Nonlinear Discriminative Common Vector Method

Hakan CEVIKALP[1], Marian NEAMTU[2], and Mitch WILKES[1]
[1] Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, Tennessee, USA
[2] Department of Mathematics, Vanderbilt University, Nashville, Tennessee, USA
hakan.cevikalp@vanderbilt.edu, neamtu@math.vanderbilt.edu, mitch.wilkes@vanderbilt.edu

ABSTRACT

In this paper we propose a new method called the Kernel Discriminative Common Vector (Kernel DCV) method. Firstly the original input space is mapped nonlinearly to a higher-dimensional feature space through a kernel mapping. Then, the linear Discriminative Common Vector (DCV) method is applied in the transformed space. The proposed method employs the projection vectors from the null space of the within-class scatter matrix of the transformed samples for feature extraction. The same discriminative common vector for all samples in each class is obtained after feature extraction. Therefore, a 100% recognition rate is always guaranteed for the training set samples. The experiments on the test sets also show that the generalization ability of the proposed method compares favorably with the other 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.

Keywords: Common vectors, discriminative common vectors, face recognition, feature extraction, kernel functions.

1. INTRODUCTION

Feature extraction has been one of the most fundamental issues of pattern recognition. In feature extraction problems, the aim is to select the variables that contain the most discriminatory information. Most of the feature extraction methods, such as Principal Component Analysis (PCA), the Fisher's Linear Discriminant Analysis (FLDA) [1], the Direct-LDA method [2], the PCA+Null Space method [3], and the DCV method [4], 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. Also linear feature extraction techniques are often used as pre-processors before more complex nonlinear classifiers. However, sometimes linear methods may not provide sufficient nonlinear discriminant power for classification of nonlinearly distributed classes. Thus, the kernel approaches, such as the Kernel PCA [5], the Kernel Fisher's Discriminant Analysis (Kernel FDA) [6], and the Kernel Generalized Discriminant Analysis (Kernel GDA) [7], have been proposed to overcome this limitation. In these methods, 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.

In this paper we propose a new method called the Kernel DCV method, which applies the linear DCV method in the nonlinearly transformed higher-dimensional space. The Kernel DCV is based on a novel variation of the FLDA criterion described in the next section. The proposed method extracts optimal features for discrimination in the nonlinearly transformed higher-dimensional space since the modified FLDA criterion, $J_{MFLDA}(W_{opt}) = \arg\max_W \frac{|W^T S_B W|}{|W^T S_T W|}$, attains its maximum. Here S_B represents the between-class scatter matrix of the training set samples and S_T represents the total scatter matrix of the samples.

The remainder of the paper is organized as follows. In Section 2, the DCV and the PCA+Null Space methods are reviewed. In Section 3, the Kernel DCV method is introduced. In Section 4, we describe the data sets and experimental results. Finally, our conclusions are formulated in Section 5.

2. OPTIMAL PROJECTION VECTORS FOR FEATURE ETRACTION

The modified FLDA criterion, $J_{MFLDA}(W_{opt}) = \arg\max_W \frac{\left|W^T S_B W\right|}{\left|W^T S_T W\right|}$, attains its maximum, 1, in the special case of $w^T S_W w = 0$ and $w^T S_B w \neq 0$, for all $w \in R^d \setminus \{0\}$, where S_W is the within-class scatter matrix. However, a projection vector *w*, satisfying the above conditions, does not necessarily maximize the between-class scatter. In this case, a better criterion is given in [1], namely

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

Therefore, the optimal projection vectors come from the null space of the within-class scatter matrix, S_W. To find the orthonormal optimal projection vectors *w* in the null space of S_W, we project the training set samples onto the null space of S_W and then obtain the projection vectors by performing PCA. After this operation we obtain a set of orthonormal vectors, which is a basis for a space, which we called the *optimal discriminant subspace*. This subspace is the intersection of the null space of the within-class scatter matrix S_W and the range

space of the total scatter matrix S_T. The criterion given in (1) attains its maximum for any orthonormal vector set that spans 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 [4]. In this section the optimal discriminant subspace is explained in detail first and then the DCV and the PCA+Null Space methods are reviewed.

2.1 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

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

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

and

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

where μ is the mean of all samples, and μ_i is the mean of samples in the *i*-th class. The matrices $A_W \in R^{dxM}$, $A_B \in R^{dxC}$, and $A_T \in R^{dxM}$ are defined as

$$A_W = [x_1^1 - \mu_1 \ldots x_{N_1}^1 - \mu_1 \; x_1^2 - \mu_2 \ldots x_{N_C}^C - \mu_C], \quad (5)$$

$$A_B = [\sqrt{N_1}(\mu_1 - \mu) \quad \ldots \quad \sqrt{N_C}(\mu_C - \mu)], \quad (6)$$

and

$$A_T = [x_1^1 - \mu \ldots x_{N_1}^1 - \mu \; x_1^2 - \mu_2 \ldots x_{N_C}^C - \mu]. \quad (7)$$

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, which form orthonormal bases for the null spaces of the corresponding scatter matrices. As explained previously, if the projection directions are chosen from the null space of S_W, the modified FLDA criterion attains its maximum, 1. Therefore, optimal projection vectors can be obtained by applying PCA to the samples which are projected onto the null space of S_W. The fact that the final 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 ($k = r_T + 1, \ldots, d$, 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 produce a unique common vector such that

$$x = \overline{U}\overline{U}^T x_m^i, \quad i = 1, \ldots, C, \qquad m = 1, \ldots, N_i, \quad (8)$$

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

Proof: See [8].

This lemma shows that the null space of 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 for discrimination will be the intersection of the null space of S_W and the range space of S_T.

There are basically two approaches to find the optimal projection vectors that span the optimal discriminant subspace. This is a result of the fact that the projection matrices (also called orthogonal projection operators) 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 $N(S_W)$ and $R(S_T)$ respectively. In this case, the projection matrix of the intersection $N(S_W) \cap R(S_T)$ is found by the equation

$$P_{opt} = P^{(1)}P^{(2)} = P^{(2)}P^{(1)}, \quad (9)$$

where P_{opt} is the projection matrix of the optimal discriminant subspace.

Theorem 1: Let $H = N(S_W) \cap R(S_T)$ represents the intersection of the null space $N(S_W)$ of the within-class scatter matrix and the range space $R(S_T)$ of the total scatter matrix. Then, the projection matrices $P^{(1)}$ and $P^{(2)}$ of the subspaces $N(S_W)$ and $R(S_T)$ commute and the projection matrix P_{opt} of the intersection *H* can be found by the following formula:

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

Proof: See [8].

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.

2.2 The Discriminative Common Vector Method

The DCV method is a computationally efficient and stable method for finding the optimal projection vectors. In this

method we first project the training set samples onto $N(S_W)$ and then we perform PCA in the transformed space. There are two different algorithms that accomplish this task. We now recall the first algorithm which uses the range space of S_W [4]. This algorithm can be summarized as follows:

Step 1: Projection of the training set samples onto $N(S_W)$:

i) Compute the nonzero eigenvalues and corresponding eigenvectors α_k of S_W by using the matrix $A_W{}^T A_W \in R^{MxM}$, where $S_W = A_W A_W{}^T \in R^{dxd}$ and A_W is given by (5) [4]. Set $Q = [\alpha_1 \quad ... \quad \alpha_r]$, where *r* is the rank of S_W.

ii) Project the training set samples onto $N(S_W)$ by

$$x_{com}^i = x_m^i - QQ^T x_m^i, \quad i = 1,...,C, \quad m = 1,...,N_i. \quad (10)$$

In this way, it turns out, we obtain the same unique vector that represents each class for all the samples in that class, i.e., the vector on the right-hand side of (10) is independent of the sample index *m* [4]. These vectors are called the *common vectors*.

Step 2: Obtaining the optimal projection vectors w_k :

The optimal projection vectors are those that maximize the total scatter across all common vectors. Therefore, the optimal projection vectors can be obtained by computing the nonzero eigenvalues and the corresponding eigenvectors of the matrix

$$S_{com} = \sum_{i=1}^{C} (x_{com}^i - \mu_{com})(x_{com}^i - \mu_{com})^T = A_{com} A_{com}^T, \quad (11)$$

where μ_{com} is the mean of all common vectors. The matrix $A_{com} \in R^{dxC}$ is defined as

$$A_{com} = [x_{com}^1 - \mu_{com} \quad ... \quad x_{com}^C - \mu_{com}]. \quad (12)$$

The nonzero eigenvalues and the corresponding eigenvectors w_k can be computed easily by using the matrix $A_{com}{}^T A_{com} \in R^{CxC}$ instead of $S_{com} \in R^{dxd}$. Then, use these eigenvectors to form the projection vector matrix $W = [w_1 \quad ... \quad w_{r_c}]$, which will be used to obtain the feature vectors of the samples. Here, $r_C \le C-1$ is the rank of S_{com}.

Since the optimal projection vectors w_k come from $N(S_W)$, it follows that when the training set samples x_m^i of the *i*-th class are projected onto the linear span of the projection vectors w_k, the feature vector $\Omega_i = [< x_m^i, w_1 > \quad ... \quad < x_m^i, w_{r_c} >]^T$ of the projection coefficients $< x_m^i, w_k >$ will also be independent of the sample index *m*. Thus, for each class we have $\Omega_i = W^T x_m^i$. The fact that the vectors Ω_i ($i = 1,...,C$) do not depend on the index *m* guarantees 100% accuracy in the recognition of the samples in the training set. The vectors Ω_i are called the *discriminative common vectors*.

To recognize a test sample, x_{test}, the feature vector of the test sample is found by the equation

$$\Omega_{test} = W^T x_{test} \quad (13)$$

and Ω_{test} is compared with the discriminative common vector Ω_i of each class using the Euclidean distance. The discriminative common vector found to be the closest to Ω_{test} is used to identify the test sample.

2.3 The PCA+Null Space Method

In this method, in order to obtain optimal projection vectors, the training set samples are first projected onto the range space of S_T through PCA, and then the vectors that span the null space of the new within-class scatter matrix in the tansformed space are computed. The algorithm can be summarized as follows:

Step 1: Compute the nonzero eigenvalues and corresponding eigenvectors u_k of S_T by using the matrix $A_T{}^T A_T \in R^{MxM}$, where $S_T = A_T A_T{}^T \in R^{dxd}$ and A_T is given by (7). Set $U = [u_1 \quad ... \quad u_{r_T}]$, where r_T is the rank of S_T. Then transform the training set samples as $U^T x_m^i$. Compute the new within-class scatter matrix in the transformed space by

$$\widetilde{S}_W = U^T S_W U. \quad (14)$$

Step 2: Find the orthonormal vector set that spans the null space of $\widetilde{S}_W$. This can be done through an eigen-decomposition of $\widetilde{S}_W$. The eigenvectors corresponding to the zero eigenvalues of $\widetilde{S}_W$ span the null space of $\widetilde{S}_W$. Let *V* be the matrix whose columns are the computed eigenvectors such that $V^T \widetilde{S}_W V = 0$. Then, the final transformation matrix will be $\widetilde{W} = UV$.

The optimal projection vector matrix $\widetilde{W}$ obtained by the PCA+Null Space method and the optimal projection vector matrix *W* obtained by the DCV method span the same optimal discriminant subspace and hence the matrices obey the equation $WW^T = \widetilde{W}\widetilde{W}^T$.

3. 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

$$S_W^\Phi = \sum_{i=1}^{C}\sum_{m=1}^{N_i}(\Phi(x_m^i)-\mu_i^\Phi)(\Phi(x_m^i)-\mu_i^\Phi)^T = (\Phi-\Phi G)(\Phi-\Phi G)^T, \quad (15)$$

$$S_B^\Phi = \sum_{i=1}^{C} N_i(\mu_i^\Phi-\mu^\Phi)(\mu_i^\Phi-\mu^\Phi)^T = (\Phi U-\Phi L)(\Phi U-\Phi L)^T, \quad (16)$$

and

$$S_T^\Phi = \sum_{i=1}^{C}\sum_{m=1}^{N_i}(\Phi(x_m^i)-\mu^\Phi)(\Phi(x_m^i)-\mu^\Phi)^T = (\Phi-\Phi 1_M)(\Phi-\Phi 1_M)^T = S_W^\Phi + S_B^\Phi, \quad (17)$$

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 $G_i \in R^{N_i xN_i}$ 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 $u_i \in R^{N_i x1}$ 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

$$\widetilde{K} = K - 1_M K - K 1_M + 1_M K 1_M \in R^{MxM} = P\Lambda P^T \quad (18)$$

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

$$K^{ij} = (k_{mn}^{ij})_{\substack{m=1,...,N_i\\ n=1,...,N_j}} = <\Phi(x_m^i), \Phi(x_n^j)> = k(x_m^i, x_n^j)_{\substack{m=1,...,N_i\\ n=1,...,N_j}}. \quad (19)$$

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

$$\widetilde{S}_T^\Phi = ((\Phi-\Phi 1_M)P\Lambda^{-1/2})^T S_T^\Phi (\Phi-\Phi 1_M)P\Lambda^{-1/2} = \Lambda^{-1/2}P^T P\Lambda P^T P\Lambda^{-1/2} = \Lambda \quad (20)$$

and

$$\widetilde{S}_W^\Phi = ((\Phi-\Phi 1_M)P\Lambda^{-1/2})^T S_W^\Phi (\Phi-\Phi 1_M)P\Lambda^{-1/2} = \Lambda^{-1/2}P^T \widetilde{K}_W \widetilde{K}_W{}^T P\Lambda^{-1/2}, \quad (21)$$

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

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,

$$V^T \widetilde{S}_W^\Phi V = 0. \quad (22)$$

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,

$$V^T\widetilde{S}_B^\Phi V = V^T\widetilde{S}_T^\Phi V = V^T\Lambda V = L\widetilde{\Lambda}L^T. \quad (23)$$

Then the final projection matrix *W* will be

$$W = (\Phi-\Phi 1_M)P\Lambda^{-1/2}VL. \quad (24)$$

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.

4. EXPERIMENTAL RESULTS

In our experiments we used the ORL (Olivetti-Oracle Research Lab) face database [9] to test the proposed method. The ORL face database contains *C*=40 individuals with 10 images per person. The images are taken at different time instances with

different lighting conditions (slightly), facial expressions, and facial details. All individuals are in up-right, frontal position (with tolerance for some side movement). The size of the each image is 92x112. Some individuals from the ORL face database are shown in Fig. 1.

An appropriate selection of kernel functions for special tasks is still an open problem [10]. We have used polynomial kernels $k(x,y)=(<x,y>)^k$, with degrees $k=2,3$, and the Gaussian kernel $k(x,y)=\exp(-\|x-y\|^2/\gamma)$ in our experiments. The parameter γ was chosen as 1.06e8 based on empirical observations. For the linear PCA and the Kernel PCA methods the most significant eigenvectors were chosen such that corresponding eigenvalues contain 95 % of the total energy. A nearest-neighbor algorithm was employed using the Euclidean distance for classification except for the methods that produce the discriminative common vectors in which case the feature vector of the test sample is only compared to the discriminative common vectors 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.

We selected randomly five samples from each class for training and the rest were used for testing. Thus, a training set of *M*=200 images and a test set of 200 images were created. We have not applied any preprocessing to the images. Then recognition rates were computed. This process was repeated six times and the recognition rates were found by averaging the recognition rates in each run. The computed recognition rates for the linear methods and the Kernel methods are shown in Table I. The best recognition result was obtained by the DCV method among the linear methods and similarly the best recognition results for all polynomial kernels with different degrees were obtained by the Kernel DCV method among the kernel methods. There is not a significant difference between recognition rates of the DCV and the Kernel DCV methods for this database. However, the recognition rates of the Kernel DCV method may be improved for different kernels that fulfill Mercer's theorem [11]. But, 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 best recognition rates were obtained for the Gaussian kernel for all the Kernel methods. The supervised methods (all methods except PCA and the Kernel PCA methods) typically outperformed the unsupervised methods (PCA and the Kernel PCA 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 decreased, which shows that the second-order data correlation is usually enough for a good recognition performance.

Fig.1. Three individuals from the ORL face database.

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 [12], 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 39 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 $t=199$ projection vectors, and we computed the corresponding recognition rates. The results for the training and test sets are illustrated in Fig. 2. 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.

TABLE I
Recognition Rates of the ORL Face Database

Linear Methods	**Recognition Rates (%) & Standard Deviations**		
PCA	93.66, $\sigma=2.01$		
FLDA	93.33, $\sigma=2.62$		
Direct-LDA	96.58, $\sigma=1.39$		
DCV	97, $\sigma=1.41$		
Kernel Methods	**Recognition Rates (%) & Standard Deviations**		
	k = 2	k = 3	k = 4
Kernel PCA	93.33, $\sigma=1.21$	92.75, $\sigma=1.40$	93.75, $\sigma=1.25$
Kernel FDA	96.33, $\sigma=1.57$	95.41, $\sigma=1.59$	96.50, $\sigma=1.18$
Kernel GDA	94.16, $\sigma=0.98$	93.58, $\sigma=1.20$	96.66, $\sigma=0.93$
Kernel DCV	97, $\sigma=1.67$	95.91, $\sigma=1.88$	97.50, $\sigma=0.94$

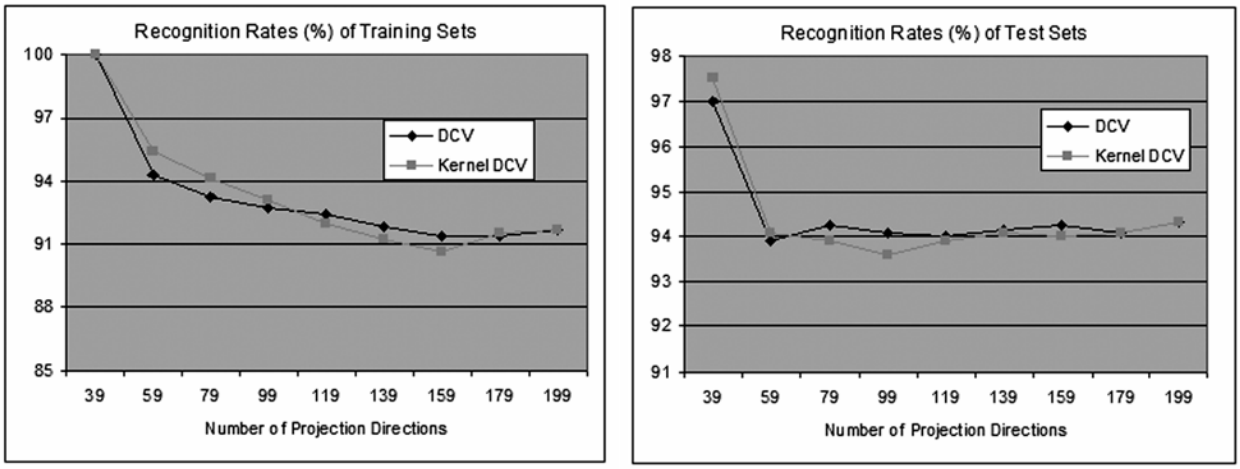

Fig. 2. Recognition rates (%) as a function of projection vectors that are used for feature extraction

5. CONCLUSION

In this paper we proposed a new method that uses the kernel functions for recognition. We first showed that the optimal projection vectors come from the intersection space of the null space of the within-class scatter matrix S_w and the total scatter matrix S_T. Then we proposed an algorithm to find 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. Test results show that the generalization ability of the proposed method compares favorably with the other kernel approaches. Also the fact that the test sample feature vectors are compared to only the discriminative common vectors, instead of all training set sample feature vectors, makes the proposed method ideal for real-time applications.

6. REFERENCES

[1] 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.

[2] 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.

[3] 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.

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

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

[6] 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.

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

[8] H. Cevikalp, M. Neamtu, and M. Wilkes, "Discriminative common vector method with kernels," *IEEE Transaction on Neural Networks*, in review.

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

[10] 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.

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

[12] 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.

[13] H. Cevikalp and M. Wilkes, "Face Recognition by Using Discriminative Common Vectors," 17th Int. Conf. on Pattern Recognition (ICPR), Cambridge, UK, August 2004.