Discriminative Common Vector Method With Kernels

Hakan Cevikalp, *Member, IEEE*, Marian Neamtu, and Mitch Wilkes, *Member, IEEE*

***Abstract*—In some pattern recognition tasks, the dimension of the sample space is larger than the number of samples in the training set. This is known as the "small sample size problem." 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 attempt to answer the question of "How should one choose the optimal projection vectors for feature extraction in the small sample size case?" Based on our findings, we propose a new method called the kernel discriminative common vector method. In this method, we first nonlinearly map the original input space to an implicit higher dimensional feature space, in which the data are hoped to be linearly separable. Then, the optimal projection vectors are computed in this transformed space. The proposed method yields an optimal solution for maximizing a modified Fisher's linear discriminant criterion, discussed in the paper. Thus, under certain conditions, a 100% recognition rate is guaranteed for the training set samples. Experiments on test data also show that, in many situations, the generalization performance of the proposed method compares favorably with other kernel approaches.**

***Index Terms*—Discriminative common vectors, feature extraction, Fisher's linear discriminant analysis, kernel methods, small sample size.**

I. Introduction

FISHER'S linear discriminant analysis (FLDA) is a supervised method that has been successfully applied in many classification problems, such as image recognition, multimedia information retrieval, and medical applications [1]. The method employs the FLDA criterion, which attempts to maximize the ratio $J_{\mathrm{FLDA}}(W_{\mathrm{opt}}) = \max(|W^T S_B W|/|W^T S_W W|)$, where W is the matrix whose columns are the projection vectors used for feature extraction, 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 nonsymmetric, its eigendecomposition may be unstable. To circumvent this problem, the simultaneous diagonalization algorithm is often employed [2], [3]. A 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. The rank of S_W cannot exceed the number of samples in the training set; thus S_W will be singular if the dimensionality of the sample space is larger than the size of the training set. This problem is also known as the *small sample size problem* [2]. To deal with this situation, the perturbation method has been used in [4] and [5], in which S_W is perturbed so as to become nonsingular. Swets and Weng [3] proposed a two-stage PCA + FLDA method, also known as the Fisherface method, in which principal component analysis (PCA) is first used for dimension reduction in order to make S_W nonsingular before applying FLDA.

Recently, Yu and Yang proposed the direct-LDA method to cope with the small sample size problem [6]. In this method, which also employs the simultaneous diagonalization for finding projection vectors in the range of S_B, the null space of S_B is first discarded, and then the projection vectors minimizing the within-class scatter in the transformed space are selected from the range of S_B. However, the range of S_B does not necessarily include the optimal projection vectors for discrimination [7]–[10].

Chen *et al.* proposed the null space method for the small sample size case, based on the modified FLDA criterion $J_{\mathrm{MFLDA}}(W_{\mathrm{opt}}) = \max(|W^T S_B W|/|W^T S_T W|)$, where S_T is the total scatter matrix of the training set data [11]. In this method, all training samples are first projected onto the null space of S_W. 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 one; therefore the null space method extracts features, which are optimal from a discrimination point of view. It turns out that the resulting orthonormal projection vectors span the space obtained as the intersection of the null space of S_W and the range of S_T. We call this space the *optimal discriminant subspace* since it is spanned by vectors that extract 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 preprocessing step was used to extract geometric features and to reduce the dimension of the original sample space. Then, they applied the null space method in the reduced space. Vapnik [12] suggests that when solving a given problem, one should avoid solving a more general problem as an intermediate step. Abiding by this principle, we showed that any preprocessing step, reducing the original dimension of the null space, is likely to reduce the performance of the method and therefore should be avoided [13]. In [13], we proposed the discriminative common vector (DCV) method for finding optimal orthonormal projection vectors in the optimal discriminant subspace. This method is equivalent to the null space method, with the exception that the dimension reduction 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

Manuscript received December 29, 2004; revised February 28, 2006.

H. Cevikalp is with the Department of Electrical and Electronics Engineering, Eskisehir Osmangazi Universitesi, 26480 Meselik, Eskisehir, Turkey (e-mail: hakan.cevikalp@gmail.com).

M. Neamtu is with the Center for Constructive Approximation and the Department of Mathematics, Vanderbilt University, Nashville, TN 37240 USA (e-mail: neamtu@math.vanderbilt.edu).

M. Wilkes is with the Center for Intelligent Systems and the Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN 37235 USA (e-mail: mitch.wilkes@vanderbilt.edu).

Digital Object Identifier 10.1109/TNN.2006.881485

algorithm uses the range 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 [9] for finding optimal projection vectors spanning the optimal discriminant subspace. In this method, PCA is first applied to remove the null space of the total scatter matrix S_T. Then, optimal projection vectors are found in the complementary lower dimensional space using the null space method. However, this method is computationally more expensive compared to the DCV method (see [13] for a comparison of these methods).

In some cases, linear methods may not provide a sufficient discriminating power for classifying linearly nonseparable classes (e.g., exclusive-or problem). Therefore, discriminant analysis techniques utilizing kernels have been recently proposed in [14]–[16]. Their main idea is to transform the input data into a higher dimensional space by a nonlinear mapping function and then apply the linear discriminant analysis techniques in that space. These methods are formulated in terms of dot products of the mapped samples, and kernel functions are used to compute these dot products. Therefore, the nonlinear mapping function and the mapped samples are not used explicitly, which makes the methods computationally feasible. However, the singularity problem of the involved matrices is typically encountered in this approach since the dimensionality of the mapped space is usually larger than the size of the training set (in particular, this problem always arises if one uses the Gaussian kernels). Two different techniques have been adopted to solve this problem. Mika *et al.* use the original FLDA criterion in the nonlinearly mapped space and add a small perturbation matrix to the involved singular matrix [14]. Yang *et al.* use the modified FLDA criterion instead of the original FLDA criterion in the mapped space [16]. They first project the data onto the range of the total scatter matrix of the mapped samples using the kernel PCA method [17]; then they apply the LDA method, which maximizes the modified FLDA criterion in this reduced space. The first approach above is called the kernel Fisher's discriminant analysis (kernel FDA) method, and the latter approach is called the kernel PCA + LDA (KPCA + LDA) method.

In this paper, we propose a new method, coined 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 kernel 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 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. Last, our conclusions are presented in Section V.

II. Optimal Projection Vectors

The modified FLDA criterion aims to maximize the ratio $J_{\mathrm{MFLDA}}(W_{\mathrm{opt}}) = \max(|W^T S_B W|/|W^T S_T W|)$. However, this criterion is not appropriate since the maximization does not have a unique solution in 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 matrices W. However, it is not reasonable to use matrices W with a small number of projection vectors since they may not be sufficient for optimal feature extraction. On the other hand, the following criterion, called the *null space based FLDA* function criterion, has a unique maximum for the projection vectors with unit length and also maximizes the modified FLDA criterion:

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

Note that the value of $J_{\mathrm{NSFLDA}}(W)$ is dependent on the lengths of the projection vectors; thus one should normalize the columns of W to have norm of one, to make the maximization of $J_{\mathrm{NSFLDA}}(W)$ a well-posed problem. To find optimal projection vectors maximizing this criterion, we first project the training set samples onto $N(S_W)$ and then obtain the projection vectors by performing PCA. As a result, we obtain a set of orthonormal vectors that forms a basis for the space, which we call the optimal discriminant subspace. This subspace is defined as the intersection of $N(S_W)$ and the range $R(S_T)$ 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 for finding this optimal subspace and an orthonormal basis for it. Some of them are given in [13].

Next, we investigate two methods, the DCV and the PCA + null space methods, both of which employ orthonormal basis vectors of the optimal discriminant subspace for feature extraction. We also give a proof of the fact that the basis vectors obtained by these methods span the same subspace, and therefore both of these methods yield equivalent solutions.

A. The Optimal Discriminant Subspace Concept

Let the training set be composed of C classes, where the ith class contains N_i samples, and let x_m^i be a d-dimensional column vector, which denotes the mth sample from the ith class. Thus, there are a total of $M = \sum_{i=1}^{C} N_i$ samples in the training set. The within-class scatter matrix S_W, the between-class scatter matrix S_B, and the total scatter matrix S_T are defined as

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

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

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

where μ is the mean of all samples and μ_i is the mean of samples in the ith class.

If the dimension d of the sample space is larger than $M-1$, the ranks of S_W, S_B, and S_T can be at most $M-C, C-1$, and $M-1$, respectively. As a result, all scatter matrices will be rank deficient in this case. Note that the dimension of the range space of a scatter matrix is equal to its rank and that the range and the null space of scatter matrices are orthogonal complements of each other. Moreover, the eigenvectors of a scatter matrix, corresponding to zero eigenvalues, span the null space of this scatter matrix. As explained previously, if the projection directions are chosen from $N(S_W)$, the modified FLDA criterion attains its maximum: one. Therefore, we must project the training set data onto $N(S_W)$. Then, the optimal projection vectors maximizing the null space based FLDA criterion given in (1) 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, whose proof is given in the Appendix.

Lemma 1: Suppose $\bar{U}$ is a matrix whose column vectors u_k ($k = r_T+1, \ldots, d$, 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 give rise to a unique common vector x such that

$$x = \bar{U}\bar{U}^T x_m^i, \quad i = 1, \ldots, C, \quad m = 1, \ldots, N_i \tag{5}$$

where x is independent of indexes i and m.

This lemma shows that $N(S_T)$ does not contain any discriminative information, which can be used for obtaining the optimal projection vectors maximizing the null space based FLDA criterion function. Therefore, this null space can be discarded. The appropriate subspace for extracting discriminating features will then be the intersection of $N(S_W)$ and $R(S_T)$.

There are numerous algorithms for finding the optimal discriminant subspace and optimal projection vectors that span it. The following observation by Therrien [19] can be used to find these optimal projection vectors and the corresponding optimal discriminant subspace.

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

$$\Psi = \sum_{i=1}^{n} a_i P^{(i)} \tag{6}$$

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

In our case, we can choose $L^{(1)}$ and $L^{(2)}$ 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 could be too large (e.g., images of size 256 by 256 result in projection matrices of size 65 536 by 65 536). There are computationally more suitable ways to find the optimal projection vectors by using smaller sets of basis vectors. This follows from 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

$$P_{\text{opt}} = P^{(1)}P^{(2)} = P^{(2)}P^{(1)} \tag{7}$$

where P_{opt} is the projection matrix of the optimal discriminant subspace [20]. A consequence of this result 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. All projections are performed economically by using basis vectors instead of projection operators. 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 lemma.

Lemma 2: Let $L^{(1)}, L^{(2)}$ be subspaces of $\Re^d$, $L^{(1)^\perp}, L^{(2)^\perp}$ be their orthogonal complements and $P^{(1)}, P^{(2)}$ be the orthogonal projection matrices onto $L^{(1)}$ and $L^{(2)}$, respectively. If $L^{(1)^\perp} \perp L^{(2)^\perp}$, then $P^{(1)}$ and $P^{(2)}$ commute, that is, $P^{(1)}P^{(2)} = P^{(2)}P^{(1)}$.

Proof: If $L^{(1)^\perp} \perp L^{(2)^\perp}$, then clearly $(I - P^{(1)})(I - P^{(2)}) = 0$ and $(I - P^{(2)})(I - P^{(1)}) = 0$. Thus

$$(I - P^{(1)})(I - P^{(2)}) = (I - P^{(2)})(I - P^{(1)}) = 0 \tag{8}$$

$$I - P^{(1)} - P^{(2)} - P^{(1)}P^{(2)} = I - P^{(1)} - P^{(2)} - P^{(2)}P^{(1)} \tag{9}$$

which implies $P^{(1)}P^{(2)} = P^{(2)}P^{(1)}$. □

We are now ready to prove the following theorem.

Theorem 1: Let $P^{(1)}$ and $P^{(2)}$ be the projection matrices of the subspaces $R(S_T)$ and $N(S_W)$, respectively. Then $P^{(1)}$ and $P^{(2)}$ commute, i.e.,

$$P^{(1)}P^{(2)} = P^{(2)}P^{(1)}.$$

Proof: Let $L^{(1)} = R(S_T)$ and $L^{(2)} = N(S_W)$. Clearly, $L^{(1)^\perp} = N(S_T)$ and $L^{(2)^\perp} = R(S_W)$. By [13, Lemma 1], we have

$$\begin{aligned} N(S_T) &= N(S_B + S_W) \\ &= N(S_B) \cap N(S_W) \end{aligned} \tag{10}$$

and, in particular, $N(S_T) \subset N(S_W)$. This, together with the fact that $N(S_W) \perp R(S_W)$, shows that

$$N(S_T) \perp R(S_W) \text{ or } L^{(1)^\perp} \perp L^{(2)^\perp}. \tag{11}$$

The assertion of the theorem now follows from Lemma 2. □

In [21]–[23], the authors claim that the direct-LDA method finds the projection vectors in the intersection space of $N(S_W)$ and $R(S_B)$, and they conclude that the projection vectors found by this method are optimal and equivalent to the projection vectors found by the null space method as well as the DCV and the PCA + null space methods. However, these assertions are incorrect in our opinion. First, the projection directions obtained by the direct-LDA method are elements of $R(S_B)$, and hence they are not necessarily in the intersection of $R(S_B)$ and $N(S_W)$. In fact, the intersection of $R(S_B)$ and $N(S_W)$

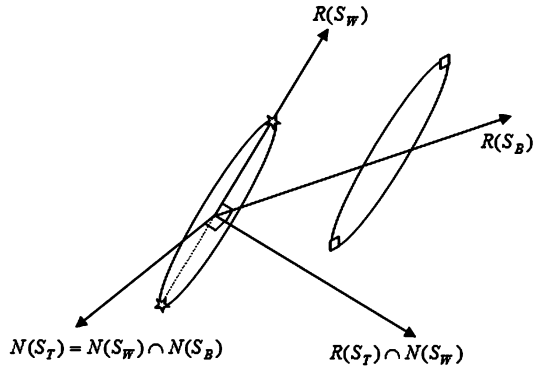


Fig. 1. Illustration of the optimal discriminant subspace.

is often trivial. Indeed, in all the face databases considered in this paper, this intersection turns out to be trivial. This means that the intersection of $R(S_B)$ and $N(S_W)$ should not be used for recognition. Second, in light of Theorem 1, the null space method (equivalently the DCV and the PCA + null space methods) finds the projection vectors in the intersection of $R(S_T)$ and $N(S_W)$. Thus, the null space method yields different results from the direct-LDA method. Last, it should be noted that, in general, the intersection of $N(S_W)$ and $R(S_B)$ is not the same as the optimal discriminant subspace $R(S_T) \cap N(S_W)$. These facts are also illustrated in Fig. 1. In Fig. 1, two classes with identical covariance matrices having two samples each in a three-dimensional space are plotted. $R(S_W)$ 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 $N(S_T)$ 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, i.e., the origin.

The projection vectors found by the direct-LDA method and the null space method also differ in terms of their orthogonality properties. In particular, the projection vectors found by the direct-LDA method satisfy $w_i^T S_W w_j = \delta_{ij}$, whereas the projection vectors found by the null space method are such that $w_i^T w_j = \delta_{ij}$, where δ_{ij} is the Kronecker's delta.

B. Distinctness of Discriminative Common Vectors

If all samples in each class are projected onto $N(S_W)$, they give rise to a unique vector called the *common vector*

$$x_{\text{com}}^i = P^{(2)} x_m^i, \quad i = 1, \ldots, C, \quad m = 1, \ldots, N_i \tag{12}$$

where $P^{(2)}$ is the orthogonal projection operator of $N(S_W)$ [13]. A natural question arises whether the common vectors $x_{\text{com}}^i, i = 1, \ldots, C$, are distinct, i.e., whether each of these vectors can be uniquely associated with the ith class—or, put yet another way, whether there is a one-to-one correspondence between the common vectors and the classes. For if this is not the case, i.e., if $x_{\text{com}}^i = x_{\text{com}}^j$, for some $i \neq j$, then the DCV method would not be able to discriminate between the two classes i and j, which would render this method less useful.

The next result shows that this situation is in practice very unlikely, even though it is possible in theory. First we state the following *necessary* condition for the common vectors to be distinct.

Observation 2: Let $i \neq j$. For the common vectors $x_{\text{com}}^i, x_{\text{com}}^j$ to be distinct it is necessary that the samples x_m^i, x_n^j in the corresponding two classes are such that one cannot find real numbers α_m, β_n satisfying $\sum_{m=1}^{N_i} \alpha_m = 1, \sum_{n=1}^{N_j} \beta_n = 1$, and such that

$$\sum_{m=1}^{N_i} \alpha_m x_m^i = \sum_{n=1}^{N_j} \beta_n x_n^j. \tag{13}$$

To explain this, let us first reformulate the above observation. Recall that the *affine hull* $\text{aff}(X)$ of a finite set X of vectors is the set (called an affine space)

$$\text{aff}(X) := \left\{ \sum_{x \in X} \lambda_x x, \sum_{x \in X} \lambda_x = 1, \lambda_x \in \Re \right\}. \tag{14}$$

Thus, the above observation can be rephrased by saying that a necessary condition for the common vectors of classes i and j to be distinct is that

$$A_i \cap A_j = \emptyset \tag{15}$$

where A_i, A_j are the affine hulls of the vectors in the ith and jth classes, respectively. It is known that the common vectors x_{com}^i can be obtained by projecting any $x \in A_i$ onto $N(S_W)$ (for example, x can be chosen to be μ_i) [13]. Consequently, we have

$$x_{\text{com}}^i = P^{(2)} x \tag{16}$$

whenever $x \in A_i$ and

$$x_{\text{com}}^j = P^{(2)} x \tag{17}$$

for $x \in A_j$. Thus, if $A_i \cap A_j \neq \emptyset$, then clearly $x_{\text{com}}^i = x_{\text{com}}^j$ since above one can take $x \in A_i \cap A_j$, which would give $x_{\text{com}}^i = P^{(2)} x = x_{\text{com}}^j$.

Unfortunately, the above observation does not constitute a sufficient condition for the common vectors to be distinct. This can be easily seen by taking classes of vectors satisfying $A_i \cap A_j = \emptyset$, for all $i \neq j$, but such that $N(S_W) = \{0\}$, in which case all common vectors will be the trivial vectors. To arrive at a sufficient condition, it will therefore be necessary to impose a condition on linear separability of the considered classes.

For the purpose of the following result, we will say that the given classes $i = 1, \ldots, C$ are *linearly separable* if for each pair $i \neq j$ there exists a hyperplane $H \subset \Re^d$ strictly separating the affine spaces A_i and A_j, such that $A_k \cap H = \emptyset$, for all $k \neq i, j$. As usual, A_i and A_j are said to be strictly separated by H if A_i and A_j are on the opposite sides of H and if $A_i \cap H = A_j \cap H = \emptyset$. Thus, this concept of separability is stronger than the usual "one against one" separability but weaker than the "one against all" separability [24]. As is well known, the above definition is equivalent to saying that there exists a linear functional φ on $\Re^d$

such that $\varphi A_i < \varphi H < \varphi A_j$ and $\varphi A_k \neq \varphi H, k \neq i, j$, where φ is such that φH is constant.

We are now ready to prove the following sufficient condition for existence of distinct common vectors.

Theorem 2: Suppose the classes $i = 1, \ldots, C$ are linearly separable. Then, the corresponding common vectors are distinct.

Proof: We will show that for any pair $i \neq j$, we have $x^i_{\text{com}} \neq x^j_{\text{com}}$. To this end, let φ be the linear functional whose existence is guaranteed by the definition of separability. Let l be the unique one-dimensional subspace of $\Re^d$ such that $l \perp H$, and let P_l be the orthogonal projection operator onto this subspace. Clearly, also $l \perp A_i, A_j$. We have, by separability

$$\varphi P_l A_i = \varphi A_i < \varphi H < \varphi A_j = \varphi P_l A_j \tag{18}$$

or, in particular, $P_l A_i \neq P_l A_j$. Next, let S_i be the scatter matrix of the ith class. It is known that the range of S_i is equal to the linear span of the vectors $x^i_m - \mu_i, m = 1, \ldots, N_i$, where μ_i is the mean of the vectors in the ith class [13], and thus $R(S_i) = A_i - \mu_i$. Consequently, l is a subspace of every $N(S_i), i = 1, \ldots, C$, since $l \perp A_i$ and $N(S_i)$ is the orthogonal complement of $R(S_i)$. Moreover, it was shown in [13] that $N(S_W) = \bigcap_{i=1}^{C} N(S_i)$. Hence, $l \subset N(S_W)$. Combining this with the fact that $P_l A_i \neq P_l A_j$, we thus obtain

$$x^i_{\text{com}} = P^{(2)} A_i \neq P^{(2)} A_j = x^j_{\text{com}} \tag{19}$$

since clearly, if the orthogonal projections onto a subspace l are distinct, then so are the projections onto a larger space $N(S_W)$. □

Note that if there are only two classes with corresponding affine hulls A_1 and A_2, then linear separability is equivalent to the condition $A_1 \cap A_2 = \varnothing$, which is a simple consequence of the Hahn–Banach theorem. Thus, for two classes, the above necessary condition is also sufficient.

Corollary 1: If all samples $x^i_m, i = 1, \ldots, C, m = 1, \ldots, N_i$, are linearly independent, then the common vectors $x^i_{\text{com}}, i = 1, \ldots, C$, are distinct.

The crux of the proof of this assertion consists in showing that linear independence of the samples implies linear separability of the classes, in the sense above. Since the proof of this fact is elementary, it will be omitted.

If the common vectors are distinct, then clearly so are the discriminative common vectors. The above conditions are typically satisfied for the data sets in high-dimensional sample spaces. For instance, for a typical face recognition problem with 256-level grayscale face images of size 128×128, the volume of the sample space is $(16\,384)^{256}$. Since the dimension is so high, it is very likely that the training set samples will be linearly independent, and therefore the DCV method can be applied safely for pattern recognition. It has been reported that the generalization performance of the DCV method is superior to competing methods for high-dimensional pattern classification tasks. In fact, the generalization performance is related to the dimensionality of $N(S_W)$ in the sense that the higher dimensions yield better results [13].

In some cases, the dimensionality of the sample space may not be large enough to ensure that the discriminative common vectors are distinct. There are three basic approaches to cope with this situation. First, we can discard all dependent samples. A second solution is to add new orthonormal projection vectors from outside the optimal discriminant subspace to the projection vectors spanning the optimal discriminant subspace. In this case, since the new projection vectors will be from $R(S_W)$, the feature vectors will no longer yield the same discriminative common vectors. Therefore, a 100% recognition accuracy is no longer guaranteed since some training samples might be misclassified in this case. A third solution would be to map the training samples into a higher dimensional space, in which the new discriminative common vectors of classes are distinct, as in the kernel DCV method introduced in the next section.

III. Kernel Discriminative Common Vector Method

Sometimes the discriminative common vectors are not distinct in the original sample space. In such cases one can map the original sample space to a higher dimensional space $\Im$, where the new discriminative common vectors in the mapped space are distinct from one another. This is because a mapping $\phi : \Re^d \to \Im$ can map two vectors that are linearly dependent in the original sample space onto two vectors that are linearly independent in $\Im$ [25]. Note that the mapped space could have arbitrarily large, possibly infinite, dimensionality, which suggests the use of the DCV method. Tsuda proved that if the kernel matrix K, given in (24), is positive definite, then all mapped samples are linearly independent [26]. Therefore, even though the data samples may be linearly dependent in the original sample space, it is guaranteed that the discriminative common vectors are distinct in $\Im$ as long as the kernel matrix K is positive definite. Therefore, a 100% recognition rate can be obtained for linearly nonseparable classes when applying the linear DCV method in $\Im$.

Let $\Phi = [\phi(x^1_1)\phi(x^1_2)\ldots\phi(x^1_{N_1})\phi(x^2_1)\ldots\phi(x^C_{N_C})]$ represent the matrix whose columns are the transformed training samples in $\Im$. The within-class scatter matrix S^Φ_W, the between-class scatter matrix S^Φ_B, and the total scatter matrix S^Φ_T in $\Im$ are given by

$$\begin{aligned} S^\Phi_W &= \sum_{i=1}^{C} \sum_{m=1}^{N_i} (\phi(x^i_m) - \mu^\Phi_i)(\phi(x^i_m) - \mu^\Phi_i)^T \\ &= (\Phi - \Phi G)(\Phi - \Phi G)^T \end{aligned} \tag{20}$$

$$\begin{aligned} S^\Phi_B &= \sum_{i=1}^{C} N_i (\mu^\Phi_i - \mu^\Phi)(\mu^\Phi_i - \mu^\Phi)^T \\ &= (\Phi U - \Phi L)(\Phi U - \Phi L)^T \end{aligned} \tag{21}$$

$$\begin{aligned} S^\Phi_T &= \sum_{i=1}^{C} \sum_{m=1}^{N_i} (\phi(x^i_m) - \mu^\Phi)(\phi(x^i_m) - \mu^\Phi)^T \\ &= (\Phi - \Phi 1_M)(\Phi - \Phi 1_M)^T = S^\Phi_W + S^\Phi_B \end{aligned} \tag{22}$$

where μ^Φ is the mean of all samples and μ^Φ_i is the mean of samples of the ith class in $\Im$. Here, $G = \text{diag}[G_1, \ldots, G_C] \in \Re^{M \times M}$ is a block-diagonal matrix and each $G_i \in \Re^{N_i \times N_i}$ is a matrix with all its elements equal to $1/N_i$; $U = \text{diag}[u_1, \ldots, u_C] \in \Re^{M \times C}$ is a block-diagonal matrix and each $u_i \in \Re^{N_i \times 1}$ is a vector with all its elements equal to $1/\sqrt{N_i}$; $L = [l_1, \ldots, l_C] \in \Re^{M \times C}$ is a matrix where each

$l_i \in \Re^{M\times 1}$ is a vector with entries $\sqrt{N_i}/M$; $1_M \in \Re^{M\times M}$ is a matrix with entries $1/M$.

In the transformed space, S_W^Φ is typically singular due to the high dimensionality of the mapped space. Thus the optimal projection vectors that maximize the null space based FLDA criterion are in the intersection of the null space $N(S_W^\Phi)$ of S_W^Φ and the range $R(S_T^\Phi)$ of S_T^Φ. Similar to the linear case, there are mainly two approaches for computing 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 follow this approach use the mapping function ϕ explicitly. Therefore, the second approach is more appropriate. The training set samples can be 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. Consequently, we obtain the discriminative common vectors that represent each class. This algorithm can be summarized as follows.

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

$$\begin{aligned}\tilde{K} &= K - 1_M K - K 1_M + 1_M K 1_M \\ &= U\Lambda U^T \in \Re^{M\times M}\end{aligned} \tag{23}$$

where Λ is the diagonal matrix of nonzero eigenvalues and U is the matrix of normalized eigenvectors associated to Λ. Here the kernel matrix $K \in \Re^{M\times M}$ is given by $K = \Phi^T\Phi = (K^{ij})_{\substack{i=1,\ldots,C\\ j=1,\ldots,C}}$, where each matrix $K^{ij} \in \Re^{N_i\times N_j}$ is defined as

$$\begin{aligned}K^{ij} &= (k_{mn}^{ij})_{\substack{m=1,\ldots,N_i\\ n=1,\ldots,N_j}} = \langle \phi(x_m^i), \phi(x_n^j)\rangle \\ &= k(x_m^i,\, x_n^j)_{\substack{m=1,\ldots,N_i\\ n=1,\ldots,N_j}}\end{aligned} \tag{24}$$

where $k(\cdot)$ represents the kernel function. The matrix that transforms the training set samples onto $R(S_T^\Phi)$ is $(\Phi - \Phi 1_M)U\Lambda^{-1/2}$ [14]. Then the new total and the within-scatter matrices in the reduced space can be shown to be

$$\begin{aligned}\tilde{S}_T^\Phi &= ((\Phi - \Phi 1_M)U\Lambda^{-1/2})^T S_T^\Phi (\Phi - \Phi 1_M)U\Lambda^{-1/2} \\ &= \Lambda^{-1/2}U^T U\Lambda U^T U\Lambda U^T U\Lambda^{-1/2} = \Lambda\end{aligned} \tag{25}$$

$$\begin{aligned}\tilde{S}_W^\Phi &= ((\Phi - \Phi 1_M)U\Lambda^{-1/2})^T S_W^\Phi (\Phi - \Phi 1_M)U\Lambda^{-1/2} \\ &= \Lambda^{-1/2}U^T \tilde{K}_W \tilde{K}_W^T U\Lambda^{-1/2}\end{aligned} \tag{26}$$

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 $\tilde{S}_W^\Phi$. This can be done by eigendecomposition. The normalized eigenvectors corresponding to zero eigenvalues of $\tilde{S}_W^\Phi$ form an orthonormal basis for the null space of $\tilde{S}_W^\Phi$. Let V be a matrix whose columns are the computed eigenvectors, corresponding to zero eigenvalues, such that

$$V^T \tilde{S}_W^\Phi V = 0. \tag{27}$$

Step 3 (Optional): Remove the null space of $V^T\tilde{S}_B^\Phi V$ 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

$$V^T\tilde{S}_B^\Phi V = V^T\tilde{S}_T^\Phi V = V^T\Lambda V = L\tilde{\Lambda}L^T. \tag{28}$$

The final projection matrix W will then be

$$W = (\Phi - \Phi 1_M)U\Lambda^{-1/2}VL. \tag{29}$$

There are at most $C-1$ projection vectors. The feature vector Ω_{test} of a test sample is obtained as

$$\Omega_{\text{test}} = W^T(\phi(x_{\text{test}}) - \mu^\Phi) \tag{30}$$

where W is the matrix whose columns are the projection vectors $w_j, j = 1,\ldots,C-1$. Then each entry of the feature vector of the test sample can be obtained by

$$\begin{aligned}&\langle w_j, \phi(x_{\text{test}}) - \mu^\Phi\rangle \\ &\quad = \langle w_j, \phi(x_{\text{test}}) - \Phi 1_M'\rangle \\ &\quad = (U\Lambda^{-1/2}VL)^T(K_{\text{test}} - K1_M' \\ &\qquad - 1_M K_{\text{test}} + 1_M K 1_M')\end{aligned} \tag{31}$$

where $1_M' \in \Re^{M\times 1}$ is a vector with all terms equal to $1/M$ and $K_{\text{test}} \in \Re^{M\times 1}$ is a vector with entries $< \phi(x_m^i), \phi(x_{\text{test}}) >_{\substack{i=1,\ldots,C\\ m=1,\ldots,N_i}}$.

All mathematical properties of the linear DCV carry over to the kernel DCV method with the modifications that they now apply to the mapped samples, $\phi(x_m^i), i = 1,\ldots,C, m = 1,\ldots,N_i$, in $\Im$. After performing the feature extraction, all training set samples in each class typically give rise to distinct discriminative common vectors. Therefore, as in the linear DCV case, a 100% recognition accuracy with respect to the training data is also guaranteed for this method. In practice, if we cannot easily find kernel functions which guarantee that the discriminative common vectors in $\Im$ are distinct, we can add new projection vectors from outside the optimal discriminant subspace, as described in Section II. However, in our experience, it was very rare that any of the kernels ever exhibited this problem and, in particular, the Gaussian kernels were never observed to have this problem.

As stated previously, the KPCA + LDA method is equivalent to applying the kernel PCA method, followed by linear discriminant analysis [16]. Following these operations, we also obtain projection vectors that give rise to discriminative common vectors for each class. Therefore, this method also guarantees a 100% recognition accuracy if the discriminative common vectors are distinct in the mapped space. It should be noted that the discriminative common vectors obtained by the KPCA + LDA are different from those obtained by the proposed method since the projection vectors of the proposed method are orthonormal, i.e., $w_i^T w_j = \delta_{ij}$. Additionally, the projection vectors are orthogonal with respect to S_T^Φ and S_B^Φ if the optional Step 3 in the kernel DCV algorithm is carried out. More formally

$$W^T S_T^\Phi W = W^T S_B^\Phi W = \tilde{\Lambda} \tag{32}$$

where $\tilde{\Lambda}$ is the diagonal matrix given in (28). On the other hand, the projection vectors of KPCA + LDA are conjugate with respect to S_T^{Φ}, i.e., $w_i^T S_T^{\Phi} w_j = \delta_{ij}$. These projection directions will be orthonormal if the total scatter of the mapped samples is isotropic in the transformed space $\Im$. Keen readers can refer to [27] for a comparison of orthonormal and conjugate projection vectors in linear discriminant analysis. The property of the existence of such discriminative common vectors for KPCA + LDA 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 KPCA + LDA methods practical for real-time applications. Note that these methods do not offer any advantages over other competing methods during the computation of the feature vectors of a test sample. Thus, if one uses a single representative prototype feature vector (e.g., mean of the feature vectors) for each class during classification of a kernel method, the real-time performance of this method will be similar to the kernel DCV and the KPCA + LDA methods.

A. Comparison of the Linear DCV and the Kernel DCV Methods

Mapping samples to a higher dimensional space via nonlinear mapping function ϕ has some advantages over the linear DCV method. The differences between the two methods can be summarized as follows.

i) The DCV method extracts linear features from the original sample space, whereas the kernel DCV method extracts features from an implicit higher dimensional space. It is possible to extract nonlinear features using the kernel DCV method since the mapped space is nonlinearly related to the original sample space. Additionally, we have the flexibility of creating different nonlinear decision boundaries by simply changing the kernel functions. However, these improvements are achieved at the expense of more intense computations.

ii) The DCV method can be applied only in the small sample size case, and the dimensionality of the null space of the within-class scatter matrix must be large in comparison with the training set size for good recognition rates. However, these limitations do not apply to the proposed kernel method. We can apply the kernel DCV method to the data sets, in which the number of the samples is larger than the dimensionality of the sample space, using kernel functions ensuring high dimensionality of the mapped space.

IV. Experimental Results

All supervised linear and kernel feature extraction methods discussed in this paper can be classified into two groups. The methods in the first group (FLDA, direct-LDA, and kernel FDA) use projection directions from $R(S_W)$ or $R(S_W^{\Phi})$ for feature extraction, i.e., the projection vectors satisfy $W^T S_W W \neq 0$ for linear methods and satisfy $W^T S_W^{\Phi} W \neq 0$ for nonlinear methods. On the other hand, projection vectors of the methods in the second group (DCV, PCA + null space, kernel DCV, and KPCA + LDA) come from $N(S_W)$ or $N(S_W^{\Phi})$ and they satisfy $W^T S_W W = 0$ or $W^T S_W^{\Phi} W = 0$. As explained before, projection directions of the methods of the second category span 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 training data 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 [28]. In our experiments, we first tested 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 $R(S_W)$ or $R(S_W^{\Phi})$. In addition to the supervised feature extraction methods, we also tested the support vector machine (SVM) classifier to give a better assessment of the recognition accuracy of the proposed method. The nearest neighbor (NN) and the nearest mean (NM) algorithms [29] were employed using the Euclidean distance during classification of data samples in feature extraction methods, except for the methods that employ the discriminative common vectors (DCV, kernel DCV, and KPCA + LDA), in which case the feature vector of the test sample was compared only to the discriminative common vectors by using the Euclidean distance for those methods.

The dimensionality of the sample space and the size of the training set are two important factors that affect recognition rates of methods [30]. Therefore, experiments were performed on data sets from two different populations with different training set sizes and dimensionalities. We selected two databases from the first population and one database from the second population. The size of the training set is larger than the dimensionality of the sample space for the databases 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 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 some 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 were first projected onto an $M - C$ dimensional space through PCA, for which S_W is nonsingular [3]. Then, the FLDA method was applied to the projected samples. We adopted the "one against one" procedure to extend classic two-class SVM problem to the multiclass recognition problem since this procedure was reported to be more suitable for practical uses [24]. The "one against one" procedure constructs $C(C-1)/2$ classifiers where each classifier is trained on data samples from two classes. Then, the so-called "max wins" voting approach was utilized during the testing phase [24].

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 [31].

We used polynomial kernels $k(x,y) = (\langle x,y\rangle)^n$, with degrees $n = 2, 3$ and the Gaussian kernel $k(x,y) = \exp(-\|x-y\|^2/q)$ for all data sets. We employed a small set of randomly created training and test sets to compute the best Gaussian parameters q for each database. These data sets were only used for parameter selection and were not employed for testing the generalization performances of the methods. We first computed the minimum and the maximum values of Gaussian parameters that produce acceptable recognition rates by globally searching over a wide range of the parameter space. Then, we linearly divided the interval determined by the minimum and the maximum values of parameters into some subintervals and computed the recognition rates. Finally, we carried out a local search in the neighborhood of the Gaussian parameter that yielded the best recognition rate and computed the final best Gaussian parameter. This process was repeated for every method. For the SVM case, this search was carried out in a two-dimensional parameter space since it was also necessary to tune the regularization parameter γ, which is a positive constant used as an upper bound needed in relaxing constraints [17].

A. Experiments on the Scenarios Without Small Sample Size Problem

In this group of experiments we tested the proposed algorithm with two databases. The first database is the well-known Fisher's Iris database [1], and the second database is the digit data set consisting of handwritten numerals (0–9) extracted from a collection of utility maps [32]. The number of samples is larger than the dimensionality of the sample space for both databases.

1) Experiments on the Fisher's Iris Database: The Iris flower database contains four measurements on 50 Iris specimens for each of three species: Iris setosa, Iris versicolor, and Iris virginica for a total of 150 samples in the database. It was reported that the first class is linearly separable from the other two classes and that the latter two are not linearly separable from each other.

We first conducted experiments to visualize the extracted features. We applied the proposed method and the other feature extraction methods discussed in the paper to the Iris database and plotted the extracted features. The data samples were centered before the feature extraction. We used the Gaussian kernel with $q = 0.7$ for all kernel methods. The feature vectors obtained by the feature extraction algorithms are illustrated in Fig. 2. As can be seen in the figure, all samples are separable for the kernel methods, whereas they are not separable for the linear methods.

In the second set of experiments, we tested the generalization performances of the methods by adopting the leave-one-out strategy [2]. We also tested the SVM classifier beside the feature extraction methods. The recognition rates and the Gaussian parameters, which were found by the search procedure described previously, are given in Table I. The regularization parameter γ of SVM classifier was chosen as five. Note that we used the Gaussian kernel only since the small sample size does not occur in the mapped space for the polynomial kernel functions. In this case, the kernel DCV and the KPCA + LDA methods cannot be used for recognition.

In terms of classification performance, the linear FLDA method followed by the NM classifier and the SVM classifier achieved the best recognition rates for the Iris database. The proposed method outperformed both the kernel FDA and the KPCA + LDA methods among the kernel methods. The kernel feature extraction methods did not show any improvement over the linear FLDA method on this database. These recognition rates can be compared to those reported in [16] and [33]. Although our proposed method was outperformed by the SVM classifier, the application of the multiclass SVM classifier is much more complicated than the kernel DCV method. It is because one must adjust some parameters for each binary classifier of SVM as opposed to the kernel DCV method in which the solution can be found in a closed form.

2) Experiments on the Digit Dataset of Handwritten Numerals: This database includes $C = 10$ 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 forming distinct databases. In our experiments, we used two separate data sets consisting of 76 Fourier coefficients and 240 pixel averages.

We randomly chose 100 samples from each class for training and used the remaining samples for testing. Thus, a training set of $M = 1000$ samples and a test set of 1000 samples were created for each database. This process was repeated 45 times, and 45 different training and test sets were created. The first five data sets were used for parameter selection, and the rest were used for performance evaluation. Thus, the final recognition rates for the experiment were found by averaging these 40 rates obtained in each trial. The regularization parameter of SVM classifier was set to five for polynomial kernels and three for the Gaussian kernel for the 76 Fourier coefficients data set and it was chosen to be ten for the polynomial kernels and five for the Gaussian kernel for the 240 pixel averages data set. The means and the standard deviations of computed recognition rates on these databases are given in Tables II and III.

As can be seen from Table II, the best recognition rate among the linear methods was obtained by the direct-LDA method followed by the NN classifier for the Fourier coefficients database. The SVM classifier using the Gaussian kernel achieved the highest recognition rate over all methods. The proposed method achieved either competitive or the best recognition rates among the kernel feature extraction methods for this database. Both the kernel FDA and the KPCA + LDA methods outperformed the FLDA for all kernel functions used here.

We also performed statistical significance tests to evaluate the differences between the recognition rates of the proposed method and the other competing methods from Table II. This test is a null hypothesis statistical test [34]. If the resulting significance is below the desired significance level, the null hypothesis is rejected and the performance difference between two methods is considered to be statistically significant. The details of the test can be found in the Appendix. The results of testing for significance (with significance level of 0.05) in the observed recognition rates are given in Table IV for the Fourier coefficients database. We compared the proposed method only to the other kernel methods and to the linear method that achieved the best recognition rate between the linear methods. In terms of

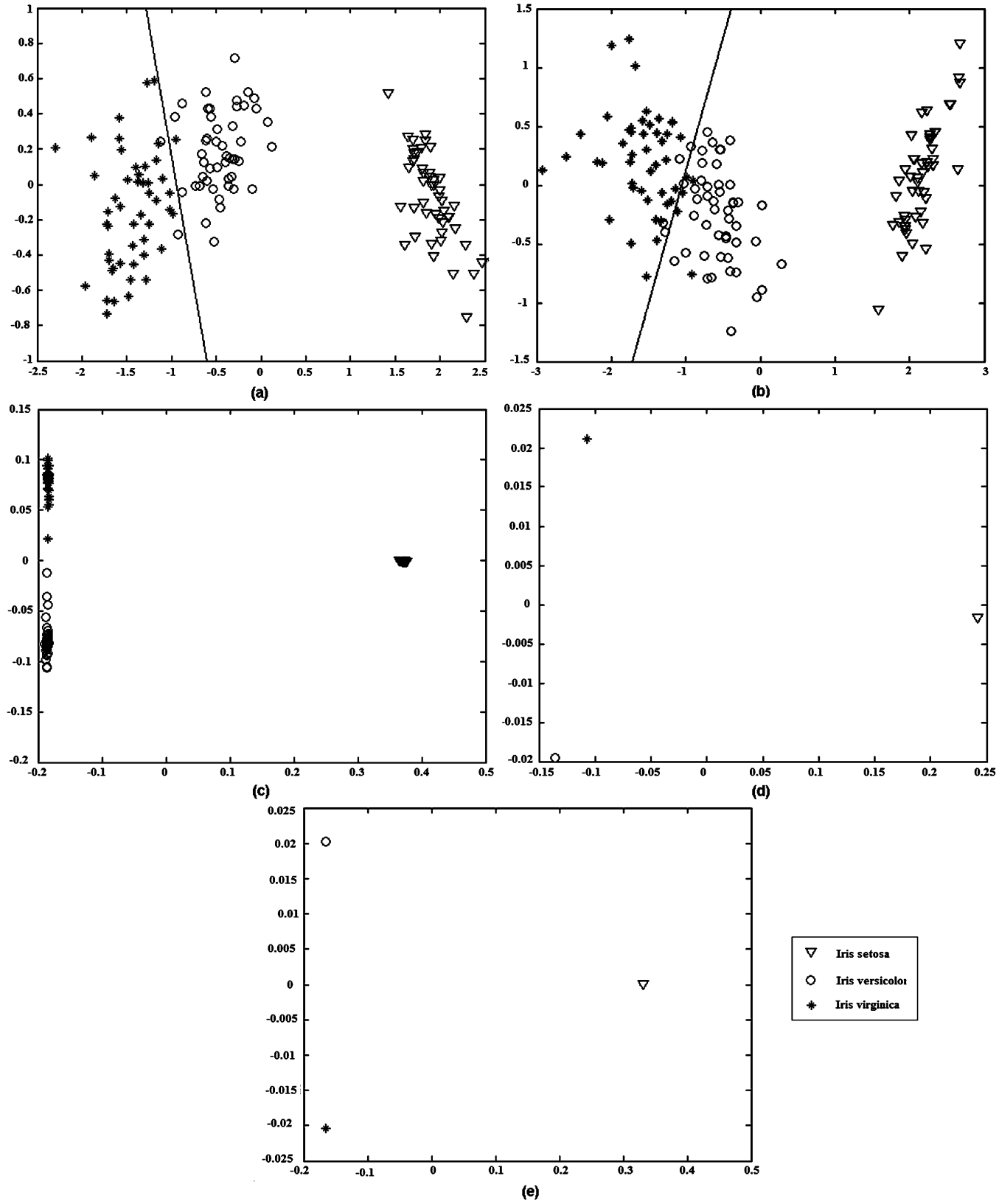


Fig. 2. Feature vectors obtained by the linear and kernel methods: (a) FLDA, (b) direct-LDA (c) kernel FDA, (d) KPCA + LDA, and (e) kernel DCV. The lines represent the decision boundaries of nonseparable classes obtained by the nearest mean classifier.

recognition performance, the term "0" implies the two methods are statistically equivalent; "1" implies the proposed method performs better; and "−1" implies the proposed method is worse than the compared method in the table. The recognition rates obtained by using the Gaussian kernels were generally observed to be the best overall. With regard to the Gaussian kernels, the pro-

TABLE I
RECOGNITION RATES OF METHODS ON THE FISHER'S IRIS DATABASE

Methods & Gaussian Kernel Parameters	Recognition Rates (%)	
	NN	NM
FLDA	96.67	**98**
Direct-LDA	92.67	94
Kernel FDA, $q = 0.7$	95.33	95.33
KPCA+LDA, $q = 0.2$	94.67	
Kernel DCV, $q = 0.1$	96	
SVM, $q = 3$	**98**	

posed method was found to be significantly better than the direct-LDA, FLDA, and kernel FDA methods with a significance level 0.05.

Similarly to the previous case, the best recognition rate among the linear methods was obtained by the direct-LDA method followed by the NN classifier for the pixel averages database as can be seen in Table III. The proposed method achieved the highest recognition rates in all cases. As in the previous case, both the kernel FDA and the KPCA + LDA methods outperformed the FLDA. Additionally, we performed statistical significance tests to evaluate the differences between the recognition rates of the proposed method and the other competing methods on the pixel averages database. The results of the significance test are given in Table V. The results show that the proposed method significantly outperforms the SVM classifier, the direct-LDA method, and the FLDA method in all cases with a significance level of 0.05 on the pixel averages database. The proposed method also significantly outperformed the KPCA + LDA method when the polynomial kernel with degree of two was used.

In general, the test results show that the proposed method generalizes well compared to other kernel approaches for data sets with large number of samples studied here since for both data sets, the proposed method achieves either competitive or the best recognition results. We also conducted 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. Only one randomly created training and test set were used for both data sets in these experiments. We used the Gaussian kernels, with the parameters as given in the tables, since these yielded the highest recognition rates. The variation of the PCA + null space method from [10] was employed to add new projection directions coming from outside the optimal discriminant subspace to the set of projection vectors spanning the optimal discriminant subspace. We split the new within-class scatter matrix $\tilde{S}_W^{\Phi}$ (the within-class scatter matrix of the samples obtained after the kernel PCA process) into its null space $N(\tilde{S}_W^{\Phi}) = \text{span}\,\{\xi_{r+1}, \ldots, \xi_t\}$ and orthogonal complement of the null space (i.e., range space) $R(\tilde{S}_W^{\Phi}) = \text{span}\,\{\xi_1, \ldots, \xi_r\}$ (where r is the rank of $\tilde{S}_W^{\Phi}$ and $t = \text{rank}(S_T^{\Phi})$ is the dimension of the reduced space after the kernel PCA step). Subsequently, all the projection vectors maximizing the between-class scatter in the null space are chosen. These are the projection vectors spanning the optimal discriminant subspace and there are nine 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 outside the optimal discriminant subspace also degrades the real-time performance since the data samples do not give rise to unique discriminative common vectors after feature extraction. As a result, if one does not utilize a single representative prototype feature vector for each class during classification, 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.

B. Experiments on the Scenarios With Small Sample Size Problem

In this group of experiments, we used the Olivetti-Oracle Research Lab (ORL) face database [35]. The ORL face database contains $C = 40$ individuals with ten images per person. The images are taken at different times with varying lighting conditions, facial expressions, and facial details. All individuals are in an upright, frontal position (with tolerance for some side movement). The size of the each image is 92×112 pixels. Some individuals from the ORL face database are shown in Fig. 4.

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. This process was repeated 45 times, and 45 different training and test sets were created. The first five data sets were used for parameter selection, and the rest were used for performance evaluation. We did not apply any preprocessing to the images. The recognition rates for the experiment were found by averaging the recognition rates of each trial. We set the SVM regularization parameter to $\gamma = 10$ for polynomial kernels and to $\gamma = 5$ for the Gaussian kernel in all cases. The computed recognition rates and standard deviations for the linear and kernel methods are given in Tables VI and VII, 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 linear 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.

Similar to the large sample size case, we also performed statistical significance tests to evaluate the differences between the recognition rates of the proposed method and the other competing methods for the ORL face database. The results are given in Table VIII. Our proposed method either matches or significantly outperforms the other kernel methods, and it also shows an improvement over the linear DCV when $N = 7$ samples were used for training. However, it statistically performs worse than the linear DCV method for the polynomial kernel with degree of

TABLE II
RECOGNITION RATES OF METHODS ON THE 76 FOURIER COEFFICIENTS DATABASE

Linear Methods	**Recognition Rates (%) and Standard Deviations**					
	NN			**NM**		
FLDA	80.03, $\sigma = 0.99$			80.22, $\sigma = 0.81$		
Direct-LDA	81.11, $\sigma = 0.94$			79.32, $\sigma = 0.83$		
Kernel Methods & Gaussian Kernel Parameters	**Recognition Rates (%) and Standard Deviations**					
	Polynomial kernel functions with different degrees				Gaussian kernel function	
	$n = 2$		$n = 3$			
	NN	**NM**	**NN**	**NM**	**NN**	**NM**
Kernel FDA, $q = 0.46$	82.59, $\sigma = 1.07$	82.85, $\sigma = 0.79$	83.60, $\sigma = 0.86$	83.91, $\sigma = 0.84$	84.72, $\sigma = 0.84$	84.90, $\sigma = 0.71$
KPCA+LDA, $q = 0.38$	80.85, $\sigma = 1.03$		82.04, $\sigma = 0.98$		84.90, $\sigma = 0.78$	
Kernel DCV, $q = 0.46$	82.62, $\sigma = 0.83$		83.42, $\sigma = 0.92$		85.12, $\sigma = 0.64$	
SVM, $q = 0.38$	**84.93**, $\sigma = 0.82$		**84.86**, $\sigma = 0.76$		**85.18**, $\sigma = 0.80$	

TABLE III
RECOGNITION RATES OF METHODS ON THE 240 PIXEL AVERAGES DATABASE

Linear Methods	**Recognition Rates (%) and Standard Deviations**					
	NN			**NM**		
FLDA	94.08, $\sigma = 0.65$			94.70, $\sigma = 0.65$		
Direct-LDA	95.94, $\sigma = 0.55$			93.24, $\sigma = 0.57$		
Kernel Methods & Gaussian Kernel Parameters	**Recognition Rates (%) and Standard Deviations**					
	Polynomial kernel functions with different degrees				Gaussian kernel function	
	$n = 2$		$n = 3$			
	NN	**NM**	**NN**	**NM**	**NN**	**NM**
Kernel FDA, $q = 1200$	97.85, $\sigma = 0.38$	97.85, $\sigma = 0.38$	98.07, $\sigma = 0.36$	98.07, $\sigma = 0.36$	98.18, $\sigma = 0.34$	98.14, $\sigma = 0.33$
KPCA+LDA, $q = 1200$	97.69, $\sigma = 0.41$		98.06, $\sigma = 0.32$		98.18, $\sigma = 0.32$	
Kernel DCV, $q = 1200$	**97.97**, $\sigma = 0.33$		**98.14**, $\sigma = 0.35$		**98.20**, $\sigma = 0.32$	
SVM, $q = 30$	97.64, $\sigma = 0.46$		97.78, $\sigma = 0.39$		97.94, $\sigma = 0.35$	

TABLE IV
STATISTICAL SIGNIFICANCE COMPARISON OF RECOGNITION PERFORMANCES ON THE FOURIER COEFFICIENTS DATABASE

Kernel Functions	**KDCV/KFDA**		**KDCV/KPCA+LDA**	**KDCV/SVM**	**KDCV/Direct-LDA**
	NN	**NM**			
$n = 2$	0	0	1	-1	1
$n = 3$	0	-1	1	-1	1
GK	1	0	0	0	1

three for $N = 3$. This can be attributed to the nature of the face images in the database. The images of individuals are mostly in frontal position, and the lighting conditions are similar. Therefore the face images in the database are linearly separable. In such cases, using higher order correlations via kernels may degrade the performance as in our case since the problem is close to linearly separable. These results on the ORL face database can be compared to those reported in [36] and [37].

Finally, we carried out some experiments in order to judge whether the performance of the DCV and the kernel DCV

TABLE V
STATISTICAL SIGNIFICANCE COMPARISON OF RECOGNITION PERFORMANCES ON THE PIXEL AVERAGES DATABASE

Kernel Functions	KDCV/KFDA		KDCV/KPCA+LDA	KDCV/SVM	KDCV/Direct-LDA
	NN	NM			
$n = 2$	0	0	1	1	1
$n = 3$	0	0	0	1	1
GK	0	0	0	1	1

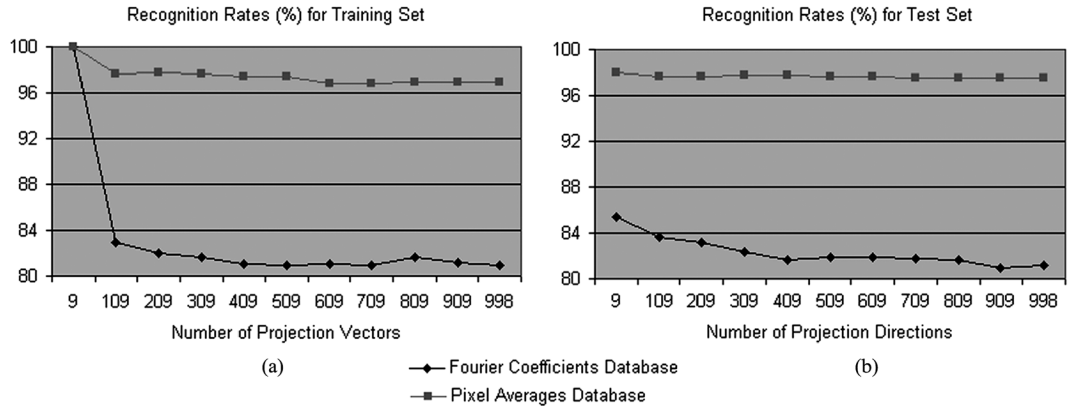


Fig. 3. Recognition rates as a function of projection vectors that are used for feature extraction: (a) training set results and (b) test set results.

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

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 set using $N = 5$ samples for training. The Gaussian kernel with parameter $q = 1.06e8$ 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. The results are given in Fig. 5. As can be seen, adding new projection vectors degraded the performance of the method similar to the large sample size case.

In general, 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.

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, kernel methods typically produced better results than linear methods. Although the supervised kernel feature extraction methods did not show any improvement over the linear FLDA method for the Iris database, they outperformed the FLDA significantly for the digit data sets. In many cases the proposed method outperformed other kernel methods. It should be noted that the number of training samples is large compared to the dimensionality of the sample space for the Iris database. Therefore, it is better to estimate distribution functions of classes for these situations. Then, more sophisticated classifiers can be constructed by using the estimated density functions. However, the estimation of density functions may not be reliable in cases where the dimensionality of the sample space and the number of samples per class are comparable in size, as was the case with the digit data sets. It has been reported that the number of samples in each class must be at least ten times the dimensionality of the sample space for a reliable density estimation [33]. Thus, our proposed kernel method will be suitable for these cases as demonstrated in experimental studies. Unlike the results obtained for the data sets from the first population, in general there is not a significant difference between the recognition rates of the linear and the kernel methods for the face database since the face samples are linearly separable. The DCV method outperformed all other linear methods in all cases for the face database. Similarly, the kernel DCV method outperformed all other kernel methods in all cases. The proposed method, kernel DCV, offered an improvement over its linear counterpart only in one case. However, the kernel DCV method might improve the recognition results of the linear DCV method on different face databases having nonlinear and complex distributions.

TABLE VI
RECOGNITION RATES OF LINEAR METHODS ON THE ORL FACE DATABASE

Number of training samples in each class	Recognition Rates (%) & Standard Deviations				
	FLDA		Direct-LDA		DCV
	NN	NM	NN	NM	
$N=3$	86.35, $\sigma=3.17$	86.10, $\sigma=3.37$	85.31, $\sigma=3.16$	84.84, $\sigma=2.81$	**90.70,** $\sigma=2.49$
$N=5$	92.13, $\sigma=2.47$	92.50, $\sigma=2.27$	95.40, $\sigma=1.64$	94.85, $\sigma=2.04$	**96.11,** $\sigma=1.73$
$N=7$	94.54, $\sigma=2.31$	94.97, $\sigma=2.18$	97.75, $\sigma=1.33$	97.43, $\sigma=1.54$	**97.77,** $\sigma=1.33$

TABLE VII
RECOGNITION RATES OF KERNEL METHODS ON THE ORL FACE DATABASE

Number of training samples	Kernel functions	Classifier	Recognition Rates (%) & Standard Deviations			
			Kernel FDA $q_3=q_5=3.18e7$, $q_7=7.96e7$	KPCA+LDA $q_3=7.96e7$, $q_5=q_7=3.18e7$	Kernel DCV $q_3=q_5=1.06e8$, $q_7=1.06e8$	SVM $q_3=3500$, $q_5=q_7=4000$
$N=3$	$n=2$	NN	89.39, $\sigma=2.80$	87.16, $\sigma=3.14$	**90.54,** $\sigma=2.72$	88.75, $\sigma=2.94$
		NM	89.39, $\sigma=2.80$			
	$n=3$	NN	87.33, $\sigma=3.25$	85.13, $\sigma=3.62$	**89.09,** $\sigma=2.93$	87.68, $\sigma=3.11$
		NM	87.33, $\sigma=3.25$			
	GK	NN	90.40, $\sigma=2.55$	91.42, $\sigma=2.57$	**91.50,** $\sigma=2.56$	89.35, $\sigma=2.81$
		NM	89.91, $\sigma=2.73$			
$N=5$	$n=2$	NN	95.36, $\sigma=1.75$	93.58, $\sigma=1.99$	**96.32,** $\sigma=1.68$	95.22, $\sigma=2.01$
		NM	95.41, $\sigma=1.77$			
	$n=3$	NN	94.46, $\sigma=1.71$	92.91, $\sigma=2.22$	**95.45,** $\sigma=1.83$	94.81, $\sigma=1.97$
		NM	94.46, $\sigma=1.71$			
	GK	NN	96.42, $\sigma=1.52$	96.46, $\sigma=1.39$	**96.71,** $\sigma=1.53$	95.46, $\sigma=1.86$
		NM	96.02, $\sigma=1.50$			
$N=7$	$n=2$	NN	97.29, $\sigma=1.95$	96.10, $\sigma=1.98$	**97.85,** $\sigma=1.46$	97.34, $\sigma=1.31$
		NM	97.29, $\sigma=1.95$			
	$n=3$	NN	96.73, $\sigma=1.41$	95.56, $\sigma=1.92$	**97.67,** $\sigma=1.52$	97.06, $\sigma=1.49$
		NM	96.73, $\sigma=1.41$			
	GK	NN	98.08, $\sigma=1.39$	97.89, $\sigma=1.25$	**98.40,** $\sigma=1.18$	97.56, $\sigma=1.25$
		NM	97.60, $\sigma=1.40$			

The recognition rates of the kernel methods might be improved for different kernels that fulfill Mercer's theorem [38]. 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 KPCA + LDA method have 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.

V. CONCLUSION

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 range of the total scatter matrix S_T. We then proposed an algorithm for finding these projection vectors in a nonlinearly mapped higher dimensional space. Under certain conditions, when the training set samples are projected onto the computed projection vectors, all training set samples in each class produce a distinct vector, called the discriminative common vector, representing the classes. Thus a 100% recognition rate is guaranteed for the training set samples even though they are not linearly separable in the original sample space. To assess

TABLE VIII
STATISTICAL SIGNIFICANCE COMPARISON OF RECOGNITION PERFORMANCES ON THE ORL FACE DATABASE

Number of training samples	Kernel functions	KDCV/KFDA		KDCV/PCA+LDA	KDCV/SVM	KDCV/DCV
		NN	NM			
$N = 3$	$n = 2$	0	0	1	1	0
	$n = 3$	1	1	1	1	-1
	GK	0	1	0	1	0
$N = 5$	$n = 2$	1	1	1	1	0
	$n = 3$	1	1	1	0	0
	GK	0	1	0	1	0
$N = 7$	$n = 2$	0	0	1	0	0
	$n = 3$	1	1	1	0	0
	GK	0	1	0	1	1

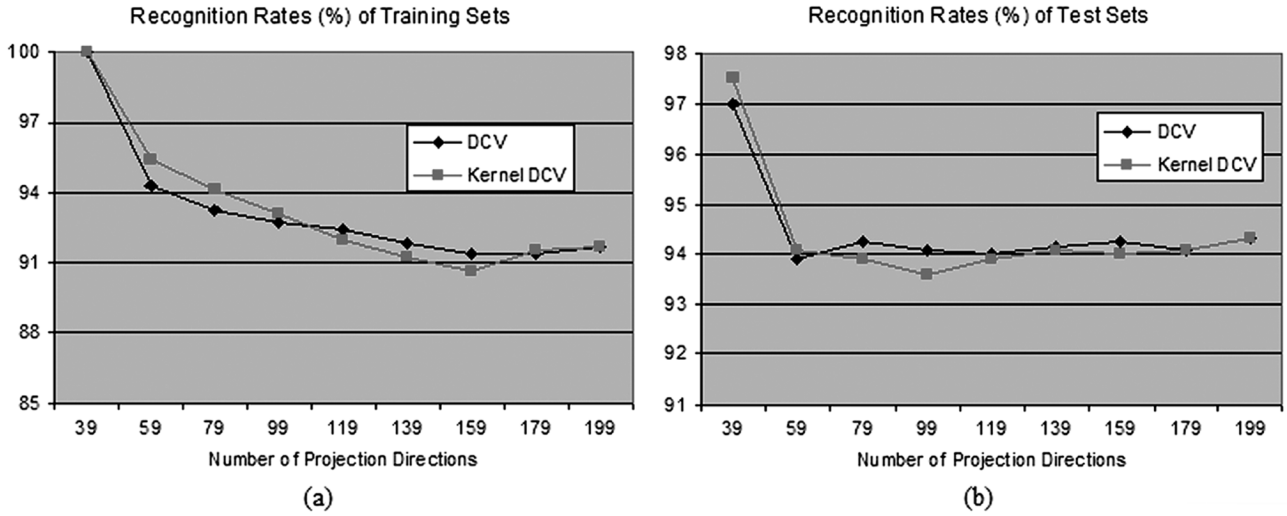


Fig. 5. Recognition rates as a function of projection vectors that are used for feature extraction: (a) training set results and (b) test set results.

the performance of the proposed method, we performed several tests. First, we compared the proposed method with methods that use projection directions from outside the optimal discriminant subspace. The proposed method outperformed other kernel feature extraction 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 spanning the optimal discriminant subspace. We then used these new vectors for feature extraction. However, this process degraded the performance of the proposed method. The experimental test results also show that the generalization ability of the proposed method is comparable to all tested kernel approaches. Finally, the fact that the test sample feature vectors are compared only to the discriminative common vectors, as opposed to all training set sample feature vectors, makes the proposed method ideal for real-time applications.

APPENDIX I

Proof of Lemma 1: By definition, a vector $u \in \Re^d$ is in $N(S_T)$ if $S_T u = 0$. Let μ be the mean vector of the samples in the training set, $1_M \in \Re^{M \times M}$ be the matrix with all elements equal to M^{-1}, and $X \in \Re^{d \times M}$ 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

$$0 = \sum_{i=1}^{C} \sum_{m=1}^{N_i} u^T (x_m^i - \mu)(x_m^i - \mu)^T u = u^T X (I - 1_M)(I - 1_M)^T X^T u = \|(I - 1_M) X^T u\|^2 \quad (33)$$

where $\| \cdot \|$ denotes the Euclidean norm. Thus, (33) 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

$$(x_m^i)^T u_k = \mu^T u_k, \quad i = 1, \ldots, C, \quad m = 1, \ldots, N_i, \quad k = r_T + 1, \ldots, d. \quad (34)$$

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

$$x = \sum_{k=r_T+1}^{d} \langle x_m^i, u_k \rangle u_k = \sum_{k=r_T+1}^{d} \langle \mu, u_k \rangle u_k, \quad i = 1, \ldots, C, \quad m = 1, \ldots, N_i \quad (35)$$

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

Statistical Significance Test Involving Differences of Means and Proportions: Consider that two classes X_1 and X_2 come from two populations with means $\bar{X}_1, \bar{X}_2$ and standard deviations σ_1, σ_2 obtained by N_1 and N_2 trials, respectively. Then, we have to decide between two hypotheses

$$\mathrm{H}_0 : \mu_1 = \mu_2$$
$$\mathrm{H}_1 : \mu_1 \neq \mu_2.$$

Under hypothesis H_0, both classes come from the same population. The mean and standard deviation of the difference in means are given by

$$\mu_{\bar{X}_1 - \bar{X}_2} = 0 \text{ and } \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\sigma_1^2/N_1 + \sigma_2^2/N_2}.$$

Then

$$z = (\bar{X}_1 - \bar{X}_2)/\sigma_{\bar{X}_1 - \bar{X}_2}.$$

For a two-tailed test, the results are significantly different at a 0.05 level if z lies outside the range -1.96 to 1.96. Hence we conclude that the difference in performance of the two methods is significantly different if z lies outside the range -1.96 to 1.96 with a significance level of 0.05.

ACKNOWLEDGMENT

The authors would like to thank C. Liu for providing the SVM classifier algorithms and for his assistance in the application of the SVM classifier.

REFERENCES

[1] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugen.*, vol. 7, pp. 179–188, 1936.

[2] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York: Academic, 1990, pp. 31, 34, 39–40, 220-221.

[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] W. Zhao, R. Chellappa, and A. Krishnaswamy, "Discriminant analysis of principal components for face recognition," in *Proc. 3rd IEEE Int. Conf. Automat. Face Gesture Recognit.*, Apr. 1998, pp. 336–341.

[5] 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 Recognit.*, vol. 24, pp. 317–324, 1991.

[6] 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.

[7] C. Lee and D. Landgrebe, "Analyzing high-dimensional multispectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 31, no. 7, pp. 792–800, Jul. 1993.

[8] Y. Bing, J. Lianfu, and C. Ping, "A new LDA-based method for face recognition," in *Proc. 16th Int. Conf. Pattern Recognit.*, Aug. 2002, vol. 1, pp. 168–171.

[9] R. Huang, Q. Liu, H. Lu, and S. Ma, "Solving the small size problem of LDA," in *Proc. 16th Int. Conf. Pattern Recognit.*, Aug. 2002, vol. 3, pp. 29–32.

[10] 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 Anal. Applicat.*, vol. 6, pp. 47–54, Apr. 2003.

[11] 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 Recognit.*, vol. 33, pp. 1713–1726, 2000.

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

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

[14] 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, 1999, pp. 41–48.

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

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

[17] B. Schölkopf, "Support Vector Learning," Ph.D. dissertation, Informatik der Technischen Universität, Berlin, Germany, 1997.

[18] 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.

[19] C. W. Therrien, "Eigenvalue properties of projection operators and their applications to the subspace method of feature extraction," *IEEE Trans. Comput.*, vol. COM-24, no. 9, pp. 944–948, Sep. 1975.

[20] J. Xu and L. Zikatanov, "The method of alternating projections and the method of subspace corrections in Hilbert space," *J. Amer. Math. Soc.*, vol. 15, pp. 573–597, 2002.

[21] J. Lu, 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.

[22] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Regularized discriminant analysis for the small sample size problem in face recognition," *Pattern Recognit. Lett.*, vol. 24, pp. 3079–3087, 2003.

[23] J. Lu, 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.

[24] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, Mar. 2002.

[25] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowledge Discovery*, vol. 2, pp. 121–167.

[26] K. Tsuda, "Subspace classifier in reproducing kernel Hilbert space," in *Proc. Int. Joint Conf. Neural Netw.*, 1999, vol. 5, pp. 3454–3457.

[27] Y. Hamamoto, T. Kanaoka, and S. Tomita, "On a theoretical comparison between the orthonormal disriminant vector method and discriminant analysis," *Pattern Recognit.*, vol. 26, pp. 1863–1867, 1993.

[28] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Oxford Univ. Press, 1995, pp. 11–14.

[29] A. Webb, *Statistical Pattern Recognition*, 2nd ed. New York: Wiley, 2002, pp. 20–21.

[30] L. O. Jimenez and D. A. Landgrebe, "Supervised classification in high dimensional space: Geometrical, statistical, and asymptotical properties of multivariate data," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 28, no. 1, pp. 39–54, Feb. 1998.

[31] F. Perez-Cruz and O. Bousquet, "Kernel methods and their potential use in signal processing," *IEEE Signal Process. Mag.*, vol. 21, no. 3, pp. 57–65, May 2004.

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

[33] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.

[34] M. R. Spiegele, *Theory and Problems of Statistics*. New York: McGraw-Hill, 1961, p. 181.

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

[36] K. I. Kim, K. Jung, and H. J. Kim, "Face recognition using kernel principal component analysis," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 40–42, Feb. 2002.

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

[38] B. Schölkopf and A. J. Smola, *Learning with Kernels*. Cambridge, MA: MIT Press, 2002, pp. 37–39.

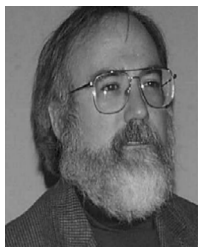
[39] J. Peltonen and S. Kaski, "Discriminative components of data," *IEEE Trans. Neural Netw.*, vol. 16, no. 1, pp. 68–83, Jan. 2005.
[40] M. M. S. Lee, S. S. Keerthi, C. J. Ong, and D. DeCoste, "An efficient method for computing leave-one-out error in support vector machines with Gaussian kernels," *IEEE Trans. Neural Netw.*, vol. 15, no. 3, pp. 750–757, May 2004.
[41] 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.

Hakan Cevikalp (S'01–M'06) received the M.S. degree from the Department of Electrical and Electronics Engineering, Eskisehir Osmangazi Universitesi, Eskisehir, Turkey, in 2001 and the Ph.D. degree from the Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN, in 2005.

He is currently a Research Assistant with the Electrical and Electronics Engineering Department, Eskisehir Osmangazi University. His research interests include pattern recognition, neural networks, image and signal processing, optimization, and computer vision.

Marian Neamtu received the M.S. degree in mechanical engineering from the Slovak Technical University, Slovakia, in 1988 and the Ph.D. degree in mathematics from the University of Twente, The Netherlands, in 1991.

Currently, he is an Associate Professor of mathematics at Vanderbilt University, Nashville, TN. His main research interests are in numerical analysis approximation theory, computer-aided geometric design, and related areas of applied mathematics.



Mitch Wilkes (S'79–M'87) received the B.S.E.E. degree from Florida Atlantic University, Boca Raton. He received the M.S.E.E. and Ph.D. degrees in electrical engineering from Georgia Institute of Technology, Atlanta, in 1984 and 1987, respectively.

He is an Associate Professor of electrical engineering and computer science at the School of Engineering, Vanderbilt University, Nashville, TN. He is also an Assistant Director of the Center for Intelligent Systems and the Assistant Director of the Intelligent Robotics Laboratory. He has more than 90 publications. His research interests include intelligent robotics and control, signal processing, and image processing.