. The projection directions obtained by the Direct-LDA method come from $R(S_B)$, and the intersection of $R(S_B)$ and $N(S_W)$ is in fact often trivial. Indeed, in all the face database examples considered in this paper, the intersection was trivial. Therefore, the intersection space of $R(S_B)$ and $N(S_W)$ cannot be used for recognition. This fact is illustrated in Fig. 2. In Fig. 2, two classes with the same covariance matrices having two samples each in a 3-dimensional space are plotted. Class means are represented with circles. $R(S_W)$ and $R(S_B)$ are shown in the figure. In this example $R(S_T)$ is the plane spanned by the vectors representing $R(S_W)$ and $R(S_B)$, and $N(S_T)$ is the line perpendicular to this plane. Note that it is also the intersection of $N(S_B)$ and $N(S_W)$. The optimal discriminant subspace, $R(S_T) \cap N(S_W)$, is the line in this plane that is perpendicular to $R(S_W)$. $N(S_W)$ is the plane spanned by the vectors representing $N(S_T)$ and $R(S_T) \cap N(S_W)$. As can be seen in the figure, the intersection of $N(S_W)$ and $R(S_B)$ is the trivial space.

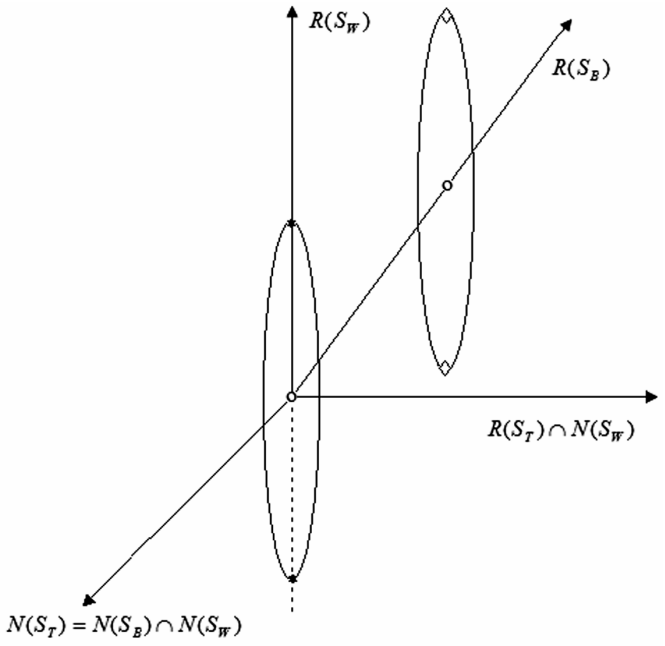

Fig. 2. Illustration of the optimal discriminant subspace.

The projection vectors found by the Direct-LDA method and the Null Space method also differ in terms of orthogonality properties. The projection vectors found by the Direct-LDA method satisfy the orthogonality property, $w_i^T S_W w_j = \delta_{ij}$, whereas the projection vectors found by the Null Space method satisfy the property, $w_i w_j = \delta_{ij}$, where δ_{ij} is the Kronecker's delta.

B. Numerical Example

In this subsection we present a numerical example to show techniques to compute the optimal projection vectors from the optimal discriminant subspace. The samples of each class given below are randomly chosen from the Gaussian distributions with different means and same identity covariance matrix. Let

$x_1^1 = [0.7310 \quad 0.0403 \quad 0.5689 \quad -0.3775 \quad -1.4751 \quad 0.7812]^T$,

$x_2^1 = [0.5779 \quad 0.6771 \quad -0.2556 \quad -0.2959 \quad -0.2340 \quad -0.2656]^T$;

$x_1^2 = [2.1184 \quad 3.4435 \quad 2.6232 \quad 2.9409 \quad 2.2120 \quad 2.5690]^T$,

$x_2^2 = [2.3148 \quad 1.6490 \quad 2.7990 \quad 1.0079 \quad 2.2379 \quad 0.8122]^T$;

$x_1^3 = [-3.0078 \quad -0.9177 \quad -1.6101 \quad -2.6355 \quad -1.5563 \quad -2.8217]^T$,

$x_2^3 = [-2.7420 \quad -2.1315 \quad -1.9120 \quad -2.5596 \quad -2.9499 \quad -1.0137]^T$.

Thus there are $C = 3$ classes, each of which contains 2 samples in a 6-dimensional sample space.

The within-class scatter matrix is

$$S_W = S_1 + S_2 + S_3 = \begin{bmatrix} 0.0663 & -0.3863 & 0.0403 & -0.1860 & -0.2777 & 0.1479 \\ -0.3863 & 2.5495 & -0.2370 & 1.7143 & 1.2177 & 0.1457 \\ 0.0403 & -0.2370 & 0.4009 & -0.2150 & -0.2990 & 0.0042 \\ -0.1860 & 1.7143 & -0.2150 & 1.8745 & -0.0273 & 1.7239 \\ -0.2777 & 1.2177 & -0.2990 & -0.0273 & 1.7416 & -1.9322 \\ 0.1479 & 0.1457 & 0.0042 & 1.7239 & -1.9322 & 3.7255 \end{bmatrix}.$$

The eigenvalues and corresponding eigenvectors of S_W are

$\lambda_1 = 5.5764$, $\alpha_1 = [0.0152 \quad 0.1408 \quad -0.0030 \quad 0.4428 \quad -0.3656 \quad 0.8064]^T$;

$\lambda_2 = 4.3672$, $\alpha_2 = [0.1215 \quad -0.7426 \quad 0.1043 \quad -0.4227 \quad -0.4721 \quad 0.1459]^T$;

$\lambda_3 = 0.4147$, $\alpha_3 = [0.0387 \quad -0.2703 \quad -0.9231 \quad 0.0397 \quad 0.2352 \quad 0.1279]^T$;

$\lambda_4 = 0$, $\alpha_4 = [0.9099 \quad 0.0229 \quad -0.0283 \quad 0.2937 \quad -0.1483 \quad -0.2498]^T$;

$\lambda_5 = 0$, $\alpha_5 = [0.0630 \quad -0.5236 \quad 0.3629 \quad 0.3660 \quad 0.6496 \quad 0.1851]^T$;

$\lambda_6 = 0$, $\alpha_6 = [0.3893 \quad 0.2843 \quad 0.0668 \quad -0.6351 \quad 0.3798 \quad 0.4642]^T$.

If we project samples onto $N(S_W)$, we obtain the same unique vector for all samples of the same class. We call these vectors as "*common vectors*" [14]. The common vectors of the classes are

$$\begin{aligned} x_{com}^1 &= <x_1^1, \alpha_4>\alpha_4 + <x_1^1, \alpha_5>\alpha_5 + <x_1^1, \alpha_6>\alpha_6 = x_1^1 - <x_1^1, \alpha_1>\alpha_1 - <x_1^1, \alpha_2>\alpha_2 - <x_1^1, \alpha_3>\alpha_3 \\ &= <x_2^1, \alpha_4>\alpha_4 + <x_2^1, \alpha_5>\alpha_5 + <x_2^1, \alpha_6>\alpha_6 = x_2^1 - <x_2^1, \alpha_1>\alpha_1 - <x_2^1, \alpha_2>\alpha_2 - <x_2^1, \alpha_3>\alpha_3 \\ &= [0.6131 \quad 0.4971 \quad -0.2522 \quad -0.3373 \quad -0.4085 \quad -0.0993]^T, \\ x_{com}^2 &= <x_1^2, \alpha_4>\alpha_4 + <x_1^2, \alpha_5>\alpha_5 + <x_1^2, \alpha_6>\alpha_6 = x_1^2 - <x_1^2, \alpha_1>\alpha_1 - <x_1^2, \alpha_2>\alpha_2 - <x_1^2, \alpha_3>\alpha_3 \\ &= <x_2^2, \alpha_4>\alpha_4 + <x_2^2, \alpha_5>\alpha_5 + <x_2^2, \alpha_6>\alpha_6 = x_2^2 - <x_2^2, \alpha_1>\alpha_1 - <x_2^2, \alpha_2>\alpha_2 - <x_2^2, \alpha_3>\alpha_3 \\ &= [2.6391 \quad -0.5379 \quad 0.9154 \quad 0.0059 \quad 2.0184 \quad 0.9593]^T, \end{aligned}$$

and

$$x_{com}^3 = < x_1^3, \alpha_4 > \alpha_4 + < x_1^3, \alpha_5 > \alpha_5 + < x_1^3, \alpha_6 > \alpha_6 = x_1^3 - < x_1^3, \alpha_1 > \alpha_1 - < x_1^3, \alpha_2 > \alpha_2 - < x_1^3, \alpha_3 > \alpha_3$$
$$= < x_2^3, \alpha_4 > \alpha_4 + < x_2^3, \alpha_5 > \alpha_5 + < x_2^3, \alpha_6 > \alpha_6 = x_2^3 - < x_2^3, \alpha_1 > \alpha_1 - < x_2^3, \alpha_2 > \alpha_2 - < x_2^3, \alpha_3 > \alpha_3$$
$$= [\text{-}3.1844 \quad 0.9010 \quad \text{-}1.0584 \quad \text{-}0.6488 \quad \text{-}2.1055 \quad \text{-}0.6994]^T.$$

The optimal projection vectors are those that maximize the scatter across the common vectors. In other words, the optimal projection vectors are the eigenvectors corresponding to the nonzero eigenvalues of S_{com}, where $S_{com} = \sum_{i=1}^{3}(x_{com}^i - \mu_{com})(x_{com}^i - \mu_{com})^T$ and $\mu_{com} = \sum_{i=1}^{3} x_{com}^i / 3$ The nonzero eigenvalues and the corresponding eigenvectors of S_{com} are

$\lambda_1 = 30.1010$, $w_1 = [0.7560 \quad \text{-}0.1830 \quad 0.2528 \quad 0.0842 \quad 0.5283 \quad 0.2119]^{\mathrm{T}}$;

$\lambda_2 = 0.6670$, $w_2 = [\text{-}0.6418 \quad \text{-}0.3751 \quad 0.2628 \quad 0.0432 \quad 0.5364 \quad 0.2981]^{\mathrm{T}}$.

The projection matrix of the subspace spanned by the optimal projection vectors is

$$P_{opt} = [w_1 \quad w_2][w_1 \quad w_2]^T = \begin{bmatrix} 0.9834 & 0.1024 & 0.0225 & 0.0359 & 0.0551 & \text{-}0.0311 \\ 0.1024 & 0.1741 & \text{-}0.1448 & \text{-}0.0316 & \text{-}0.2978 & \text{-}0.1506 \\ 0.0225 & \text{-}0.1448 & 0.1329 & 0.0326 & 0.2745 & 0.1319 \\ 0.0359 & \text{-}0.0316 & 0.0326 & 0.0089 & 0.0676 & 0.0307 \\ 0.0551 & \text{-}0.2978 & 0.2745 & 0.0676 & 0.5668 & 0.2718 \\ \text{-}0.0311 & \text{-}0.1506 & 0.1319 & 0.0307 & 0.2718 & 0.1337 \end{bmatrix}.$$

As explained before, optimal projection vectors form an orthonormal basis for the intersection subspace of $N(S_W)$ and $R(S_T)$. Thus, Observation 1 can also be used to find the projection matrix P_{opt} of this intersection subspace. Let $P^{(1)}$ and $P^{(2)}$ represent the projection matrices of $R(S_T)$ and $N(S_W)$ respectively. Then

$$\Psi = 0.5P^{(1)} + 0.5P^{(2)} = \begin{bmatrix} 0.9917 & 0.0512 & 0.0112 & 0.0180 & 0.0276 & \text{-}0.0156 \\ 0.0512 & 0.5871 & \text{-}0.0724 & \text{-}0.0158 & \text{-}0.1489 & \text{-}0.0753 \\ 0.0112 & \text{-}0.0724 & 0.5665 & 0.0163 & 0.1372 & 0.0659 \\ 0.0180 & \text{-}0.0158 & 0.0163 & 0.5045 & 0.0338 & 0.0153 \\ 0.0276 & \text{-}0.1489 & 0.1372 & 0.0338 & 0.7834 & 0.1359 \\ \text{-}0.0156 & \text{-}0.0753 & 0.0659 & 0.0153 & 0.1359 & 0.5669 \end{bmatrix},$$

where

$$P^{(1)} = \begin{bmatrix} 0.9999 & 0.0039 & -0.0006 & -0.0071 & 0.0013 & 0.0038 \\ 0.0039 & 0.8186 & 0.0269 & 0.3340 & -0.0623 & -0.1799 \\ -0.0006 & 0.0269 & 0.9960 & -0.0495 & 0.0092 & 0.0266 \\ -0.0071 & 0.3340 & -0.0495 & 0.3853 & 0.1146 & 0.3312 \\ 0.0013 & -0.0623 & 0.0092 & 0.1146 & 0.9786 & -0.0618 \\ 0.0038 & -0.1799 & 0.0266 & 0.3312 & -0.0618 & 0.8216 \end{bmatrix},$$

and

$$P^{(2)} = \begin{bmatrix} 0.9835 & 0.0985 & 0.0231 & 0.0431 & 0.0538 & -0.0349 \\ 0.0985 & 0.3556 & -0.1717 & -0.3655 & -0.2355 & 0.0294 \\ 0.0231 & -0.1717 & 0.1369 & 0.0821 & 0.2653 & 0.1052 \\ 0.0431 & -0.3655 & 0.0821 & 0.6236 & -0.0470 & -0.3005 \\ 0.0538 & -0.2355 & 0.2653 & -0.0470 & 0.5882 & 0.3336 \\ -0.0349 & 0.0294 & 0.1052 & -0.3005 & 0.3336 & 0.3122 \end{bmatrix}.$$

The eigenvectors corresponding to the eigenvalue 1 are

$e_1 = [0.9630 \quad 0.1968 \quad -0.0649 \quad 0.0143 \quad -0.1254 \quad -0.1175]^{\mathrm{T}}$, and

$e_2 = [-0.2369 \quad 0.3680 \quad -0.3588 \quad -0.0935 \quad -0.7423 \quad -0.3463]^{\mathrm{T}}$.

These vectors also span the same space spanned by the optimal projection vectors computed before, since the projection matrix found by using these vectors is the same as P_{opt} computed before, i.e., $P_{opt} = [e_1 \quad e_2][e_1 \quad e_2]^T$.

Now let $P^{(3)}$ be the projection matrix of the range space of S_B. We need to compute the following matrix to find the intersection of the null space of S_W and the range space of S_B,

$$\tilde{\Psi} = 0.5P^{(2)} + 0.5P^{(3)} = \begin{bmatrix} 0.8436 & 0.0966 & 0.0258 & 0.1110 & -0.0888 & 0.1507 \\ 0.0966 & 0.2561 & 0.0068 & -0.1011 & -0.0025 & 0.0623 \\ 0.0258 & 0.0068 & 0.1833 & 0.1324 & 0.2927 & 0.0909 \\ 0.1110 & -0.1011 & 0.1324 & 0.4020 & 0.0736 & -0.0833 \\ -0.0888 & -0.0025 & 0.2927 & 0.0736 & 0.5699 & 0.1569 \\ 0.1507 & 0.0623 & 0.0909 & -0.0833 & 0.1569 & 0.2451 \end{bmatrix}.$$

There is no eigenvalue of $\tilde{\Psi}$ that corresponds to 1. Thus, the intersection of $N(S_W)$ and $R(S_B)$ is trivial, which clearly indicates that the optimal projection vectors are not in this intersection. Hence the intersection of $N(S_W)$ and $R(S_B)$ alone cannot be used for recognition tasks.

We can also compute the projection matrix of the optimal discriminant subspace directly with the following formula,

$$P_{opt} = P^{(1)}P^{(2)} = P^{(2)}P^{(1)}$$

since $P^{(1)}$ and $P^{(2)}$ commute. Thus, the optimal projection vectors that span the optimal discriminant subspace can also be obtained by the PCA+Null space method. Note also that the projection matrix $P^{(2)}$ of $N(S_W)$ and $P^{(3)}$ of $R(S_B)$ do not commute, i.e., $P^{(2)}P^{(3)} \neq P^{(3)}P^{(2)}$. That is why the Direct-LDA method does not extract features from the intersection of $N(S_W)$ and $R(S_B)$.

Now we can use the optimal projection vectors for dimension reduction. In this case every sample in each class produces the same feature vector, called the *discriminative common vector*. In particular,

$$\Omega_1 = [< x_1^1, w_1 > \quad < x_1^1, w_2 >]^T = [< x_2^1, w_1 > \quad < x_2^1, w_2 >]^T = [\text{-}0.9094 \quad 0.0436]^T,$$
$$\Omega_2 = [< x_1^2, w_1 > \quad < x_1^2, w_2 >]^T = [< x_2^2, w_1 > \quad < x_2^2, w_2 >]^T = [0.1174 \quad 3.5951]^T,$$
$$\Omega_3 = [< x_1^3, w_1 > \quad < x_1^3, w_2 >]^T = [< x_2^3, w_1 > \quad < x_2^3, w_2 >]^T = [0.0618 \quad \text{-}4.1549]^T.$$

As a consequence a 100 % recognition rate is guaranteed for the vectors in the training set in the reduced 2-dimensional space.

III. THE KERNEL DISCRIMINATIVE COMMON VECTOR METHOD

In the Kernel approaches we transform the training set samples into an implicit higher-dimensional space $\Im$ through nonlinear kernel mapping. Let

$\Phi(x_1^1), \Phi(x_2^1),..., \Phi(x_{N_1}^1), \Phi(x_1^2),..., \Phi(x_{N_C}^C)$ represent the transformed samples in $\Im$. The within-class scatter matrix S_W^Φ, the between-class scatter matrix S_B^Φ, and the total scatter matrix S_T^Φ in $\Im$ are given by

$$\begin{aligned} S_W^\Phi &= \sum_{i=1}^{C} \sum_{m=1}^{N_i} (\Phi(x_m^i) - \mu_i^\Phi)(\Phi(x_m^i) - \mu_i^\Phi)^T \\ &= (\Phi - \Phi G)(\Phi - \Phi G)^T, \end{aligned} \tag{21}$$

$$\begin{aligned} S_B^\Phi &= \sum_{i=1}^{C} N_i (\mu_i^\Phi - \mu^\Phi)(\mu_i^\Phi - \mu^\Phi)^T \\ &= (\Phi U - \Phi L)(\Phi U - \Phi L)^T, \end{aligned} \tag{22}$$

and

$$\begin{aligned} S_T^\Phi &= \sum_{i=1}^{C} \sum_{m=1}^{N_i} (\Phi(x_m^i) - \mu^\Phi)(\Phi(x_m^i) - \mu^\Phi)^T \\ &= (\Phi - \Phi 1_M)(\Phi - \Phi 1_M)^T = S_W^\Phi + S_B^\Phi, \end{aligned} \tag{23}$$

where μ^Φ is the mean of all samples, μ_i^Φ is the mean of samples in the *i*-th class, and Φ is the matrix whose columns are the transformed training set samples in $\Im$. Here $G = diag[G_1,..., G_C] \in R^{MxM}$ is a block-diagonal matrix and each $G_i \in R^{N_i x N_i}$ is a matrix with all elements equal to $1/N_i$; $U = diag[u_1,..., u_C] \in R^{MxC}$ is a block-diagonal matrix and each $u_i \in R^{N_i x 1}$ is a vector with all elements equal to $1/\sqrt{N_i}$; $L = [l_1,..., l_C] \in R^{MxC}$ is a matrix where each $l_i \in R^{Mx1}$ is a vector with entries $\sqrt{N_i}/M$; $1_M \in R^{MxM}$ is a matrix with entries $1/M$.

In the transformed space, S_W^Φ is typically singular. Thus the optimal projection vectors that maximize the modified FLDA criterion are in the intersection of the null space $N(S_W^\Phi)$ of S_W^Φ and the range space $R(S_T^\Phi)$ of S_T^Φ. Similar to the linear case, there are mainly two approaches to compute these optimal projection vectors. We can either first project the training set samples onto $N(S_W^\Phi)$ and then apply PCA, or we can first apply PCA to project the training set samples

onto $R(S_T^{\Phi})$ and then find an orthonormal basis for the new null space of the within-class scatter matrix of the transformed samples. However, the first approach is not feasible since the algorithms that accomplish this task work in a higher-dimensional space. Therefore, it is better to follow the second approach. The training set samples can be easily projected onto $R(S_T^{\Phi})$ through the Kernel PCA. Then we can find the vectors that span the new null space of the within-class scatter matrix of the transformed samples. After this operation, we obtain the discriminative common vectors that represent each class. The algorithm can be summarized as follows:

Step 1: Project the training set samples onto $R(S_T^{\Phi})$ through the Kernel PCA. Let

$$\tilde{K} = K - 1_M K - K 1_M + 1_M K 1_M \in R^{MxM} = P \Lambda P^T \tag{24}$$

where the diagonal elements of Λ are nonzero and $K \in R^{MxM}$ is given by $K = \Phi^T \Phi = (K^{ij})_{\substack{i=1,\ldots,C \\ j=1,\ldots,C}}$,where each matrix $K^{ij} \in R^{N_i x N_j}$ can be defined as

$$K^{ij} = (k_{mn}^{ij})_{\substack{m=1,\ldots,N_i \\ n=1,\ldots,N_j}} = < \Phi(x_m^i), \Phi(x_n^j) > = k(x_m^i, x_n^j)_{\substack{m=1,\ldots,N_i \\ n=1,\ldots,N_j}} . \tag{25}$$

The matrix that transforms the training set samples onto $R(S_T^{\Phi})$ is $(\Phi - \Phi 1_M) P \Lambda^{-1/2}$. Then the new total and the within-scatter matrices in the reduced space will be

$$\begin{aligned} \tilde{S}_T^{\Phi} &= ((\Phi - \Phi 1_M) P \Lambda^{-1/2})^T S_T^{\Phi} (\Phi - \Phi 1_M) P \Lambda^{-1/2} \\ &= \Lambda^{-1/2} P^T P \Lambda P^T P \Lambda P^T P \Lambda^{-1/2} = \Lambda \end{aligned} \tag{26}$$

and

$$\begin{aligned} \tilde{S}_W^{\Phi} &= ((\Phi - \Phi 1_M) P \Lambda^{-1/2})^T S_W^{\Phi} (\Phi - \Phi 1_M) P \Lambda^{-1/2} \\ &= \Lambda^{-1/2} P^T \tilde{K}_W \tilde{K}_W{}^T P \Lambda^{-1/2}, \end{aligned} \tag{27}$$

where $\tilde{K}_W = K - KG - 1_M K + 1_M KG = (K - 1_M K)(I - G)$.

Step 2: Find vectors that span the null space of $\widetilde{S}_W^{\Phi}$. This can be performed by an eigen-decomposition. The normalized eigenvectors corresponding to the zero eigenvalues of $\widetilde{S}_W^{\Phi}$ form an orthonormal basis for the null space of $\widetilde{S}_W^{\Phi}$. Let *V* be a matrix whose columns are the computed eigenvectors corresponding to the zero eigenvalues such that,

$$V^T \widetilde{S}_W^{\Phi} V = 0 . \tag{28}$$

Step 3 (optional) : Remove the null space of $V^T \widetilde{S}_B^{\Phi} V$, if it exists and rotate the projection directions so that the new total and between-scatter matrices are diagonal (i.e., the scatter matrices of the feature vectors of the training set samples are uncorrelated). That is,

$$V^T \widetilde{S}_B^{\Phi} V = V^T \widetilde{S}_T^{\Phi} V = V^T \Lambda V = L \widetilde{\Lambda} L^T . \tag{29}$$

Then the final projection matrix *W* will be

$$W = (\Phi - \Phi 1_M) P \Lambda^{-1/2} V L . \tag{30}$$

There are at most *C-1* projection vectors. After performing the feature extraction, all the training set samples in each class produce the discriminative common vector of that class. Therefore, similar to the linear DCV case a 100% recognition accuracy is also guaranteed for this method.

As we stated previously, the Kernel GDA method is equivalent to applying the Kernel PCA method followed by the linear discriminant analysis [18]. After this operation, we also obtain projection vectors that give rise to discriminative common vectors for each class, satisfying the orthogonality constraint $w_i^T S_T^{\Phi} w_j = \delta_{ij}$. Therefore this method also guarantees a 100% recognition accuracy. It should be noted that the discriminative common vectors obtained by the Kernel GDA are different from the ones obtained by the proposed method since the projection vectors of the proposed method are othonormal, i.e., $w_i^T w_j = \delta_{ij}$. This property of the existence

of such discriminative common vectors for the Kernel GDA does not seem to have been noticed in the literature. Thus, the feature vector of a test sample must only be compared to the discriminative common vector of each class during classification, which makes the Kernel DCV and the Kernel GDA methods practical for real-time applications.

IV. EXPERIMENTAL RESULTS

All supervised linear and kernel methods discussed in this paper can be classified in two groups. The methods in the first group (FLDA, Direct-LDA, and Kernel FDA) use the projection directions satisfying the conditions $W^T S_B W \neq 0$ and $W^T S_W W \neq 0$ or $W^T S_B^{\Phi} W \neq 0$ and $W^T S_W^{\Phi} W \neq 0$ for feature extraction. On the other hand, the methods in the second group (DCV, PCA+Null Space, Kernel DCV, and Kernel GDA) use the projection directions that satisfy the conditions $W^T S_B W \neq 0$ and $W^T S_W W = 0$ or $W^T S_B^{\Phi} W \neq 0$ and $W^T S_W^{\Phi} W = 0$. As explained before, the projection directions from the second category come from the optimal discriminant subspace and all training set samples can be classified correctly by using these projection directions for feature extraction. However, the goal of a recognition method is not only to classify all the training data themselves, but also to classify well the test data samples that are not used for training. In other words, we want the recognition method to produce a correct input-output mapping. This is known as the *generalization ability* of a method [1]. In our experiments, we first explored the generalization abilities of those methods coming from the two different general categories separately, and then we investigated whether the performance of the methods from the second category can be improved by adding some projection directions from the first category. In addition to the supervised methods, we also tested the unsupervised methods, the PCA and the Kernel PCA, to give a better assessment of the accuracy of the proposed method.

The dimensionality of the sample space and the size of the training set are two important factors that affect the recognition rates of the methods [25]. Therefore, experiments were performed on data sets from two different populations with different training set sizes and dimensionalities. We have selected two data sets from the first population and one data set from the second population. The size of the training set is larger than the dimensionality of the sample space for the data sets from the first population, unlike in the case of the second population. Therefore, S_W is nonsingular for the data sets from the first population and it is singular for the data set of the second population. In the first group of experiments, since S_W is nonsingular, we cannot apply the linear DCV method. However, it is possible to apply the Kernel DCV method since, as we noted, the training set samples are first transformed into a higher-dimensional space for which S_W^{Φ} is singular. For the second group of experiments, the FLDA method cannot be applied directly. Therefore, we applied the approach suggested by Swets and Weng in which the training set samples are first projected onto an *M-C* dimensional space through PCA, for which S_W is nonsingular [5]. Then, the FLDA method is applied to the projected samples. For the linear PCA and the Kernel PCA methods, the most significant eigenvectors were chosen in such a way that the corresponding eigenvalues contain 95% of the total energy [5].

An appropriate selection of kernel functions for special tasks is still an open problem since different kernel functions give rise to different constructions of the implicit feature space [26]. We have used polynomial kernels $k(x, y) = (< x, y >)^k$, with degrees $k = 2,3,4$ and the Gaussian kernel $k(x, y) = \exp(- \| x - y \|^2 / \gamma)$ for all data sets. The parameter γ was chosen based on empirical observations for each database.

A. Experiments with Large Number of Training Samples

In this group of experiments we used the digit data set, consisting of handwritten numerals (0-9) extracted from a collection of utility maps [27]. There are $C = 10$ classes, each having 200 patterns. Sample patterns are available in the form of binary images. These characters are represented in terms of different feature sets. In our experiments we used only a subset of the original data set consisting of 76 Fourier coefficients and 240 pixel averages.

The odd-numbered samples were used for the training set and the even-numbered samples were used for testing. Thus, a training set of $M = 1000$ samples and a test set of 1000 samples were created. A nearest-neighbor algorithm was employed using the Euclidean distance for classification, except for the methods that employ the discriminative common vectors (DCV, Kernel DCV, and Kernel GDA), in which case the feature vector of the test sample was compared to the discriminative common vectors only by using the Euclidean distance for those methods. The discriminative common vector found to be the closest to the feature vector of the test sample was used to identify the test sample. Recognition results of the test sets for these data sets are given in Tables I and II.

As can be seen from the results, the best recognition rates among the linear methods were obtained by the PCA method for both test sets. For the Fourier Coefficient Database, the best recognition rates among all methods for the test set were obtained by the Kernel DCV and the Kernel FDA methods using the Gaussian kernel. For the Pixel Averages Database, the best recognition rates were obtained by the Kernel DCV and the Kernel GDA methods with the Gaussian kernel. Although the Kernel PCA method did not outperform the classical linear counterpart for the test sets, both the Kernel FDA and the Kernel GDA methods significantly outperformed the FLDA method for both data sets.

TABLE I
Recognition Rates of the 76 Fourier Coefficients Database

Linear Methods	**Recognition Rates (%)**			
PCA	82.5			
FLDA	80.5			
Direct-LDA	80.8			
Kernel Methods	**Recognition Rates (%)**			
	Polynomial kernel functions with different degrees			Gaussian kernel function
	k = 2	k = 3	k = 4	$\lambda = 0.38$
Kernel PCA	81.1	80.6	80.3	81.6
Kernel FDA	82.5	82.6	83.8	85.4
Kernel GDA	80.8	83.3	82.4	85.2
Kernel DCV	83	83.5	83.4	85.4

TABLE II
Recognition Rates of the 240 Pixel Averages Database

Linear Methods	**Recognition Rates (%)**			
PCA	97.3			
FLDA	93.2			
Direct-LDA	95.2			
Kernel Methods	**Recognition Rates (%)**			
	Polynomial kernel functions with different degrees			Gaussian kernel function
	k = 2	k = 3	k = 4	$\gamma = 1200$
Kernel PCA	96.9	97	95.8	97.5
Kernel FDA	97.4	97.6	97.6	97.9
Kernel GDA	97.2	97.5	98	98
Kernel DCV	97.6	97.6	97.7	98

The results show that the proposed method generalizes well compared to other kernel approaches for data sets with large number of samples since for both data sets, the proposed method gives the best recognition results. We also performed some experiments to see if the recognition performance of the Kernel DCV method can be increased by incorporating some projection directions from outside the optimal discriminant subspace into the Kernel DCV framework. In these experiments we used the Gaussian kernels, with the parameters as given in the tables, since these yielded the highest recognition rates. We employed the variation of PCA+Null Space method from [11], to add the projection directions coming from outside the optimal discriminant subspace. We split the new within-class scatter matrix, $\widetilde{S}_W^{\Phi}$ (the within-class scatter matrix of the samples obtained after the Kernel PCA process), into its null space $N(\widetilde{S}_W^{\Phi}) = span\{\xi_{r+1},...,\xi_t\}$ and orthogonal complement (i.e., range space) $R(\widetilde{S}_W^{\Phi}) = span\{\xi_1,...,\xi_r\}$ (where r is the rank of S_W^{Φ}, and $t = rank(S_T^{\Phi})$ is the dimension of the reduced space after Kernel PCA step). Subsequently, all the projection vectors maximizing the between-class scatter in the null space are chosen. These are the projection vectors from the optimal discriminant subspace and there are 9 of them. Then, beginning with these optimal projection vectors, we gradually added new projection vectors from the range space until we reached to the number of $t = 998$ projection vectors, and we computed the corresponding recognition rates. The results for the training and test sets are illustrated in Fig. 3. As can be seen from the figure, adding new projection directions from outside the optimal discriminant subspace does not increase the performance; in fact the performance can be seen to degrade. Adding projection directions from the outside the optimal discriminant subspace also degrades the real-time performance since the added projections no longer produce a unique discriminative common vector for each class. As a result, the comparisons must be made over all feature vectors

of the training set, rather than just over a much smaller number of discriminative common vectors, leading to an increase in the computational cost.

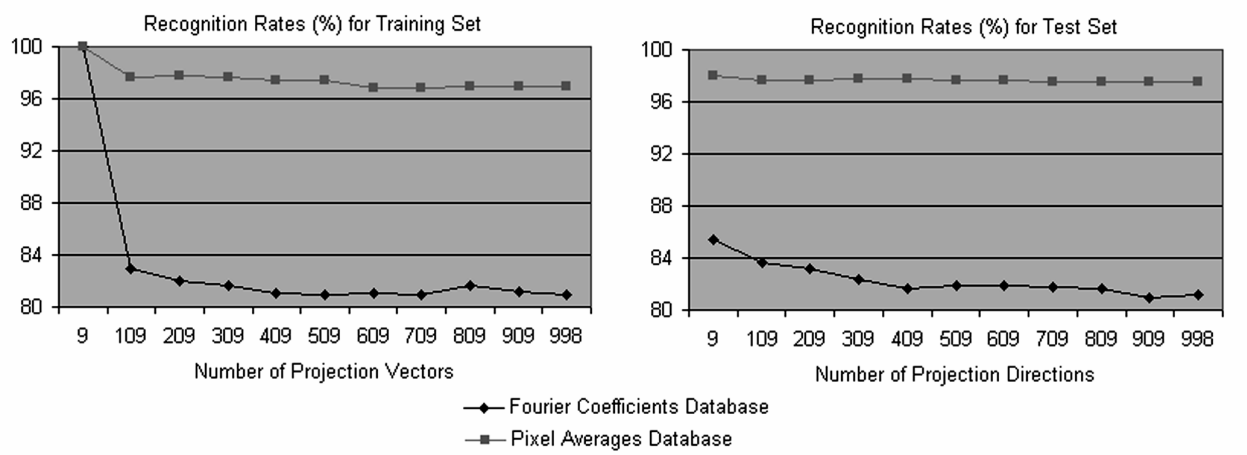

Fig. 3. Recognition rates (%) as a function of projection vectors that are used for feature extraction.

B. Experiments with High-Dimensional Sample Space

In this group of experiments we used the ORL (Olivetti-Oracle Research Lab) face database [28]. The ORL face database contains C=40 individuals with 10 images per person. The images are taken at different times with varying lighting conditions, facial expressions, and facial details. All individuals are in an up-right, frontal position (with tolerance for some side movement). The size of the each image is 92x112 pixels. Some individuals from the ORL face database are shown in Fig. 4.

Fig. 4. Three sample sets from the ORL face database.

We randomly selected $N = 3,5,7$ samples from each class for training and the remaining $(10 - N)$ samples of each class were used for testing. We have not applied any pre-processing to

the images. A nearest-neighbor algorithm was utilized using the Euclidean distance for classification, except for the methods that employ discriminative common vectors. The recognition rates were then computed. This process was repeated six times and the recognition rates for the experiment were found by averaging these rates in each run. The recognition rates for the linear and kernel methods are given in Tables III and IV, respectively. The best recognition was obtained by the DCV method among the linear methods in all cases. The recognition performance of the DCV method is especially superior to the other methods when $N = 3$ samples are used for training. As the number of training samples is increased, the difference between the recognition rates of the DCV method and other linear methods decreases. Similarly, the best recognition results among the Kernel methods were obtained by the Kernel DCV method for all cases. Although the best recognitions among all the methods were obtained by the Kernel DCV method, there was not a significant difference between the recognition rates of the linear DCV and the Kernel DCV methods for this database. An interesting observation is that as the degree of the polynomial kernel is increased, the recognition rates of the test set decrease, which shows that the second-order data correlation is sufficient for good recognition performance.

We also carried out some experiments in order to judge whether the performance of the DCV and the Kernel DCV methods can be increased by adding projection directions from outside the optimal discriminant subspace. The same procedure was followed as in the previous subsection. These experiments were performed on the data sets using $N = 5$ samples for training. The Gaussian kernel with parameter $\gamma = 1.06e8$ was used for the Kernel DCV method. For both methods, starting with 39 optimal projection vectors, we gradually added new projection vectors from outside the optimal discriminant subspace, until we reached the number $t = 199$ of

projection vectors. This procedure was repeated 6 times and the recognition rates were found by averaging the computed recognition rates in each run. The results are given in Fig. 5. As can be seen, adding new projection vectors degraded the performance of the method similar to the previous case.

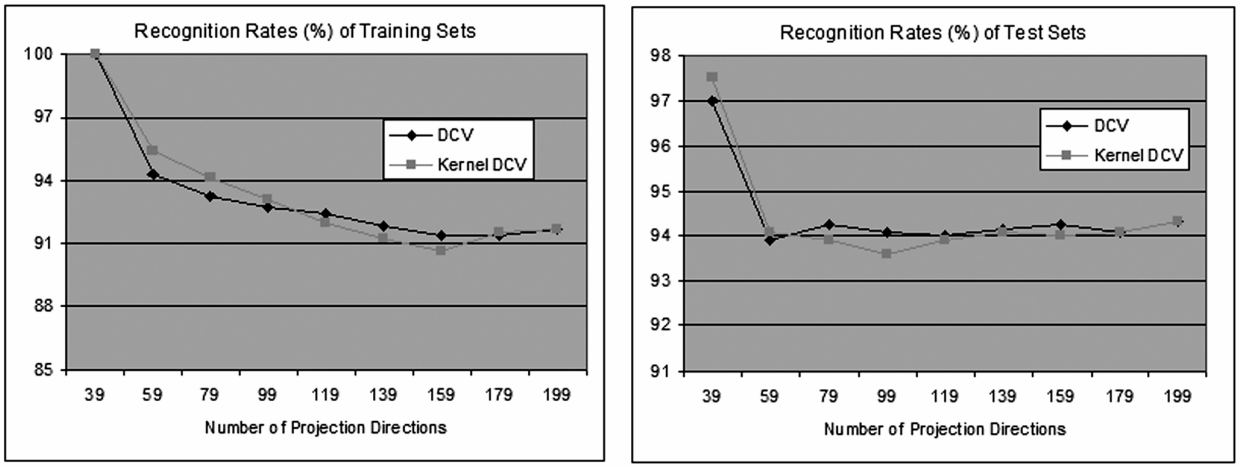

Fig. 5. Recognition rates (%) as a function of projection vectors that are used for feature extraction.

TABLE III
Recognition Rates of the ORL Face Database for Linear Methods

Number of training samples in each class	**Recognition Rates & Standard Deviations**			
	PCA	**FLDA**	**Direct-LDA**	**DCV**
$N = 3$	87.15% $\sigma = 4.03$	86.76% $\sigma = 2.81$	86.61% $\sigma = 3.44$	91.31% $\sigma = 2.01$
$N = 5$	93.66% $\sigma = 2.01$	93.33% $\sigma = 2.62$	96.58% $\sigma = 1.39$	97% $\sigma = 1.41$
$N = 7$	96.94% $\sigma = 1.25$	95.27 % $\sigma = 1.94$	98.33% $\sigma = 1.17$	98.47% $\sigma = 1.10$

These results show that the proposed method leads to a reliable input-output mapping for the data sets with a high-dimensional space by using only a few training set samples.

TABLE IV

Recognition Rates (%) of the ORL Face Databasee for Kernel Methods

Kernel Methods	Number of Training Samples in Each Class											
	$N=3$				$N=5$				$N=7$			
	k = 2	k = 3	k = 4	γ =1.06e8	k = 2	k = 3	k = 4	γ =1.06e8	k = 2	k = 3	k = 4	γ =1.06e8
Kernel PCA	86.20 $\sigma=3.08$	84.15 $\sigma=3.28$	82.86 $\sigma=3.23$	86.25 $\sigma=3.91$	93.33 $\sigma=1.21$	92.75 $\sigma=1.40$	92.16 $\sigma=1.69$	93.75 $\sigma=1.25$	96.94 $\sigma=1.55$	96.38 $\sigma=1.63$	96.39 $\sigma=1.55$	97.08 $\sigma=1.36$
Kernel FDA	90.05 $\sigma=2.66$	87.41 $\sigma=3.14$	84.86 $\sigma=3.82$	89.99 $\sigma=2.23$	96.33 $\sigma=1.57$	95.41 $\sigma=1.59$	93.41 $\sigma=1.93$	96.50 $\sigma=1.18$	97.49 $\sigma=1.58$	96.94 $\sigma=1.63$	96.25 $\sigma=1.46$	98.47 $\sigma=0.82$
Kernel GDA	87.60 $\sigma=2.47$	85.88 $\sigma=3.50$	81.48 $\sigma=4.37$	91.31 $\sigma=2.40$	94.16 $\sigma=0.98$	93.58 $\sigma=1.20$	91.16 $\sigma=2.40$	96.66 $\sigma=0.93$	96.66 $\sigma=1.57$	95.41 $\sigma=2.15$	94.86 $\sigma=2.06$	98.19 $\sigma=0.62$
Kernel DCV	91.38 $\sigma=2.60$	89.40 $\sigma=3.23$	87.25 $\sigma=2.99$	91.60 $\sigma=2.16$	97.00 $\sigma=1.67$	95.91 $\sigma=1.88$	95.08 $\sigma=1.42$	97.50 $\sigma=0.94$	98.61 $\sigma=1.01$	97.91 $\sigma=1.26$	97.91 $\sigma=1.26$	98.75 $\sigma=1.02$

C. Discussion

We have seen in the described experiments that when the dimension of the sample space was smaller than the size of the training set, the Kernel methods typically produced better results than the linear methods. Although the Kernel PCA did not improve the classical PCA method significantly, the supervised kernel approaches, the Kernel FDA and the Kernel GDA methods, outperformed the FLDA method significantly. In most cases the proposed method outperformed the other kernel methods. Unlike the results obtained for the data sets from the first population, there is not a significant difference between the recognition rates of the linear and the kernel methods for the face database. The DCV method outperformed all other linear methods in all cases. Similarly, the Kernel DCV method outperformed all other kernel methods in all cases. This supports the conclusion that the proposed method is suitable for data sets with high-dimensional sample spaces.

The recognition results may be improved for different kernels that fulfill Mercer's theorem [29]. However, we did not attempt to find better kernels since our aim here was to compare the accuracy of the Kernel DCV method with other kernel techniques. The test results show that the projection vectors coming from the optimal discriminant subspace are the best suited set of projection directions for feature extraction. Another advantage of the Kernel DCV method is its real-time performance. The proposed method and the Kernel GDA method yield the highest real-time efficiency among the kernel methods. In these methods, after a test image is projected onto the (C-1) optimal projection vectors, the feature vector of the test sample is compared to C discriminative common vectors only, in sharp contrast to all other methods, where it must be compared to all training set feature vectors if the nearest neighbor algorithm is used. Thus, if we assume that each class has N samples and each kernel method uses (C-1) projection vectors for

feature extraction, then the computational complexity of the other kernel approaches will be N times greater than the computational complexity of the Kernel DCV and the Kernel GDA methods.

V. CONCLUSIONS

In this paper we proposed a new method that uses kernel functions for recognition. The proposed method combines kernel-based methodologies with the optimal discriminant subspace concept. We first showed that the optimal projection vectors come from the optimal discriminant subspace, which is the intersection of the null space of the within-class scatter matrix S_W and the between-class scatter matrix S_B. We then proposed an algorithm for finding these projection vectors in the nonlinearly mapped higher-dimensional space. When the training set samples are projected onto the computed projection vectors, all training set samples in each class produce a unique vector called the discriminative common vector. Thus a 100% recognition rate is guaranteed for the training set samples. To assess the performance of the proposed method, we performed several tests. First, we compared the proposed method with the methods that use projection directions from outside the optimal discriminant subspace. The proposed method outperformed all other kernel methods in most of the cases. Then, we generated a new set of projection vectors by adding new projection vectors from outside the optimal discriminant subspace to the optimal projection vectors. We then used these new vectors for feature extraction. However, this process degraded the performance of the method presented. The results show that the generalization ability of the proposed method is superior to all tested kernel approaches. Also the fact that the test sample feature vectors are compared to only the discriminative common vectors, as opposed to all training set sample feature vectors, makes the proposed method ideal for real-time applications.

REFERENCES

[1] C.M Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1995, pp. 11-17, 295-297, 318-319.

[2] A. Webb, *Statistical Pattern Recognition.* New York: Oxford University Press, 1999, pp. 327-328.

[3] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *In Annals of Eugenics*, vol. 7, pp. 179-188, 1936.

[4] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* 2nd edition, New York: Academic Press, 1990, pp. 31-34, 39-40.

[5] D. L. Swets and J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 831-836, August 1996.

[6] W. Zhao, R. Chellappa, and A Krishnaswamy, "Discriminant analysis of principal components for face recognition," in *Proceedings of 3rd IEEE International Conference on Automatic Face and Gesture Recognition*, April 1998, pp. 336-341.

[7] Z-Q Hong and J-Y Yang, "Optimal discriminant plane for a small number of samples and design method of classifier on the plane," *Pattern Recognition*, vol. 24, pp. 317-324, 1991.

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

[9] Y. Bing, J. Lianfu, and C. Ping, "A new LDA-based method for face recognition," in *Proceedings of 16th International Conference on Pattern Recognition*, August 2002, vol. 1, pp. 168-171.

[10] R. Huang, Q. Liu, H. Lu, and S. Ma, "Solving the small size problem of LDA", in *Proceedings of 16th International Conference on Pattern Recognition*, August 2002, vol. 3, pp. 29-32.

[11] J. Yang, D. Zhang and J-Y Yang, "A generalised K-L expansion method which can deal with small sample size and high-dimensional problems," *Pattern Analysis & Applications*, vol. 6, pp. 47-54, April 2003.

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

[13] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd edition, Springer-Verlag, New York, 1995, pp. 30-31.

[14] H. Cevikalp, M. Neamtu, M. Wilkes, and A. Barkana, "Discriminative common vectors for face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 4-13, January 2005.

[15] B. Schölkopf, *Support Vector Learning*, Ph. D. thesis, Informatik der Technischen Universität, Berlin, 1997.

[16] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller, "Fisher discriminant analysis with kernels," in *Neural Networks for Signal Processing IX*, Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, Eds. Piscataway, NJ: IEEE, pp. 41-48, 1999.

[17] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Computation*, vol. 12, pp. 2385-2404, 2000.

[18] J. Yang, A. F. Frangi, Z. Jin, and J-Y Yang, "Essence of kernel Fisher discriminant: KPCA plus LDA," *Pattern Recognition*, vol. 37, pp. 2097-2100, October 2004.

[19] P.N Belhumeur, J. P. Hespanha, and D. J. Kriegman, “Eigenfaces vs. Fisherfaces: recognition using class specific linear projection,” *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711-720, 1997.

[20] C. W. Therrien, “Eigenvalue properties of projection operators and their applications to the subspace method of feature extraction,” *IEEE Transactions on Computers*, 24 (9) 944-948, 1975.

[21] J. Xu and L. Zikatanov, “The method of alternating projections and the method of subspace corrections in Hilbert space,” *Journal of the American Mathematical Society*, 15 , pp. 573-597, 2002.

[22] J. Lu, K. N. Plataniotis, A. N. Venetsanopoulos, “Face recognition using LDA-based algorithms,” *IEEE Transactions on Neural Networks*, vol. 14, pp. 195-200, 2003.

[23] J. Lu, K. N. Plataniotis, A. N. Venetsanopoulos, “Regularized discriminant analysis for the small sample size problem in face recognition,” *Pattern Recongition Letters*, vol. 24, pp. 3079-3087, 2003.

[24] J. Lu, K. N. Plataniotis, A. N. Venetsanopoulos, “Face recognition using kernel direct discriminant analysis algorithms,” *IEEE Transactions on Neural Networks*, vol. 14, pp. 117-126, 2003.

[25] L. O. Jimenez and D. A. Landgrebe, “Supervised classification in high dimensional space: geometrical, statistical, and asymptotical properties of multivariate data,” *IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews*, 28-1 pp. 39-54, 1998.

[26] F. Perez-Cruz and O. Bousquet, “Kernel methods and their potential use in signal processing,” *IEEE Signal Processing Magazine*, vol. 21, no. 3, pp. 57-65, May 2004.

[27] M. van Breukelen, R.P.W. Duin, D.M.J. Tax, and J.E. den Hartog, “Handwritten digit recognition by combined classifiers,” *Kybernetika*, 34-4 pp. 381-386, 1998.

[28] The ORL Database of Faces, AT&T Laboratories Cambridge. Available: http://www.uk.research.att.com/facedatabase.html

[29] B. Schölkopf and A. J. Smola, *Learning with Kernels*, MIT Press, 2002, pp. 37-39.

[30] N. Cristianini and J. Shawe-Taylor, *An introduction to Support Vector Machines and other kernel-based learning methods*, Cambridge: Cambridge University Press, 2000.

[31] K-R Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, “An introduction to kernel-based learning algorithms,” *IEEE Transaction on Neural Networks*, vol. 12, no. 2, March 2001.