A Comparison of the Common Vector and the Discriminative Common Vector Methods for Face Recognition

Hakan CEVIKALP[1], Bilal BARKANA[1], and Atalay BARKANA[2]
[1] Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, Tennessee, USA
[2] Department of Electrical and Electronics Engineering, Anadolu University, Eskisehir, Turkey
hakan.cevikalp@vanderbilt.edu, bilal.barkana@vanderbilt.edu, abarkana@ogu.edu.tr

Abstract

The Common Vector (CV) method is a successful method which has been originally proposed for isolated word recognition problems in the case where the number of samples for each class is less than or equal to the dimensionality of the sample space. This method suggests elimination of all the features that are in the direction of the eigenvectors corresponding to the nonzero eigenvalues of the covariance matrix for each class. The feature vectors obtained after this operation are unique for each class and called common vectors. Recently, a similar method called the Discriminative Common Vector (DCV) method has been proposed for face recognition problems. Instead of using a given class' own covariance matrix, this method uses the within-class scatter matrix of all classes to obtain the common vectors. Then, PCA is applied to the common vectors to obtain the final projection vectors. In this paper we apply the CV method to the face recognition problem and compare the CV and the DCV methods in terms of recognition accuracy, training time efficiency, storage requirements, and real-time performance.

Keywords: common vectors, discriminative common vectors, face recognition, feature extraction, subspace methods.

1. Introduction

The face is an important source of information for the identification of people in communities. Face recognition is usually defined as the identification of individuals from images of their faces by using a stored database of faces labeled with people's identities. However, there are many factors that degrade the performance of face recognition techniques, such as bad illumination conditions, different facial expressions, hair styles, make up, and so on. In addition to these factors, lighting, background, and scale changes make the face recognition task even more challenging. Additional problematic conditions include noise, occlusion, and many other possible factors [1].

Identification of individuals from their faces is gradually becoming essential for today's world's demands. It has applications in areas related to the public and private security issues, electronic identity controls, law enforcement applications, access control, and so on. Numerous methods have been developed for face recognition. Among these methods appearance-based approaches directly operate on two-dimensional face images of size *w* by *h*, which are represented by vectors in a *w*x*h* dimensional space. The dimensionality of this space is typically large compared to the number of samples in the training sets. This causes situations known as the small sample size problem [2].

The DCV method is a successful linear method that has been proposed for the face recognition tasks with the small sample size problems [3]. It employs the projection vectors that come from the null space of the within-class scatter matrix of the training set samples for feature extraction. This method tries to maximize a novel criterion which is given in the next section.

The CV method has been originally proposed for isolated word recognition problems [4]. It tries to extract the features that are common for the samples of a same class. It employs the covariance matrices of classes to accomplish its goal. Although this method is applicable to the face recognition tasks with the small sample size problem, no study has been reported in literature.

In this paper we apply the CV method to face recognition and compare it to the DCV method. The remainder of the paper is organized as follows. In Section 2, the CV and the DCV methods are reviewed. In Section 3, we describe the data sets and experimental results. In section 4, two methods are compared. Finally, our conclusions are formulated in Section 5.

2. Methods

In this section we will review the CV and the DCV methods.

2.1 Common Vector Method

The CV method has been originally proposed for isolated word recognition problems, where the number of samples in each class is smaller than or equal to the dimensionality of the sample space (i.e., the covariance matrix of each class is singular). This method extracts the features that are common for all samples in each class. In order to accomplish its goal, the method eliminates all features that are in the direction of the eigenvectors corresponding to the nonzero eigenvalues of the covariance matrices of classes. Thus, the method suggests doing the opposite of the well-known so-called subspace methods [5].

Two algorithms have been proposed for this method. One algorithm uses the covariance matrices of the classes, whereas the other employs the subspace methods and the Gram-Schmidt orthogonalization procedure. In this paper we review the first algorithm from [4].

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 $d > N_i$ for $i = 1,...,C$. In this case, the covariance matrices of classes are defined as

$$\begin{aligned}\Sigma_i &= \frac{1}{N_i}\sum_{m=1}^{N_i}(x_m^i - \mu_i)(x_m^i - \mu_i)^T \\ &= A_i A_i^T \quad i = 1,...,C, m = 1,...,N_i,\end{aligned} \tag{1}$$

where μ_i is the mean of the samples in the *i*-th class and $A_i \in R^{dxN_i}$ is given by

$$A_i = [(x_1^i - \mu_i)/\sqrt{N_i} \;...\; (x_{N_1}^i \text{ - } \mu_{\mathrm{i}})/\sqrt{N_i}\;]. \tag{2}$$

Each sample in the training set is represented as,

$$x_m^i = x_{m,dif}^i + x_{com}^i + \varepsilon_m^i,\ i = 1,...,C, m = 1,...,N_i, \tag{3}$$

where x_{com}^i is a unique vector representing the *i*-th class, and ε_m^i is the error vector term. The CV method tries to minimize the criterion given below for each class,

$$F_i = \sum_{m=1}^{N_i} \| \varepsilon_m^i \|^2 = \sum_{m=1}^{N_i} \| x_m^i - x_{m,dif}^i - x_{com}^i \|^2,\ i = 1,...,C, \tag{4}$$

where $\|\,.\,\|$ denotes the Euclidean norm. It was shown that if the common vector x_{com}^i is chosen as

$$x_{com}^i = x_m^i - x_{m,dif}^i,\ \ i = 1,...,C, m = 1,...,N_i, \tag{5}$$

then F is minimized such that $F_{\min} = 0$, where $x_{m,dif}^i$ represents the projection of x_m^i onto the range space of the covariance matrix of the *i*-th class [4]. The projection can be computed by

$$x_{m,dif}^i = P_i x_m^i,\ i = 1,...,C, m = 1,...,N_i, \tag{6}$$

where P_i is the orthogonal projection operator of the range space of Σ_i. As can be seen in equations (5) and (6), x_{com}^i is unique for each class and it does not depend on the choice of the sample (i.e., x_{com}^i is independent of the sample index *m*) [4].

To recognize a test sample x_{test}, the test sample is projected onto the null space of the covariance matrix of each class separately and the projected vectors are compared to the common vector of each class by using the Euclidean distance. Then, the test sample is assigned to the class which gives the minimum distance.

The method described above can be summarized as follows:

Step 1: Compute the nonzero eigenvalues and the corresponding eigenvectors of the covariance matrix Σ_i of each class by using the matrix $A_i^T A_i \in R^{N_i x N_i}$, where $\Sigma_i = A_i A_i^T \in R^{dxd}$ and A_i is given by (2). Let λ_k^i and v_k^i be the *k*-th nonzero eigenvalue and the corresponding eigenvector of $A_i^T A_i$, where $k \le N_i - 1$. Then $u_k^i = A_i v_k^i$ will be the eigenvector that corresponds to the *k*-th nonzero eigenvalue of Σ_i. Normalize the computed eigenvectors and set $U_i = [u_1^i \quad ... \quad u_{r_i}^i]$, where r_i is the rank of Σ_i.

Step 2: Project any sample from each class onto the null space of Σ_i and compute the common vector of each class by,

$$x_{com}^i = x_m^i - P_i x_m^i = x_m^i - U_i U_i^T x_m^i,\ i = 1,...,C, m = 1,...,N_i. \tag{7}$$

Note that, common vectors x_{com}^i are unique for each class and independent of the sample index *m*.

Step 3: Project a test sample onto the null spaces of Σ_i to obtain the feature vectors by

$$\Omega_{test}^i = x_{test} - P_i x_{test},\ i = 1,...,C\ . \tag{8}$$

Compute the Euclidean distances between the test sample feature vector and the common vectors of each class by,

$$\kappa_i = \| \Omega_{test}^i - x_{com}^i \|,\ i = 1,...,C. \tag{9}$$

Assign the test sample to the class which produces the minimum distance.

There might be situations where the dimensionality of the null space of Σ_i is smaller than the dimensionality of the range space of Σ_i. In these situations common vectors can be computed directly by the following formula

$$x_{com}^i = P_i^{\perp} x_m^i,\ i = 1,...,C, m = 1,...,N_i, \tag{10}$$

where $P_i^{\perp}$ is the orthogonal projection operator onto the null space of Σ_i and $P_i + P_i^{\perp} = I$.

2.2 Discriminative Common Vector Method

The DCV method is a supervised method that has been recently proposed for face recognition problems with the small sample size problem. It tries to find the projection directions that, on one hand maximize the Euclidean distances between the samples of different classes and, on the other, minimize the distance between the samples of the same class. In order to do it, this method employs the within-class scatter matrix of the samples to obtain the feature vectors. The within-class scatter, the between-class scatter, and the total scatter matrices of the training samples are defined as

$$S_W = \sum_{i=1}^{C}\sum_{m=1}^{N_i}(x_m^i - \mu_i)(x_m^i - \mu_i)^T = A_W A_W^{\ T}, \qquad (11)$$

$$S_B = \sum_{i=1}^{C} N_i(\mu_i - \mu)(\mu_i - \mu)^T = A_B A_B^{\ T}, \qquad (12)$$

and

$$S_T = \sum_{i=1}^{C}\sum_{m=1}^{N_i}(x_m^i - \mu)(x_m^i - \mu)^T = A_T A_T^{\ T} = S_W + S_B, \qquad (13)$$

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

$$A_W = [x_1^1 - \mu_1 \ldots\ x_{N_1}^1 - \mu_1\ \ x_1^2 - \mu_2 \ldots x_{NC}^C \text{ - } \mu_C\ ], \qquad (14)$$

$$A_B = [\sqrt{N_1}(\mu_1 - \mu) \quad \ldots \quad \sqrt{N_C}(\mu_C - \mu)], \qquad (15)$$

and

$$A_T = [x_1^1 - \mu \ldots\ x_{N_1}^1 - \mu\ \ x_1^2 - \mu_2 \ldots x_{NC}^C \text{ - } \mu\ ]. \qquad (16)$$

The modified FLDA criterion, $J_{MFLDA}(W_{opt}) = \arg\max_{W} \frac{|W^T S_B W|}{|W^T S_T W|}$ [6], attains its maximum, 1, in the special case where $w^T S_W w = 0$ and $w^T S_B w \neq 0$, for all $w \in R^d \setminus \{0\}$. However, a projection vector *w*, satisfying the above conditions, does not necessarily maximize the between-class scatter. In this case, a better criterion will be

$$J(W_{opt}) = \arg\max_{|W^T S_W W|=0} |W^T S_B W| = \arg\max_{|W^T S_W W|=0} |W^T S_T W|. \qquad (17)$$

Therefore, 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 $N(S_W)$ of S_W and then apply PCA to the projected samples.

The DCV method 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 (14). Set $Q = [\alpha_1 \quad \ldots \quad \alpha_r]$, where *r* is the rank of S_W.

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

$$x_{com}^i = x_m^i - QQ^T x_m^i, \quad i = 1,\ldots,C, \quad m = 1,\ldots,N_i. \qquad (18)$$

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 (18) is independent of the sample index *m* [3]. These vectors are also 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

$$S_{com} = \sum_{i=1}^{C}(x_{com}^i - \mu_{com})(x_{com}^i - \mu_{com})^T = A_{com} A_{com}^T, \qquad (19)$$

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

$$A_{com} = [x_{com}^1 - \mu_{com} \quad \ldots \quad x_{com}^C - \mu_{com}]. \qquad (20)$$

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 $W = [w_1 \quad \ldots \quad w_{rC}]$, which will be used to obtain the feature vectors of the samples. Here, $r_C \leq C - 1$ 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 $\Omega_i = [< x_m^i, w_1 > \quad \ldots \quad < x_m^i, w_{rC} >]^T$ 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 ($i = 1,\ldots,C$) 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*. Note that after the projection, the distances between the training set samples of the same classes decreased to zero, which is the minimum distance that can be achieved.

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

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

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

3. Experimental Results

In our experiments we used the ORL (Olivetti-Oracle Research Lab) face database [7]. 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 upright, 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.

We first randomized the samples in the database and then selected $k = 3,5,7,9$ from each class for training and the rest $(10-k)$ samples of each class were used for testing. We have not applied any preprocessing to the images. Then recognition rates were computed. Euclidean distance is used to compute the distances between the sample feature vectors of test set and the common vectors for the CV method and similarly the same metric is used to compute the distances between the feature vectors of test samples and the discriminative common vectors. This process was repeated seven times and the recognition rates were found by averaging the recognition rates in each run. The results for the test sets are given in Table 1. We did not give the training set results since it is 100% for both methods.

Some of the common vectors computed for the CV and the DCV methods are plotted in Fig. 1. Fig. 1 displays the absolute values of the common vectors obtained by the CV method in image form. On the other hand, to display the common vectors obtained by the DCV method we took the logarithm of the values after taking the absolute values since common vectors displayed by only taking the absolute values were mostly dark. In Fig. 2, we plotted some of the projection vectors w_k obtained from the common vectors in the DCV method.

4. Discussion

Recognition accuracy, training cost, storage requirements, and the real-time performance are some factors that may be used to evaluate a method. We discuss here the differences of these factors between the CV and the DCV methods.

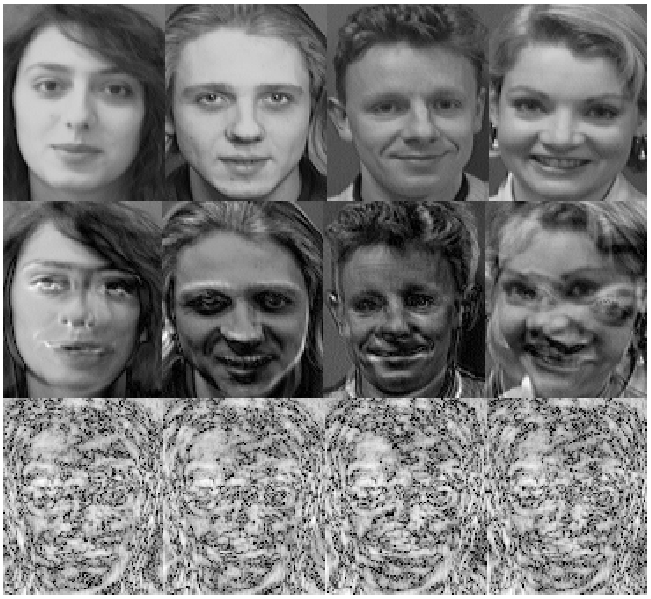

Fig.1. Common vectors obtained by the CV and the DCV methods. The first row shows some individuals from the ORL face database and the second and the third rows show the corresponding common vectors obtained by the CV and the DCV methods, respectively.

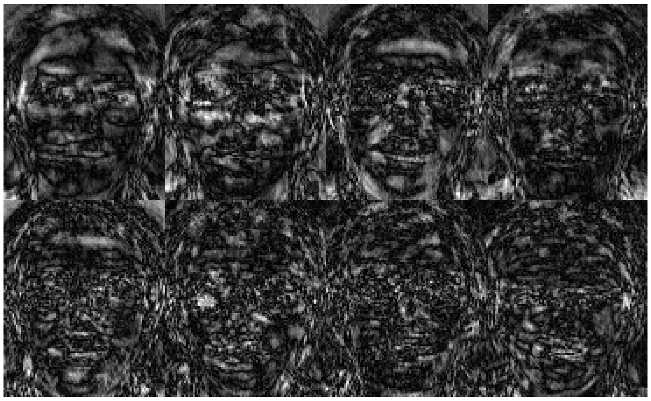

Fig. 2. Some of the projection vectors obtained from the common vectors for the DCV method.

TABLE I.
Recognition Rates (%) of the ORL face database

Number of training samples in each class	Methods	
	CV	DCV
$k = 3$	88.82 $\sigma = 3.73$	91.02 $\sigma = 1.89$
$k = 5$	95.78 $\sigma = 1.41$	96.92 $\sigma = 1.30$
$k = 7$	97.97 $\sigma = 1.16$	98.21 $\sigma = 1.39$
$k = 9$	99.28 $\sigma = 1.21$	99.28 $\sigma = 1.21$

As can be seen in Table I, the DCV method tends to yield better results compared to the CV method. The results reveal the important fact that there is a relationship between the number of training samples k in each class and the difference between the recognition rates of the CV and the DCV methods. As the number of training set samples is increased, the difference between the recognition rates decreases and finally becomes zero in this example. These observations somewhat support the hypothesis that the variations among the face samples of each class are similar. Therefore, we can assume that the scatter matrices of each face class are identical and we can replace it with the within-class scatter matrix. A similar assumption is made in the Fisher's Linear Discriminant Analysis approach. That is why we obtained better results for the DCV method in the case of having only a few training vectors in each class. As explained before, the CV method first models the variations in each class and removes them from the samples in order to obtain the common vectors. If this variation is modeled correctly, we will classify all the samples correctly. The low recognition rates of the CV method for small numbers of training set samples show that the number of training samples in each class is not enough to obtain a good model of the variations. On the other hand, the DCV method does a better job with the small number of training set samples_since it makes use of all of the vectors from all of the classes and does not perform a separate analysis on each class by itself. Some of the variations that come from the test samples of one class may be captured by the variations between the training set samples of one or more other classes.

Training cost is the amount of computations required to find the projection vectors and the sample feature vectors of the training set samples. We compare the training cost of the methods based on their computational complexities (number of flops). The CV method yields higher efficiency in terms of computation complexity since the DCV method includes an additional step of applying PCA to the common vectors.

The DCV method requires less storage space than the CV method. If we assume that all the training set sample vectors are linearly independent then the CV method requires us to store (M-C) d-dimensional projection vectors and C d-dimensional common vectors. However, we need to store only (C-1) d-dimensional projection vectors and C (C-1)-dimensional discriminative common vectors for the DCV method. Therefore if we assume that each class has N samples, the storage space of the CV method is approximately N times of the storage space of the DCV method.

The real-time performance of a method is determined by the time that is required to classify a new test image. To do this, we need to compute the feature vector of the test sample and compare it to the feature vectors of training set. We compare testing times based on computational complexities here. The DCV method is more efficient than the CV method in terms of testing time. For the CV method, we need to project our test sample onto (M-C) d-dimensional vectors to obtain feature vectors and compute the distances between the d-dimensional common vector and the feature vectors. On the other hand we need to project our test sample onto only (C-1) d-dimensional vectors to obtain the feature vector of the test sample and compare it to the C (C-1)-dimensional vectors. Assuming d>>(C-1), the difference between the testing times of the methods is determined by the number of computations that is required to project a test sample onto (M-2C+1) d-dimensional vectors.

5. Conclusion and Future Work

After comparing the CV and the DCV methods, we arrive at the following conclusions:

i) The DCV method is more efficient than the CV method in terms of recognition accuracy, storage requirements, and real-time performance for face recognition tasks. However, the training cost of the CV method is less than the DCV method.

ii) The CV method is expected to perform well if the variations among the test samples of a class are similar to the variations among the training samples of that class.

iii) The DCV method performs well if the variations among the samples of classes are similar. This enables us to classify the test samples more accurately even if they are not similar to the ones used for training.

In the near future, we plan to apply the CV and the DCV methods to recognition tasks, other than face recognition with the small sample size problem.

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

[3] 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.

[4] 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.

[5] E. Oja, *Subspace Methods of Pattern Recognition.* Letchworth, UK: Research Studies Press, 1983.

[6] 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.

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