View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by CiteSeerX

Face Recognition by Using Discriminative Common Vectors

Hakan Cevikalp and Mitch Wilkes
Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, Tennessee, USA
hakan.cevikalp@vanderbilt.edu, 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 also known as the "small sample size" problem. In this paper, we propose a new face recognition method based on the discriminative common vectors for the small sample size case. The discriminative common vectors representing the people in the face database were found by using the null space of the within-class scatter matrix. Then, these vectors were used for classification of new faces. Test results show that the proposed method is superior to other methods in terms of accuracy, efficiency, and numerical stability.

1. Introduction

Recently, due to military, commercial, and law enforcement applications, face recognition has received significant attention in several disciplines such as image processing, pattern recognition, computer vision and neural networks. Its applications include static matching of controlled format images such as passports, credit cards, photo ID's, driver's licenses, and mug shots. A more challenging application includes real-time detection and recognition of faces in surveillance video images [1].

Face recognition can be defined as the identification of individuals using a stored database of faces labeled with people's identities. It requires detection of faces, localization of them 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, lighting, noise, occlusion, scale changes and so on which affect the appearance of an individual's facial features.

Many methods have been proposed for face recognition within the last decades [1]. These methods can be divided into three categories: 3-D model-based, feature based and appearance-based methods. 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 [3]. When using appearance-based approaches, a two-dimensional image of size *w* by *h* pixels is represented by a vector in a *wh*-dimensional space. This space is called the sample space or the image space, and its dimension is typically very high. However, since the image vectors are correlated, any image in the sample space can be represented in a lower-dimensional subspace without losing a significant amount of information. The Eigenface method [4], the Fisherface method [5], the Direct-LDA method [6], and the Null Space method [7] have been proposed for finding such a lower-dimensional subspace. The method we proposed in this paper is based on the Null Space method.

2. The Null Space Method

This method tries to maximize the modified Fisher's Linear Discriminant (FLD) criterion, $J(\mathbf{W}_{opt}) = \arg\max \frac{|\mathbf{W}^T\mathbf{S}_B\mathbf{W}|}{|\mathbf{W}^T\mathbf{S}_T\mathbf{W}|}$, where $\mathbf{S}_B$ is the between-class scatter matrix, $\mathbf{S}_T$ is the total scatter matrix, and $\mathbf{W}$ is the matrix whose columns are the optimal projection vectors. The only difference between the FLD and the modified FLD criteria is that the latter uses $\mathbf{S}_T$ instead of the within-class scatter matrix $\mathbf{S}_W$ in denominator. In [8], it has been shown that the FLD criterion can be replaced by the modified FLD criterion in the course of solving the optimal projection vectors. It is not hard to see that when the projection directions satisfy the condition of $\mathbf{w}_k^T\mathbf{S}_W\mathbf{w}_k = 0$ and $\mathbf{w}_k^T\mathbf{S}_B\mathbf{w}_k \neq 0$, the modified FLD criterion attains its maximum, 1. However, a projection vector that satisfies the above condition does not necessarily maximize the between-class scatter. In this case, a better criterion will be,

$$J(\mathbf{W}_{\mathbf{opt}}) = \arg\max_{|\mathbf{W}^T\mathbf{S}_W\mathbf{W}|=0} |\mathbf{W}^T\mathbf{S}_B\mathbf{W}| = \arg\max_{|\mathbf{W}^T\mathbf{S}_W\mathbf{W}|=0} |\mathbf{W}^T\mathbf{S}_T\mathbf{W}| . \quad (1)$$

All image samples in the training set are first projected onto the null space of $\mathbf{S}_W$ to find the optimal projection vectors. As a result, the new within-class scatter matrix of the projected samples will be a zero matrix. Then, PCA is applied to the projected samples to obtain the optimal projection vectors. To project samples onto the null space of $\mathbf{S}_W$, we have to find the orthonormal vector set that

0-7695-2128-2/04 $20.00 (C) 2004 IEEE

spans the null space of $\mathbf{S_W}$. But this is almost impossible for the large values of sample space (e.g. an image of size 256x256 creates a 65536 dimensional sample space.) To overcome this problem, authors used a pixel grouping method to extract geometric features and reduce the dimension of the sample space in [7]. Then they applied the Null Space method in this new reduced space. However, we showed that the performance of the Null Space method depends on the dimension of the null space of $\mathbf{S_W}$ in the sense that larger dimension provides better performance. Thus, any kind of pre-processing that reduces the original sample space such as, a pixel grouping method, is likely to reduce achievable performance and therefore should be avoided [9]. The method proposed below solves this problem and allows us to work in the original sample space.

3. The Discriminative Common Vector Method

Let the training set be composed of *C* classes, where each class contains *N* samples, and let $\mathbf{x_m^i}$ 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. Let us assume *d>M-C*. In this case, $\mathbf{S_W}$, $\mathbf{S_B}$, and $\mathbf{S_T}$ can be found by the following equations:

$$\mathbf{S_W} = \sum_{i=1}^{C}\sum_{m=1}^{N}(\mathbf{x_m^i} - \boldsymbol{\mu}_\mathbf{i})(\mathbf{x_m^i} - \boldsymbol{\mu}_\mathbf{i})^T = \mathbf{AA^T}, \quad (2)$$

$$\mathbf{S_B} = \sum_{i=1}^{C} N(\boldsymbol{\mu}_\mathbf{i} - \boldsymbol{\mu})(\boldsymbol{\mu}_\mathbf{i} - \boldsymbol{\mu})^T, \quad (3)$$

and

$$\mathbf{S_T} = \sum_{i=1}^{C}\sum_{m=1}^{N}(\mathbf{x_m^i} - \boldsymbol{\mu})(\mathbf{x_m^i} - \boldsymbol{\mu})^T, \quad (4)$$

where $\boldsymbol{\mu}$ is the mean of all samples, and $\boldsymbol{\mu}_\mathbf{i}$ is the mean of the samples in the *i*-th class. $\mathbf{A}$ is a *d*x*M* matrix as given below:

$$\mathbf{A} = [\mathbf{x_1^1} - \boldsymbol{\mu}_\mathbf{1} \quad \ldots \quad \mathbf{x_N^1} - \boldsymbol{\mu}_\mathbf{1} \quad \mathbf{x_1^2} - \boldsymbol{\mu}_\mathbf{2} \quad \ldots \quad \mathbf{x_N^C} - \boldsymbol{\mu}_\mathbf{C}] \quad (5)$$

To find the optimal projection vectors in the null space of $\mathbf{S_W}$, we need to find the orthonormal vector set that spans the null space of $\mathbf{S_W}$. However, this task is computationally intractable because the dimension of the null space of $\mathbf{S_W}$ can be very large. Since we can easily find the orthonormal vector set that spans the complement of the null space (i.e. the range space) of $\mathbf{S_W}$ by using the smaller matrix (*M*x*M)* $\mathbf{A^TA}$, we can use this vector set to find the projections of the samples in the training set onto the null space of $\mathbf{S_W}$ [9].

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

$$V = span\{\boldsymbol{\alpha}_\mathbf{k} \mid \mathbf{S_W}\boldsymbol{\alpha}_\mathbf{k} \neq \mathbf{0}, \quad k = 1,\ldots,r, \quad \boldsymbol{\alpha}_\mathbf{k} \in R^d\}, \quad (6)$$

and

$$V^\perp = span\{\boldsymbol{\alpha}_\mathbf{k} \mid \mathbf{S_W}\boldsymbol{\alpha}_\mathbf{k} = \mathbf{0}, \quad k = r+1,\ldots,d, \quad \boldsymbol{\alpha}_\mathbf{k} \in R^d\} \quad (7)$$

where $r < d$ is the rank of $\mathbf{S_W}$, $\{\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_\mathbf{d}\}$ is an orthonormal set, and $\{\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_\mathbf{r}\}$ is the set of orthonormal vectors that span the range of $\mathbf{S_W}$.

Consider the matrices $\mathbf{Q} = [\boldsymbol{\alpha}_\mathbf{1} \quad \ldots \quad \boldsymbol{\alpha}_\mathbf{r}]$ and $\overline{\mathbf{Q}} = [\boldsymbol{\alpha}_\mathbf{r+1} \quad \ldots \quad \boldsymbol{\alpha}_\mathbf{d}]$. Let $\mathbf{P} = \mathbf{QQ^T}$ and $\overline{\mathbf{P}} = \overline{\mathbf{Q}}\overline{\mathbf{Q}}^\mathbf{T}$. Since $R^d = V \oplus V^\perp$,

$$\mathbf{x_{com}^i} = \mathbf{x_m^i} - \mathbf{Px_m^i} = \overline{\mathbf{P}}\mathbf{x_m^i}, \quad m = 1,\ldots,N, \quad i = 1,\ldots,C. \quad (8)$$

In this way, it turns out we obtain the same unique vector for all the samples of the same class, i.e., the vector on the right-hand side of (8) is independent of the sample index *m* [9]. We called these vectors as the common vectors.

The optimal projection vectors are the vectors that maximize the total scatter of the common vectors. In other words,

$$J(\mathbf{W_{opt}}) = \underset{|\mathbf{W^TS_WW}|=\mathbf{0}}{\arg\max} |\mathbf{W^TS_BW}| = \arg\max |\mathbf{W^TS_{com}W}|, \quad (9)$$

where $\mathbf{W}$ is a matrix whose columns are the orthonormal optimal projection vectors $\mathbf{w_k}$, and $\mathbf{S_{com}}$ is the scatter matrix of the common vectors. $\mathbf{S_{com}}$ can be found by the following equation,

$$\mathbf{S_{com}} = \sum_{i=1}^{C}(\mathbf{x_{com}^i} - \boldsymbol{\mu}_\mathbf{com})(\mathbf{x_{com}^i} - \boldsymbol{\mu}_\mathbf{com})^T = \mathbf{A_{com}A_{com}^T}, \quad (10)$$

where $\boldsymbol{\mu}_\mathbf{com}$ is the mean of all common vectors. In this case the optimal projection vectors $\mathbf{w_k}$ can be found by an eigen-analysis of $\mathbf{S_{com}}$. In particular, all eigenvectors corresponding to the nonzero eigenvalues of $\mathbf{S_{com}}$ will be the optimal projection vectors and they can be easily found by using the smaller, *C*x*C* matrix $\mathbf{A_{com}^TA_{com}}$.

Since the optimal projection vectors $\mathbf{w_k}$ come from the null space of $\mathbf{S_W}$, it follows that when the image samples $\mathbf{x_m^i}$ of the *i*-th class are projected onto the linear span of the projection vectors $\mathbf{w_k}$, the feature vector $\boldsymbol{\Omega}_\mathbf{i} = [<\mathbf{x_m^i}, \mathbf{w_1}> \quad \ldots \quad <\mathbf{x_m^i}, \mathbf{w_r}>]^T$ of the projection coefficients $<\mathbf{x}_{\mathbf{m}}^{\mathbf{i}},\mathbf{w}_{\mathbf{k}}>$ will also be independent of the sample index *m*. Thus, we have, $\mathbf{\Omega}_{\mathbf{i}}=\mathbf{W}^{\mathbf{T}}\mathbf{x}_{\mathbf{m}}^{\mathbf{i}}$ for each class. The fact that $\mathbf{\Omega}_{\mathbf{i}}$ does not depend on the index *m* guarantees 100% accuracy in the recognition of the samples in the training set. We called these vectors as the discriminative common vectors.

To recognize a test image $\mathbf{x}_{\mathbf{test}}$, the feature vector of the test image is found by,

$$\mathbf{\Omega}_{\mathbf{test}}=\mathbf{W}^{\mathbf{T}}\mathbf{x}_{\mathbf{test}} \tag{11}$$

and $\mathbf{\Omega}_{\mathbf{test}}$ is compared with the discriminative common vector $\mathbf{\Omega}_{\mathbf{i}}$ of each class using the Euclidean distance. The discriminative common vector found to be the closest to $\mathbf{\Omega}_{\mathbf{test}}$ is used to identify the test image.

The above method can be summarized as follows:

Step 1: Compute the nonzero eigenvalues and corresponding eigenvectors $\boldsymbol{\alpha}_{\mathbf{k}}$ of $\mathbf{S}_{\mathbf{W}}$ by using the matrix $\mathbf{A}^{\mathbf{T}}\mathbf{A}$, where $\mathbf{S}_{\mathbf{W}}=\mathbf{A}\mathbf{A}^{\mathbf{T}}$ and $\mathbf{A}$ is given by (5). Set $\mathbf{Q}=[\boldsymbol{\alpha}_1 \quad ... \quad \boldsymbol{\alpha}_r]$, where *r* is the rank of $\mathbf{S}_{\mathbf{W}}$.

Step 2: Choose any sample from each class and project it onto the null space of $\mathbf{S}_{\mathbf{W}}$ to obtain the common vectors.

$$\mathbf{x}_{\mathbf{com}}^{\mathbf{i}}=\mathbf{x}_{\mathbf{m}}^{\mathbf{i}}-\mathbf{Q}\mathbf{Q}^{\mathbf{T}}\mathbf{x}_{\mathbf{m}}^{\mathbf{i}}, \quad m=1,...,N, \quad i=1,...,C. \tag{12}$$

Step 3: Form the matrix $\mathbf{S}_{\mathbf{com}}=\mathbf{A}_{\mathbf{com}}\mathbf{A}_{\mathbf{com}}^{\mathbf{T}}$ by using the common vectors and compute the eigenvectors $\mathbf{w}_{\mathbf{k}}$ of $\mathbf{S}_{\mathbf{com}}$ that correspond to the nonzero eigenvalues by using the matrix $\mathbf{A}_{\mathbf{com}}^{\mathbf{T}}\mathbf{A}_{\mathbf{com}}$. Use these eigenvectors to form the projection matrix $\mathbf{W}=[\mathbf{w}_1 \quad ... \quad \mathbf{w}_{\mathbf{K}}]$. Here, $K \le C$-1 refers to the rank of $\mathbf{S}_{\mathbf{com}}$.

4. Obtaining the Common Vectors by Using Subspace Methods

We used the *M*x*M* matrix $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ to find the orthonormal vector set that spans the range of $\mathbf{S}_{\mathbf{W}}$ in the course of obtaining the common vector of each class. However, the computations may become numerically unstable for large values of *M*. To overcome this problem, we can use the subspace methods and the Gram-Schmidt orthogonalization procedure to obtain the common vectors.

Firstly, we choose any one of the image vectors from the *i*-th class as the subtrahend vector and then obtain the difference vectors $\mathbf{b}_{\mathbf{k}}^{\mathbf{i}}$ of the difference subspace of the *i*-th class. Thus, assuming that the first sample of each class is taken as the subtrahend vector, the difference vectors are $\mathbf{b}_{\mathbf{k}}^{\mathbf{i}}=\mathbf{x}_{\mathbf{k+1}}^{\mathbf{i}}-\mathbf{x}_{\mathbf{1}}^{\mathbf{i}}$, $k=1,...,N-1$. The difference subspace B_i of the *i*-th class is defined as $B_i=span\{\mathbf{b}_{\mathbf{1}}^{\mathbf{i}},....,\mathbf{b}_{\mathbf{N-1}}^{\mathbf{i}}\}$. These subspaces can be summed up to form the complete difference subspace

$$B=B_1+...+B_C=span\{\mathbf{b}_{\mathbf{1}}^{1},...,\mathbf{b}_{\mathbf{N-1}}^{1},\mathbf{b}_{\mathbf{1}}^{2},...,\mathbf{b}_{\mathbf{N-1}}^{C}\}. \tag{13}$$

The complement of *B* will be called the indifference subspace $B^{\perp}$. The number of linearly independent difference vectors that span the complete difference subspace *B* of the training set samples is equal to the rank of $\mathbf{S}_{\mathbf{W}}$. The linearly independent difference vectors can be orthonormalized by using the Gram-Schmidt orthogonalization procedure to obtain the orthonormal basis vector set, $\{\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_{\mathbf{r}}\}$. We can form the matrix $\mathbf{Q}=[\boldsymbol{\beta}_1 \quad ... \quad \boldsymbol{\beta}_{\mathbf{r}}]$, and then any sample from each class can now be projected onto the indifference subspace $B^{\perp}$ to obtain the corresponding common vectors of the classes by using (12). The common vectors do not depend on the choice of the subtrahend vectors and are identical to the common vectors obtained by using the null space of $\mathbf{S}_{\mathbf{W}}$ [9]. After calculating the common vectors, the optimal projection vectors can be found by performing PCA. However, the optimal projection vectors can also be found by computing the basis of the difference subspace B_{com} of the common vectors.

5. Experimental Results

We used the AR-face database [10] to test our method. 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.

We randomly selected *C*=50 individuals (30 males and 20 females) for the experiment. Only nonoccluded images ((1)-(7) and (14)-(20) as in Fig. 1) were chosen for every subject. Thus, our face database size was 700 with 14 images per subject. After, these images were converted to grayscale, we pre-processed them by aligning and scaling so that the distances between the eyes were the same for all images. Then, we cropped the resulting images ensuring that the eyes occurred in the same coordinates of the images. 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 [11] was employed using 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 I.

0-7695-2128-2/04 $20.00 (C) 2004 IEEE

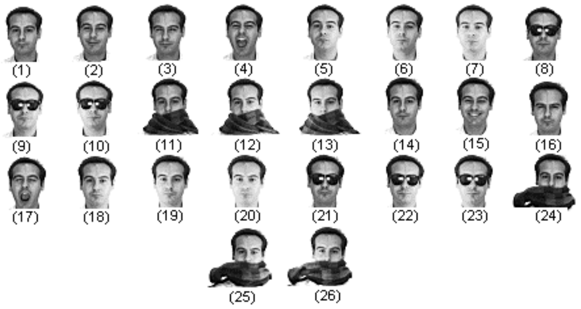

Figure 1. Images of one subject in the AR-face database. Only nonoccluded images (1)-(7) and (14)-(20) were used in our experiments.

Table I. Recognition rates for the AR-face database

Method	Recognition Rate
Eigenface	79.14
Fisherface	98.85
Direct-LDA	98.64
Discriminative Common Vector	99.35

6. Conclusion

In this paper we proposed new algorithms for obtaining the optimal projection vectors efficiently in the null space of the within-class scatter matrix of the training set samples. We showed that every sample in a particular class produces the same unique common vector when they are projected onto the null space of $\mathbf{S}_w$. Using the common vectors 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 the 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 also show that our method classifies the test set data better than other methods. Using subspace methods overcomes the numerical stability problem encountered frequently when working with high dimensional matrices. Since feature vector of the test image is only compared to a single vector for each class during classification, 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 is usually compared to all feature vectors of samples in the training set, making these methods impractical for real-time applications for large training sets.

7. 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 3^{rd} IEEE International Conference on Automatic Face and Gesture Recognition*, April 1998, pp. 336-341.

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

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] H. Cevikalp, M. Neamtu, M. Wilkes, A. Barkana, "Discriminative common vectors for face recognition," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, submitted for publication.

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

[11] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* 2^{nd} edition, New York: Academic Press, 1990, pp. 220-221.