DISCRIMINATIVE COMMON VECTORS FOR FACE RECOGNITION

Hakan Cevikalp[1], Marian Neamtu[2], Mitch Wilkes[1], and Atalay Barkana[3]

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

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

[3]Department of Electrical and Electronics Engineering, Osmangazi University, Eskisehir, Turkey.

CORRESPONDENCE ADDRESS:

Prof. Mitch Wilkes

Department of Electrical Engineering and Computer Science,

Vanderbilt University, Nashville, Tennessee, USA

Tel: (615) 343-6016

Fax: (615) 322-7062

e-mail: mitch.wilkes@vanderbilt.edu

Abstract

In face recognition tasks, the dimension of the sample space is typically larger than the number of the samples in the training set. As a consequence, the within-class scatter matrix is singular and the Linear Discriminant Analysis (LDA) method cannot be applied directly. This problem is known as the "small sample size" problem. In this paper, we propose a new face recognition method called the Discriminative Common Vector method based on a variation of Fisher's Linear Discriminant Analysis for the small sample size case. Two different algorithms are given to extract the discriminative common vectors representing each person in the training set of the face database. One algorithm uses the within-class scatter matrix of the samples in the training set while the other uses the subspace methods and the Gram-Schmidt orthogonalization procedure to obtain the discriminative common vectors. Then the discriminative common vectors are used for classification of new faces. The proposed method yields an optimal solution for maximizing the modified Fisher's Linear Discriminant criterion given in the paper. Our test results show that the Discriminative Common Vector method is superior to other methods in terms of recognition accuracy, efficiency, and numerical stability.

Index Terms: Common Vectors, Discriminative Common Vectors, Face Recognition, Fisher's Linear Discriminant Analysis, Principal Component Analysis, Small Sample Size, Subspace Methods.

I. INTRODUCTION

Recently, due to military, commercial, and law enforcement applications, there has been much interest in automatically recognizing faces in still and video images. This research spans several disciplines such as image processing, pattern recognition, computer vision and neural networks. The data come from a wide variety of sources. One group of sources is the relatively controlled format images such as passports, credit cards, photo ID's, driver's licenses, and mug shots. A more challenging class of application imagery includes real-time detection and recognition of faces in surveillance video images, which present additional constraints in terms of speed and processing requirements [1].

Face recognition can be defined as the identification of individuals from images of their faces by using a stored database of faces labeled with people's identities. This task is complex and can be decomposed into the smaller steps of detection of faces in a cluttered background, localization of these faces followed by extraction of features from the face regions, and finally recognition and verification [2]. It is a difficult problem as there are numerous factors such as 3-D pose, facial expression, hair style, make up, and so on, which affect the appearance of an individual's facial features. In addition to these varying factors, lighting, background, and scale changes make this task even more challenging. Additional problematic conditions include noise, occlusion, and many other possible factors.

Many methods have been proposed for face recognition within the last two decades [1], [3]. Among these methods, appearance-based approaches operate directly on images or appearances of face objects, and process the images as two-dimensional (2-D) holistic patterns. In these approaches, a two-dimensional image of size w by h pixels is represented by a vector in a wh-dimensional space. Therefore, each facial image corresponds to a point in this space. This space

is called the sample space or the image space, and its dimension typically is very high [4]. However, since face images have similar structure, the image vectors are correlated, and any image in the sample space can be represented in a lower-dimensional subspace without losing a significant amount of information. The Eigenface method has been proposed for finding such a lower-dimensional subspace [5]. The key idea behind the Eigenface method, which uses Principal Component Analysis (PCA), is to find the best set of projection directions in the sample space that will maximize the total scatter across all images such that $J_{PCA}(W_{opt}) = \arg\max_{W} |W^T S_T W|$ is maximized. Here S_T is the total scatter matrix of the training set samples, and W is the matrix whose columns are the orthonormal projection vectors. The projection directions are also called the eigenfaces. Any face image in the sample space can be approximated by a linear combination of the significant eigenfaces. The sum of the eigenvalues that correspond to the eigenfaces not used in reconstruction gives the mean square error of reconstruction. This method is an unsupervised technique, since it does not consider the classes within the training set data. In choosing a criterion that maximizes the total scatter, this approach tends to model unwanted within-class variations such as those resulting from differences in lighting, facial expression, and other factors [6], [7]. Additionally, since the criterion does not attempt to minimize within-class variation, the resulting classes may tend to have more overlap than other approaches. Thus, the projection vectors chosen for optimal reconstruction may obscure the existence of the separate classes.

The Linear Discriminant Analysis (LDA) method is proposed in [6] and [7]. This method overcomes the limitations of the Eigenface method by applying the Fisher's Linear Discriminant criterion. This criterion tries to maximize the ratio $J_{FLD}(W_{opt}) = \arg\max_{W} \frac{|W^T S_B W|}{|W^T S_W W|}$, where S_B is

the between-class scatter matrix, and S_W is the within-class scatter matrix. Thus, by applying this method, we find the projection directions that on one hand maximize the Euclidean distance between the face images of different classes and on the other minimize the distance between the face images of the same class. This ratio is maximized when the column vectors of the projection matrix W are the eigenvectors of $S_W^{-1}S_B$. In face recognition tasks, this method cannot be applied directly since the dimension of the sample space is typically larger than the number of samples in the training set. As a consequence, S_W is singular in this case. This problem is also known as the "small sample size problem" [8].

In the last decade numerous methods have been proposed to solve this problem. Tian *et al*. [9] used the Pseudo-Inverse method by replacing S_W^{-1} with its pseudo-inverse. The Perturbation method is used in [2] and [10], where a small perturbation matrix Δ is added to S_W in order to make it nonsingular. Cheng *et al*. [11] proposed the Rank Decomposition method based on successive eigen-decompositions of the total scatter matrix S_T and the between-class scatter matrix S_B. However, the above methods are typically computationally expensive since the scatter matrices are very large (e.g., images of size 256 by 256 yield scatter matrices of size 65,536 by 65,536). Swets and Weng [7] proposed a two stage PCA+LDA method, also known as the Fisherface method, in which PCA is first used for dimension reduction so as to make S_W nonsingular before the application of LDA. In this method the final optimal projection vector matrix becomes $W_{opt} = W_{PCA}W_{FLD}$, where $W_{PCA} = \arg\max_W |W^T S_T W|$, and $W_{FLD} = \arg\max_W \frac{|W^T W_{PCA}^T S_B W_{PCA} W|}{|W^T W_{PCA}^T S_W W_{PCA} W|}$. However, in order to make S_W nonsingular, some directions corresponding to the small eigenvalues of S_T are thrown away in the PCA step. Thus,

applying PCA for dimensionality reduction has the potential to remove dimensions that contain discriminative information [12]-[16]. Chen *et al*. [17] proposed the Null Space method based on the modified Fisher's Linear Discriminant criterion, $J_{MFLD}(W_{opt}) = \arg\max_{W} \frac{\left|W^T S_B W\right|}{\left|W^T S_T W\right|}$. This method was proposed to be used when the dimension of the sample space is larger than the rank of the within-class scatter matrix, S_W. It has been shown that the original Fisher's Linear Discriminant criterion can be replaced by the modified Fisher's Linear Discriminant criterion in the course of solving the discriminant vectors of the optimal set in [18]. In this method, all image samples are first projected onto the null space of S_W, resulting in a new within-class scatter that is a zero matrix. Then, PCA is applied to the projected samples to obtain the optimal projection vectors. Chen *et al*. also proved that by applying this method, the modified Fisher's Linear Discriminant criterion attains its maximum. However, they did not propose an efficient algorithm for applying this method in the original sample space. Instead, a pixel grouping method is applied to extract geometric features and reduce the dimension of the sample space. Then they applied the Null Space method in this new reduced space. In our experiments, we observed that the performance of the Null Space method depends on the dimension of the null space of S_W in the sense that larger dimension provides better performance. Thus, any kind of pre-processing that reduces the original sample space should be avoided.

Another novel method, the PCA+Null Space method was proposed by Huang *et al*. in [15] for dealing with the small sample size problem. In this method, at first, PCA is applied to remove the null space of S_T, which contains the intersection of the null spaces of S_B and S_W. Then, the optimal projection vectors are found in the remaining lower-dimensional space by using the Null Space method. The difference between the Fisherface method and the PCA+Null Space method

is that for the latter, the within-class scatter matrix in the reduced space is typically singular. This occurs because all eigenvectors corresponding to the nonzero eigenvalues of S_T are used for dimension reduction. Yang *et al*. applied a variation of this method in [16]. After dimension reduction, they split the new within-class scatter matrix, $\widetilde{S}_W = P_{PCA}^T S_W P_{PCA}$ (where P_{PCA} is the matrix whose columns are the orthonormal eigenvectors corresponding to the nonzero eigenvalues of S_T), into its null space $N(\widetilde{S}_W) = span\{\xi_{r+1},...,\xi_t\}$ and orthogonal complement (i.e., range space) $R(\widetilde{S}_W) = span\{\xi_1,...,\xi_r\}$ (where *r* is the rank of S_W, and $t = rank(S_T)$ is the dimension of the reduced space). Then, all the projection vectors that maximize the between-class scatter in the null space are chosen. If, according to some criterion, more projection vectors are needed, the remaining projection vectors are obtained from the range space. Although the PCA+Null Space method and the variation proposed by Yang *et al*., use the original sample space, applying PCA and using all eigenvectors corresponding to the nonzero eigenvalues make these methods impractical for face recognition applications when the training set size is large. This is due to the fact that the computational expense of training becomes very large.

Lastly, the Direct-LDA method is proposed in [12]. This method uses the simultaneous diagonalization method [8]. First, the null space of S_B is removed, and then the projection vectors that minimize the within-class scatter in the transformed space are selected from the range space of S_B. However, removing the null space of S_B by dimensionality reduction will also remove part of the null space of S_W and may result in the loss of important discriminative information [13], [15], [16]. Furthermore, S_B is whitened as a part of this method. This whitening process can be shown to be redundant and therefore should be skipped.

In this paper, a new method is proposed which addresses the limitations of other methods that use the null space of S_W to find the optimal projection vectors. Thus, the proposed method can be only used when the dimension of the sample space is larger than the rank of S_W. The remainder of the paper is organized as follows. In Section II, the Discriminative Common Vector approach is introduced. In Section III, we describe the data sets and experimental results. Finally, we formulate our conclusions in Section IV.

II. DISCRIMINATIVE COMMON VECTOR APPROACH

The idea of common vectors was originally introduced for isolated word recognition problems in the case where the number of samples in each class was less than or equal to the dimensionality of the sample space [19], [20]. These approaches extract the common properties of classes in the training set by eliminating the differences of the samples in each class. A common vector for each individual class is obtained by removing all the features that are in the direction of the eigenvectors corresponding to the nonzero eigenvalues of the scatter matrix of its own class. The common vectors are then used for recognition. In our case instead of using a given class's own scatter matrix, we use the within-class scatter matrix of all classes to obtain the common vectors. We also give an alternative algorithm based on the subspace methods and the Gram-Schmidt orthogonalization procedure to obtain the common vectors. Then, a new set of vectors, called the discriminative common vectors, which will be used for classification are obtained from the common vectors. We introduce algorithms for obtaining the common vectors and the discriminative common vectors below.

A. Obtaining the Discriminative Common Vectors by Using the Null Space of S_W

Let the training set be composed of *C* classes, where each class contains *N* 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=NC* samples in the training set. Suppose that *d>M-C.* In this case, S_W, S_B, and S_T are defined as,

$$S_W = \sum_{i=1}^{C} \sum_{m=1}^{N} (x_m^i - \mu_i)(x_m^i - \mu_i)^T , \tag{1}$$

$$S_B = \sum_{i=1}^{C} N(\mu_i - \mu)(\mu_i - \mu)^T , \tag{2}$$

and

$$S_T = \sum_{i=1}^{C} \sum_{m=1}^{N} (x_m^i - \mu)(x_m^i - \mu)^T = S_W + S_B , \tag{3}$$

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

In the special case where $w^T S_W w = 0$ and $w^T S_B w \neq 0$, for all $w \in R^d \setminus \{0\}$, the modified Fisher's Linear Discriminant criterion attains a maximum. 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 [6] and [13], namely

$$J(W_{opt}) = \arg\max_{|W^T S_W W|=0} |W^T S_B W| = \arg\max_{|W^T S_W W|=0} |W^T S_T W| . \tag{4}$$

To find the optimal projection vectors *w* in the null space of S_W, we project the face samples onto the null space of S_W and then obtain the projection vectors by performing PCA. To do so, vectors that span the null space of S_W must first be computed. However, this task is computationally intractable since the dimension of this null space can be very large. A more

efficient way to accomplish this task is by using the orthogonal complement of the null space of S_W, which typically is a significantly lower-dimensional space.

Let R^d be the original sample space, V be the range space of S_W, and $V^\perp$ be the null space of S_W. Equivalently,

$$V = span\{\alpha_k \mid S_W \alpha_k \neq 0, \qquad k = 1,...,r\} \tag{5}$$

and

$$V^\perp = span\{\alpha_k \mid S_W \alpha_k = 0, \qquad k = r+1,...,d\}, \tag{6}$$

where $r < d$ is the rank of S_W, $\{\alpha_1,....,\alpha_d\}$ is an orthonormal set, and $\{\alpha_1,....,\alpha_r\}$ is the set of orthonormal eigen vectors corresponding to the nonzero eigenvalues of S_W.

Consider the matrices $Q = [\alpha_1 \quad .... \quad \alpha_r]$ and $\overline{Q} = [\alpha_{r+1} \quad .... \quad \alpha_d]$. Since $R^d = V \oplus V^\perp$, every face image $x_m^i \in R^d$ has a unique decomposition of the form

$$x_m^i = y_m^i + z_m^i, \tag{7}$$

where $y_m^i = Px_m^i = QQ^T x_m^i \in V$, $z_m^i = \overline{P}x_m^i = \overline{Q}\,\overline{Q}^T x_m^i \in V^\perp$, and P and $\overline{P}$ are the orthogonal projection operators onto V and $V^\perp$, respectively. Our goal is to compute

$$z_m^i = x_m^i - y_m^i = x_m^i - Px_m^i. \tag{8}$$

To do this, we need to find a basis for V, which can be accomplished by an eigen-analysis of S_W. In particular, the normalized eigenvectors α_k corresponding to the nonzero eigenvalues of S_W will be an orthonormal basis for *V*. The eigenvectors can be obtained by calculating the eigenvectors of the smaller *M* by *M* matrix, $A^T A$, defined such that $S_W = AA^T$, where *A* is a *d* by *M* matrix of the form

$$A = [x_1^1 - \mu_1 \quad .... \quad x_N^1 - \mu_1 \quad x_1^2 - \mu_2 \quad .... \quad x_N^C - \mu_C]. \tag{9}$$

Let λ_k and v_k be the *k*-th nonzero eigenvalue and the corresponding eigenvector of $A^T A$, where $k \le M - C$. Then $\alpha_k = Av_k$ will be the eigenvector that corresponds to the *k*-th nonzero eigenvalue of S_W. The sought-for projection onto $V^\perp$ is achieved by using (8). In this way, it turns out, we obtain the same unique vector for all samples of the same class,

$$x_{com}^i = x_m^i - QQ^T x_m^i = \overline{Q}\,\overline{Q}^T x_m^i, \quad m=1,\ldots,N, \quad i=1,...,C, \tag{10}$$

i.e., the vector on the right-hand side of (10) is independent of the sample index *m*. We refer to the vectors x_{com}^i as the common vectors. The above fact is proved in the following theorem.

Theorem 1: Suppose $\overline{Q}$ is a matrix whose column vectors are the orthonormal vectors that span the null space $V^\perp$ of S_W. Then, the projections of the samples x_m^i of the class *i* onto $V^\perp$ produce a unique common vector x_{com}^i such that

$$x_{com}^i = \overline{Q}\,\overline{Q}^T x_m^i, \quad m=1,\ldots,N, \quad i=1,\ldots,C. \tag{11}$$

Proof: By definition, a vector $\alpha \in R^d$ is in $V^\perp$ if $S_W \alpha = 0$. Let μ_i be the mean vector of the *i*-th class, G be the *N* by *N* matrix whose entries are all N^{-1}, and X^i be the *d* by *N* matrix whose *m*-th column is the sample x_m^i. Thus, multiplying both sides of identity $S_W \alpha = 0$ by α^T and writing

$$S_W = \sum_{i=1}^{C} S_i, \tag{12}$$

with

$$S_i = \sum_{m=1}^{N} (x_m^i - \mu_i)(x_m^i - \mu_i)^T = (X^i - X^i G)(X^i - X^i G)^T, \tag{13}$$

immediately leads to

$$0 = \sum_{i=1}^{C} \alpha^T X^i (I-G)(I-G)^T (X^i)^T \alpha = \sum_{i=1}^{C} \| (I-G)(X^i)^T \alpha \|^2, \quad (14)$$

where ||.|| denotes the Euclidean norm. Thus, (14) holds if $(I-G)(X^i)^T \alpha_k = 0$, or $(X^i)^T \alpha_k = G(X^i)^T \alpha_k$. From this relation we can see that,

$$(x_m^i)^T \alpha_k = (\mu_i)^T \alpha_k, \quad m = 1,...,N, \quad i = 1,...,C, \quad k = r+1,...,d. \quad (15)$$

Thus, the projection of x_m^i onto $V^{\perp}$,

$$x_{com}^i = \sum_{k=r+1}^{d} \langle x_m^i, \alpha_k \rangle \alpha_k = \sum_{k=r+1}^{d} \langle \mu_i, \alpha_k \rangle \alpha_k, \quad (16)$$

is independent of *m*, which proves the theorem.

The theorem states that it is enough to project a single sample from each class. This will greatly reduce the computational burden of the calculations. This computational savings has not been previously reported in the literature.

After obtaining the common vectors x_{com}^i, optimal projection vectors will be those that maximize the total scatter of the common vectors,

$$J(W_{opt}) = \arg\max_{|W^T S_w W|=0} |W^T S_B W| = \arg\max_{|W^T S_w W|=0} |W^T S_T W| = \arg\max_{W} |W^T S_{com} W|, \quad (17)$$

where *W* is a matrix whose columns are the orthonormal optimal projection vectors w_k, and S_{com} is the scatter matrix of the common vectors,

$$S_{com} = \sum_{i=1}^{C} (x_{com}^i - \mu_{com})(x_{com}^i - \mu_{com})^T, \quad i=1,\ldots,C, \quad (18)$$

where μ_{com} is the mean of all common vectors, $\mu_{com} = \frac{1}{C} \sum_{i=1}^{C} x_{com}^i$.

In this case optimal projection vectors w_k can be found by an eigen-analysis of S_{com}. In particular, all eigenvectors corresponding to the nonzero eigenvalues of S_{com} will be the optimal projection vectors. S_{com} is typically a large *d* by *d* matrix and thus we can use the smaller matrix, $A_{com}^T A_{com}$, of size *C* by *C*, to find nonzero eigenvalues and the corresponding eigenvectors of $S_{com} = A_{com} A_{com}^T$, where A_{com} is the *d* by *C* matrix of the form

$$A_{com} = [x_{com}^1 - \mu_{com} \quad .... \quad x_{com}^C - \mu_{com}]. \tag{19}$$

There will be *C*-1 optimal projection vectors since the rank of S_{com} is *C*-1 if all common vectors are linearly independent. If two common vectors are identical, then the two classes which are represented by this vector cannot be distinguished. Since the optimal projection vectors w_k belong to the null space of S_W, it follows that when the image samples x_m^i of the *i*-th class are projected onto the linear span of the projection vectors w_k, the feature vector $\Omega_i = [< x_m^i, w_1 > \quad .... \quad < x_m^i, w_{C-1} >]^T$ of the projection coefficients $< x_m^i, w_k >$ will also be independent of the sample index *m*. Thus, we have

$$\Omega_i = W^T x_m^i, \quad m=1,\ldots,N, \quad i=1,\ldots,C. \tag{20}$$

We call the feature vectors Ω_i *discriminative common vectors*, and they will be used for classification of face images. The fact that Ω_i does not depend on the index *m* in (20) guarantees 100% accuracy in the recognition of the samples in the training set. This guarantee has not been reported in connection with other methods [15], [17].

To recognize a test image x_{test}, the feature vector of this test image is found by

$$\Omega_{test} = W^T x_{test}, \tag{21}$$

which is then 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 image.

Since Ω_{test} is only compared to a single vector for each class, the recognition is very efficient for real-time face recognition tasks. In the Eigenface, the Fisherface, and the Direct-LDA methods, the test sample feature vector Ω_{test} is typically compared to all feature vectors of samples in the training set, making these methods impractical for real-time applications for large training sets.

The above method can be summarized as follows:

Step 1: Compute the nonzero eigenvalues and corresponding eigenvectors of S_W by using the matrix $A^T A$, where $S_W = AA^T$ and A is given by (9). Set $Q = [\alpha_1 \quad .... \quad \alpha_r]$, where r is the rank of S_W.

Step 2: Choose any sample from each class and project it onto the null space of S_W to obtain the common vectors

$$x_{com}^i = x_m^i - QQ^T x_m^i, \qquad m = 1,...,N, \quad i = 1,...,C. \tag{22}$$

Step 3: Compute the eigenvectors w_k of S_{com}, corresponding to the nonzero eigenvalues, by using the matrix $A_{com}^T A_{com}$, where $S_{com} = A_{com} A_{com}^T$ and A_{com} is given in (19). There are at most C-1 eigenvectors that correspond to the nonzero eigenvalues. Use these eigenvectors to form the projection matrix $W = [w_1 \quad .... \quad w_{C-1}]$, which will be used to obtain feature vectors in (20) and (21).

B. Obtaining the Discriminative Common Vectors by Using Difference Subspaces and the Gram-Schmidt Orthogonalization Procedure

To find an orthonormal basis for the range of S_W, the algorithm described above uses the eigenvectors corresponding to the nonzero eigenvalues of the *M* by *M* matrix $A^T A$, where $S_W = AA^T$. Assuming that $rank(S_W) = M - C$, then $l(\frac{4M^3}{3} + 2M^3 - M^2) + 2dM(M-C) + dC$ floating point operations (flops) are required to obtain an orthonormal basis set spanning the range of S_W by using this approach. Here *l* represents the number of iterations required for convergence of the eigen-decomposition algorithm. However, the computations may become expensive and numerically unstable for large values of *M*. Since we do not need to find the eigenvalues (i.e., an explicit symmetric Schur decomposition) of S_W, the following algorithm can be used for finding the common vectors efficiently. It requires only $(2d(M-C)^2 + d(M-C))$ flops to find an orthonormal basis for the range of S_W and is based on the subspace methods and the Gram-Schmidt orthogonalization procedure.

Suppose that *d*>*M*-*C*. In this case, the subspace methods can be applied to obtain the common vectors x_{com}^i for each class *i*. To do this, we choose any one of the image vectors from the *i*-th class as the subtrahend vector and then obtain the difference vectors b_k^i of the so-called difference subspace of the *i*-th class [20]. Thus, assuming that the first sample of each class is taken as the subtrahend vector, we have $b_k^i = x_{k+1}^i - x_1^i$, $k = 1,..., N-1$.

The difference subspace B_i of the *i*-th class is defined as $B_i = span\{b_1^i, ..., b_{N-1}^i\}$. These subspaces can be summed up to form the complete difference subspace

$$B = B_1 + ... + B_C = span\{b_1^1, ..., b_{N-1}^1, b_1^2, ..., b_{N-1}^C\}. \quad (23)$$

The number of independent difference vectors b_k^i will be equal to the rank of S_W. For simplicity, suppose there are *M-C* independent difference vectors. Since by Theorem 3, *B* and the range space *V* of S_W, are the same spaces, the projection matrix onto *B* is the same as the matrix *P* (projection matrix onto the range space of S_W) defined previously in Section II-A. This matrix can be computed as

$$P = D(D^T D)^{-1} D^T, \tag{24}$$

where $D = [b_1^1 \quad .... \quad b_{N-1}^1 \quad b_1^2 \quad .... \quad b_{N-1}^C]$ is a *d* by *M-C* matrix [21]. This involves finding the inverse of an *M-C* by *M-C* nonsingular, positive definite symmetric matrix $D^T D$. A computationally efficient method of applying the projection uses an orthonormal basis for *B*. In particular, the difference vectors b_k^i can be orthonormalized by using the Gram-Schmidt orthogonalization procedure to obtain orthonormal basis vectors $\beta_1,....,\beta_{M-C}$. The complement of *B* is the indifference subspace $B^\perp$ such that

$$U = [\beta_1 \quad .... \quad \beta_{M-C}], \quad P = UU^T, \tag{25}$$

$$\overline{U} = [\beta_{M-C+1} \quad .... \quad \beta_d], \quad \overline{P} = \overline{U}\,\overline{U}^T, \tag{26}$$

where P and $\overline{P}$ are the orthogonal projection operators onto *B* and $B^\perp$, respectively. Thus matrices P and $\overline{P}$ are symmetric and idempotent, and satisfy $P + \overline{P} = I$. Any sample from each class can now be projected onto the indifference subspace $B^\perp$ to obtain the corresponding common vectors of the classes,

$$\begin{aligned} x_{com}^i &= \overline{P} x_m^i = x_m^i - P x_m^i \\ &= \overline{U}\,\overline{U}^T x_m^i = x_m^i - UU^T x_m^i, \qquad m = 1,...,N, \quad i = 1,...,C. \end{aligned} \tag{27}$$

The common vectors do not depend on the choice of the subtrahend vectors and they are identical to the common vectors obtained by using the null space of S_W. This follows from Theorem 3 below, which uses the results of Lemma 1 and Theorem 2.

Theorem 2: Let $V_i^{\perp}$ be the null space of the scatter matrix S_i, and $B_i^{\perp}$ be the orthogonal complement of the difference subspace B_i. Then $V_i^{\perp} = B_i^{\perp}$ and $V_i = B_i$.

Proof: See [20].

Lemma 1: Suppose that $S_1,....,S_C$ are positive semi-definite scatter matrices. Then

$$N(S_1 + ... + S_C) = \bigcap_{i=1}^{C} N(S_i), \tag{28}$$

where *N*( ) denotes the null space.

Proof: The null space on the left-hand side of the above identity contains elements α such that

$$(S_1 + ... + S_C)\alpha = 0 \tag{29}$$

or

$$\alpha^T (S_1 + ... + S_C)\alpha = \alpha^T S_1 \alpha + ... + \alpha^T S_C \alpha = 0, \tag{30}$$

by the positive semi-definiteness of $S_1 + ... + S_C$. Thus, again by the positive semi-definiteness, $\alpha \in N(S_1 + ... + S_C)$ if and only if

$$\alpha^T S_i \alpha = 0, \quad i=1,...,C, \tag{31}$$

or, equivalently, $\alpha \in \bigcap_{i=1}^{C} N(S_i)$.

Theorem 3: Let $S_1,....,S_C$ be positive semi-definite scatter matrices. Then

$$B = R(S_W) = R(S_1 + ... + S_C) = R(S_1) + ... + R(S_C) = B_1 + ... + B_C, \tag{32}$$

where R denotes the range.

Proof: Since it is well known that the null space and the range of a matrix are complementary spaces, using the previous Lemma 1, we have

$$\begin{aligned} R(S_1 + ... + S_C) &= (N(S_1 + ... + S_C))^{\perp} = (\bigcap_{i=1}^{C} N(S_i))^{\perp} = (N(S_1))^{\perp} + ... + (N(S_C))^{\perp} \\ &= R(S_1) + ... + R(S_C) = B_1 + ... + B_C, \end{aligned} \tag{33}$$

where the last equality is a consequence of Theorem 2.

After calculating the common vectors, the optimal projection vectors can be found by performing PCA as described previously in Section II-A. The eigenvectors corresponding to the nonzero eigenvalues of S_{com} will be the optimal projection vectors. However, optimal projection vectors can also be obtained more efficiently by computing the basis of the difference subspace B_{com} of the common vectors, since we are only interested in finding an orthonormal basis for the range of S_{com}.

The algorithm based on the Gram-Schmidt orthogonalization can be summarized as follows.

Step 1: Find the linearly independent vectors b_k^i that span the difference subspace *B* and set $B = span\{b_1^1, ..., b_{N-1}^1, b_1^2, ..., b_{N-1}^C\}$. There are totally *r* linearly independent vectors, where *r* is at most *M*-*C*.

Step 2: Apply the Gram-Schmidt orthogonalization procedure to obtain an orthonormal basis $\beta_1, ..., \beta_r$ for *B* and set $U = [\beta_1 \quad .... \quad \beta_r]$.

Step 3: Choose any sample from each class and project it onto *B* to obtain common vectors by using (27).

Step 4: Find the difference vectors that span B_{com} as

$$b_{com}^k = x_{com}^{k+1} - x_{com}^1, \quad k = 1, ..., C-1, \tag{34}$$

and apply the Gram-Schmidt orthogonalization to obtain an orthonormal basis $\tilde{w}_1,...,\tilde{w}_{C-1}$ for B_{com}. These vectors will be the optimal projection vectors to be used to form the projection matrix $\tilde{W} = [\tilde{w}_1 \quad .... \quad \tilde{w}_{C-1}]$, which will in turn be used to obtain feature vectors in (20) and (21). Note that columns of $\tilde{W}$ and columns of the projection matrix W (described in section II-A) span the same space and hence the matrices obey the equation $WW^T = \tilde{W}\tilde{W}^T$.

III. EXPERIMENTAL RESULTS

The Yale [7] and AR [22] face databases were used to test the proposed method.

A. Experiments with the Yale Face Database

The Yale face database consists of images from $C = 15$ different people, using 11 images from each person, for a total of 165 images. The images contain variations with the following facial expressions or configurations: center-light, with glasses, happy, left-light, without glasses, normal, right-light, sad, sleepy, surprised and wink. For subjects numbered 2, 3, 6, 7, 8, 9, 12 and 14, the normal facial expression and the without glasses (or with glasses if subject normally wears glasses) images were copies of each other. Thus, we removed the image without glasses (or with glasses if subject normally wears glasses) from every subject in order to make all classes have an equal number of samples and have all sample images distinct. Thus, we had 10 samples per subject yielding a face database size of 150. We preprocessed these images by aligning and scaling them so that the distances between the eyes were the same for all images, and also ensuring that the eyes occurred in the same coordinates of the image. The resulting image was then cropped. The final image size was 126x152. In addition to our proposed method, we also tested the Eigenface method, the Fisherface method, and the Direct-LDA method. We did not

test the PCA+Null Space method since it has the same recognition accuracy as our method. For the Eigenface method the images were normalized to have zero mean and unit variance, as this improved the performance of this method by reducing the within-class scatter. The recognition rates were computed by the "leave-one-out" strategy [8] since the training set size is relatively small. The nearest-neighbor algorithm was employed using Euclidean distance for classification. For the Eigenface method the most significant eigenvectors were chosen such that corresponding eigenvalues contain 95 % of the total energy [7]. For the Fisherface method, all images were first projected onto a (*M-C=134)*-dimensional space, where S_w was non-singular. The results for the Yale Database are given in Table I.

TABLE I
The Recognition Rates for the Yale Face Database

Methods	**Recognition Rate**
Eigenface	76%
Fisherface	96%
Direct-LDA	92%
Discriminative Common Vector	97.33%

B. Experiments with the AR-Face Database

The AR-face database includes 26 frontal images with different facial expressions, illumination conditions, and occlusions for 126 subjects. Images were recorded in two different sessions 14 days apart. Thirteen images were recorded under controlled circumstances in each session. The size of the images in the database is 768x576 pixels, and each pixel is represented by 24 bits of RGB color values.

We randomly selected $C = 50$ individuals (30 males and 20 females) for the experiment. Only nonoccluded images ((a)-(g) and (n)-(t) as in Fig. 1) were chosen for every subject. Thus, our face database size was 700 with 14 images per subject. Next, these images were converted to grayscale, aligned, scaled, localized and cropped using the same procedure described previously

for the Yale face database experiment. The final size of the images was 222x299. The training set consisted of $N = 7$ images randomly selected from each subject, and the rest of the images were used for the test set. Thus, a training set of $M = 350$ images and a test set of 350 images were created. A nearest-neighbor algorithm was employed using the Euclidean distance for classification. This process was repeated 4 times and the recognition rates were found by averaging the error rates of each run. The results are shown in Table II.

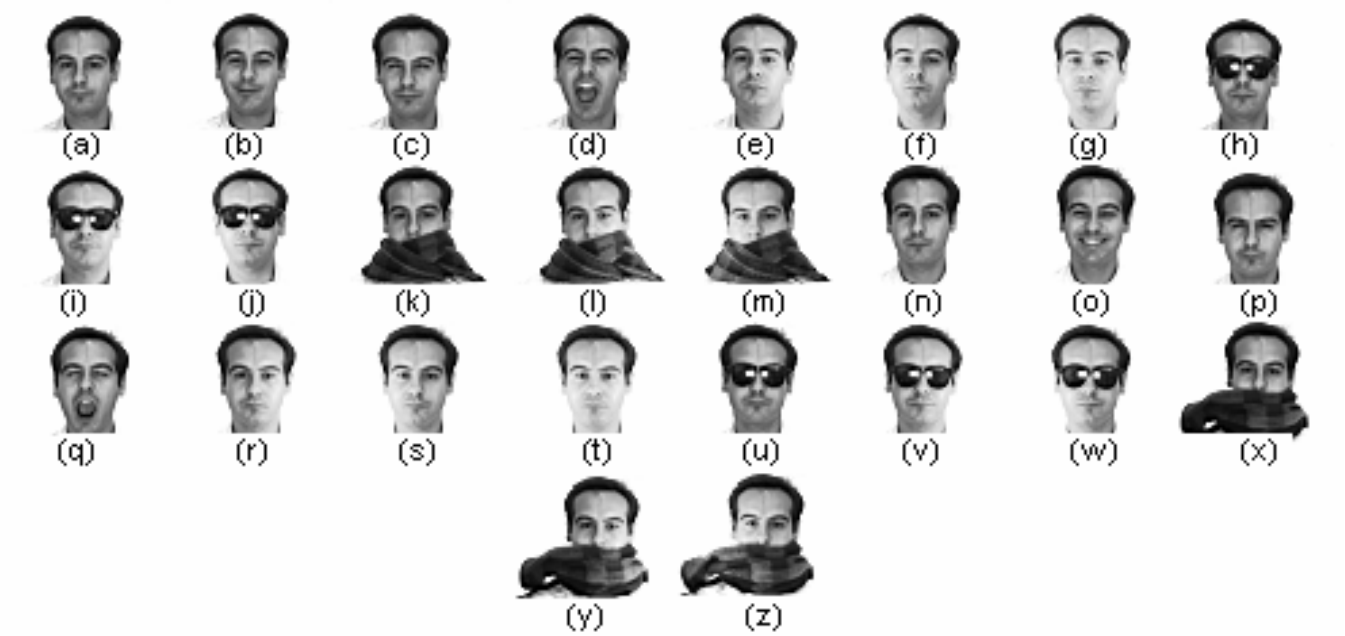

Fig. 1. Images of one subject in the AR-face database. First 13 images (a)-(m) were taken in one session and the others (n)-(z) in another session. Only nonoccluded images (a)-(g) and (n)-(t) were used in our experiments.

TABLE II
The Recognition Rates for the AR-Face Database

Methods	**Recognition Rate**
Eigenface	79.14%
Fisherface	98.85%
Direct-LDA	98.64%
Discriminative Common Vector	99.35%

The success of the proposed method depends on the size of the null space of the within-class scatter matrix, S_W. When the size of the null space is small, recognition rates are expected to be poor, since there will not be sufficient space for obtaining the optimal projection vectors. This is also mentioned in [17]. To verify this effect, we performed experiments using the pre-processed AR-face database images. We randomly selected 7 images from each class for training and used the rest for testing. Thus, a training set of 350 images and a test set of 350 images were created. To observe the decrease in performance due to a small null space, we would have to have a huge number of classes for a training set with sample space size 222x299. Unfortunately, we had a very limited number of classes in the training set. Thus, we had to take the approach of decreasing the dimensionality of the sample space by sub-sampling the images. Based on empirical observations, a new sample space size was chosen by down-sampling the images to 24x18. Then, we gradually decreased the number of classes from 50 down to 5. This procedure was repeated 8 times using randomly chosen subsets of the 50 classes, and recognition rates were found by averaging the rates of each run. The results are shown in Fig. 2. As can be seen, the performance decreases as the dimension of the null space decreases. This suggests that the initial sample space reduction step given in [17] is likely to reduce the achievable performance.

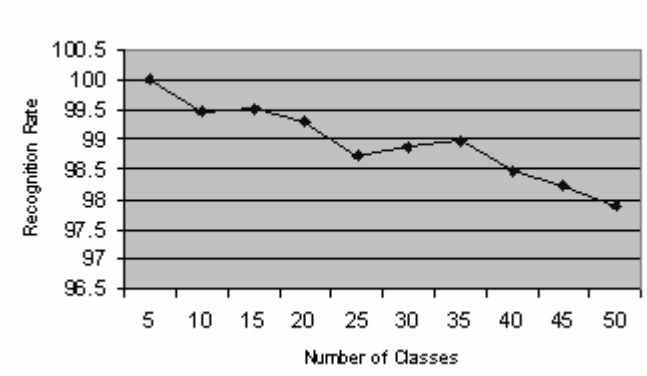

Fig. 2. The recognition rates as functions of the number of classes for subsampled images.

C. Discussion

Accuracy, training cost, execution speed, and storage requirements are some factors that may be used to judge a face recognition method. We discuss here the differences of these factors between the methods considered in this work. We also give a visual presentation of the eigenfaces and common vectors at the end of this discussion.

Experimental results show that the proposed method (as well as the PCA+Null Space method) yielded the highest performance in terms of accuracy. The Eigenface method yielded the lowest recognition rate. In particular, its recognition rate for the Yale face database was notably poor. The misclassified images for the Eigenface method were typically images that were not taken under the standard ambient light conditions used for most of the data (i.e., illumination was center-light, left-light, or right-light). Given that projection directions found by the Eigenface method are chosen for optimal reconstruction, this method is expected to work well when the testing samples of a subject are similar to the samples of the subject used for training. Since the leave-one-out method was used for testing and there was only one sample for these non-ambient light illumination conditions per class, these unusual illumination images behaved as data outliers (i.e., these images were far from the samples used for training). We would expect better results if there were more than one example with these illumination conditions. The other tested methods produced better results since projection directions minimizing the total within-class scatter were used. A significant part of the total within-class scatter was produced by the non-ambient lighting cases in all of the classes. This variation due to lighting conditions appears to produce similar deviations from class mean across all classes. Thus, we believe the resulting projection reduces variation due to lighting in all classes, even classes in which such variation did not appear in the training set.

The proposed method and the PCA+Null Space method require the same storage space, which is the smallest of all the methods studied. We need to store at most (C-1) d-dimensional projection vectors and C (C-1)-dimensional discriminative common vectors for comparison (In the PCA+Null Space method it is not necessary to save all the training sample feature vectors, only the smaller set of discriminative common vectors, although this has not been reported in the literature.) Secondly, the Direct-LDA and the Fisherface methods have the same storage requirements, which are higher than those of the proposed method and the PCA+Null Space method. For these methods we have to save at most (C-1) d-dimensional projection vectors and M (C-1)-dimensional sample feature vectors of the training set for comparison. Hence, the only difference among storage requirements of the four methods is the number of feature vectors saved for comparison (The difference is the need to store additional (M-C) (C-1)-dimensional vectors for the Direct-LDA and the Fisherface methods.) If M is small and d is large, this difference is negligible. However, if M is increased, this difference will also increase and become significant. Finally, for $n>C$-1 the Eigenface method has the largest storage space requirements. Here n is the number of the chosen significant eigenvectors and has been chosen such that the corresponding eigenvectors contain 95% of the total energy in our experiments. It was found to be a minimum of 65 for the Yale face database and 108 for the AR-face database.

Training cost is the amount of computations required to find the optimal projection vectors and the sample feature vectors of the training set for comparison. We compare the training cost of the methods based on their computational complexities (number of flops). The Direct-LDA method yields the highest efficiency in terms of computation complexity. The next efficient method is the proposed method, followed by the Eigenface method, the Fisherface method and the PCA+Null Space method. The computational comparison that is most interesting to us is

between the PCA+Null Space method and the proposed method, since these two methods yield the same accuracy and this accuracy is higher than the other methods. We estimated the computational complexities of these two algorithms and found PCA+Null Space to require approximately $(4dM^2 + 2l(\frac{4M^3}{3} + 2M^3 - M^2))$ flops and the proposed method required approximately $(2d(M-C)^2 + 4dMC)$ flops. Here l represents the number of iterations required for convergence of the eigen-decomposition algorithm. As d (the sample space size) and M (the number of training samples) get large, the proposed method requires less than half of the computations as the PCA+Null Space method.

Execution speed or testing time is the time that is required to classify a new test image. To do this, a test image must be projected onto the linear span of the projection vectors and compared to the sample feature vectors of the training set. Testing time determines the real-time efficiency of a method. We also compare testing times based on computational complexities here. Our proposed method and the PCA+Null Space method yield the highest efficiency in terms of computation. In these methods a test image is projected onto (C-1) d-dimensional vectors and compared to the C (C-1)-dimensional vector set. The Direct-LDA and Fisherface methods follow them in cost. In these methods, a test image is projected onto (C-1) d-dimensional vectors and compared to M (C-1)-dimensional vectors. As a result, the only difference between the testing times of these four methods is the time that is spent on comparison. In the Direct-LDA and the Fisherface methods, a projected test image must be compared to all sample feature vectors of the training set instead of being compared to only one representative for each class. Thus, as with the storage requirements, when the number of samples M is increased, the difference between testing times of these methods will also increase and become significant. Finally, the Eigenface method yields the maximum test time in the case n>C-1.

In summary, the proposed method becomes progressively more efficient, compared to the other methods, as the size of the sample space *M* is increased. In Table III we present the overall results of our comparisons. The top row of the table lists the four criteria on which the methods were compared. The left column of the table is a qualitative ranking of how each method performed, and the cells in the table contain methods with comparable performance.

TABLE III
Comparisons of Performance Across Methods for $n>C$-1

Performance Rank	**Accuracy**	**Training Cost**	**Testing Cost**	**Storage Requirements**
1	Discriminative Common Vector, PCA+Null Space	Direct-LDA	Discriminative Common Vector, PCA+Null Space	Discriminative Common Vector, PCA+Null Space
2	Fisherface	Discriminative Common Vector	Fisherface, Direct-LDA	Fisherface, Direct-LDA
3	Direct-LDA	Eigenface	Eigenface	Eigenface
4	Eigenface	Fisher		
5		PCA+Null Space		

The eigenfaces and common vectors obtained from the Yale and AR face databases are shown in Fig. 3 and Fig. 4 respectively. Fig. 3 displays the absolute values of the elements of the eigenfaces in an image form. If the common vectors are displayed in the same manner, the resulting image is mostly very dark and obscures the interesting details in the darker areas. Thus, the common vectors in Fig. 4 were displayed after taking the absolute value followed by the logarithm. Eigenfaces characterize the variations resulting from differences in lighting conditions, facial expression, and so on between face images. Thus, using the most significant eigenfaces (i.e., the ones corresponding to the largest eigenvalues) may not be the best choice from a discrimination point of view. In contrast common vectors represent the invariant regions of faces. Thus, the eyes, nose, part of the forehead above the eye brows, and cheeks are dominant in common vectors.

Fig. 3. Most 10 significant eigenfaces obtained from the Yale and AR face databases. The first row shows 10 significant eigenfaces obtained from one of the training set of the AR-face database and the second row shows 10 significant eigenfaces obtained from one of the training set of the Yale face database.

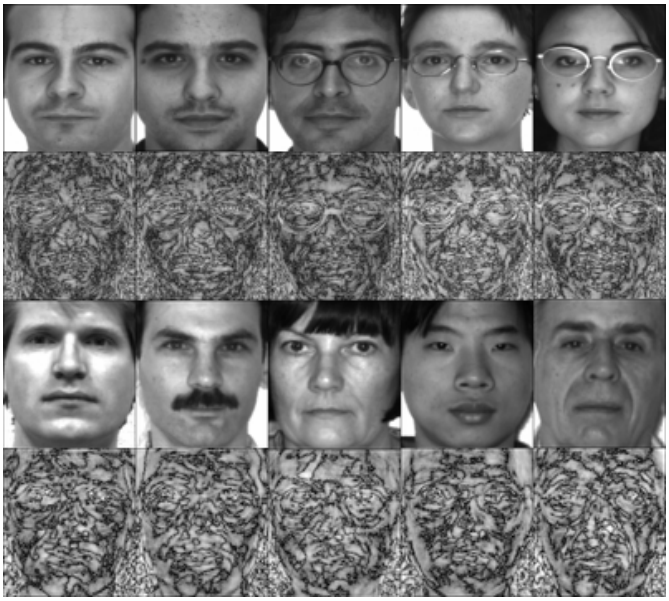

Fig. 4. Some of the common vectors obtained from the Yale and AR face databases. The first and second rows shows some individuals and corresponding common vectors from the AR-face database and the third and fourth rows show some individuals and corresponding common vectors for the Yale face database.

IV. CONCLUSIONS

In this paper we proposed a new method for addressing computational difficulties encountered in obtaining the optimal projection vectors in the null space of the within-class scatter. We showed that every sample in a given class produces the same unique common vector when they

are projected onto the null space of S_W . We also proposed an alternative algorithm for obtaining common vectors based on the subspace methods and the Gram-Schmidt orthogonalization procedure, which avoids handling large matrices and improves the stability of the computation. Using common vectors also leads to an increased computational efficiency in face recognition tasks. The Optimal projection vectors are found by using the common vectors, and the discriminative common vectors are determined by projecting any sample from each class onto the span of optimal projection vectors. There is no loss of information content in our method, in the sense that the method has 100% recognition rate for the training set data. Experimental results show that our method is superior to other methods in terms of accuracy, real-time performance, storage requirements, and numerical stability.

REFERENCES

[1] R. Chellappa, C.L. Wilson, and S. Sirohey, "Human and machine recognition of faces: a survey," *Proceedings of the IEEE*, vol. 83, pp. 705-740, May 1995.

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

[3] W. Zhao, R. Chellappa, A. Rosenfeld, and P.J. Phillips, "Face recognition: a literature survey," Technical Report CAR-TR-948, University of Maryland, College Park, 2000.

[4] M. Turk, "A random walk through eigenspace," *IEICE Trans. Inf. & Syst.*, vol. E84-D, no. 12, pp. 1586-1695, December 2001.

[5] M. Turk and A. P. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71-86, 1991.

[6] 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.

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

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

[9] Q. Tian, M. Barbero, Z. H. Gu, and S. H. Lee, "Image classification by the Foley-Sammon transform," *Opt. Eng.*, vol. 25, no. 7 , pp. 834-840, 1986.

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

[11] Y. Q. Cheng, Y. M. Zhuang, and J. Y. Yang, "Optimal fisher discriminate analysis using the rank decomposition," *Pattern Recognition*, vol. 25, pp. 101-111, 1992.

[12] 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.

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

[14] D-Q Dai and P. C. Yuen, "Regularized discriminant analysis and its application to face recognition," *Pattern Recognition*, vol. 36, pp. 845-847, 2003.

[15] 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.

[16] 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.

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

[18] K. Liu, Y-Q Cheng, and J-Y Yang, "A generalized optimal set of discriminant vectors," *Pattern Recognition*, vol. 25, no. 7, pp. 731-739, 1992.

[19] M. B. Gulmezoglu, V. Dzhafarov, M. Keskin, and A. Barkana, "A novel approach to isolated word recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 6, November 1999.

[20] M. B. Gulmezoglu, V. Dzhafarov, and A. Barkana, "The common vector approach and its relation to principal component analysis," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 6, September 2001.

[21] E. Oja, *Subspace Methods of Pattern Recognition.* Letchworth, UK: Research Studies Press, 1983, pp. 13-14.

[22] A.M. Martinez and R. Benavente, "The AR face database," CVC Tech. Report #24, 1998.

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

[24] A. Webb, *Statistical Pattern Recognition.* New York: Oxford University Press, 1999.

[25] Y. Adini, Y. Moses, and S. Ullman, "Face recognition: the problem of compensating for changes in illumination direction," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 721-732, July 1997.

[26] A. M. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228-233, February 2001.