TWO DIMENSIONAL FISHER DISCRIMINANT ANALYSIS: FORGET ABOUT SMALL SAMPLE SIZE PROBLEM

Hui Kong, Eam Khwang Teoh

Nanyang Technological University
Singapore 639798
{pg03802060, eekteoh}@ntu.edu.sg

Jian Gang Wang, Ronda Venkateswarlu

Institute for Infocomm Research
Singapore 119613
{jgwang, vronda}@i2r.a-star.edu.sg

ABSTRACT

This paper addresses the Small Sample Size (SSS) problem in Linear Discriminant Analysis (LDA) utilizing a so called Two-Dimensional Fisher Discriminant Analysis (2D-FDA) algorithm. As opposed to traditional LDA-based approaches, 2D-FDA is based on 2D image matrices rather than 1D vectors so the image matrix does not need to be transformed into a vector before feature extraction. The between-class scatter and the within-class scatter is constructed using the original image matrices. The advantage arising in this way is that the SSS problem existing in traditional linear discriminant analysis does not occur any more. To test the performance of 2D-FDA with small number of training samples, a series of experiments are conducted on two public databases: ORL and Yale face database B. In both two trials, the 2D-FDA outperforms the other linear subspace methods when there are only very limited training images for each subject.

1. INTRODUCTION

The past ten years witness the progress of the LDA-based face recognition algorithms. LDA/FDA is theoretically one of the best classification methods [1]. The two noted methods using FDA are Swets and Weng's Discriminant Eigenfeatures [2] for image retrieval and Belhumeur et al.'s Fisherface [3] for face recognition, they both achieve better results than Eigenface method [4]. Similar to Eigenface method, both transform the 2D image matrix into a 1D vector before feature extraction. However, the within-class scatter matrix is almost singular after such a transform because the number of training samples is so small compared with the dimension of the image vector. This degeneration results in great difficulty in solving the inverse of within-class scatter.

To solve the SSS problems, various schemes have been proposed so far. In Swets and Weng's Discriminant Eigenfeatures and Belhumeur et al.'s Fisherface, they both used a two stage PCA+LDA approach. Using PCA, the high dimensional face data is projected to a low dimensional space and then LDA is performed in this PCA subspace. However, the discarded subspace may also encode some information helpful for recognition, this removal may introduce a loss of discriminative information. Chen et al. [5] suggested that the null space spanned by the eigenvectors of S_w with zero eigenvalues contains the most discriminative information. A LDA method in the null space of S_w was proposed, called N-LDA. However, as explained in [5], with the existence of noise, when the training sample number is large, the null space of S_w becomes small, so much discriminative information outside this null space will be lost. Another shortcoming is this approach involves computing eigenvalue problem for a very high-dimension matrix.

Yu and Yang [6] proposed a new algorithm which incorporates the concept of null space. It first removes the null space of the between-class scatter matrix S_b and seeks a projection to minimize the within-class scatter (called Direct LDA /DLDA). Because the rank of S_b is smaller than that of S_w, removing the null space of S_b may lose part of or the entire null space of S_w, which is very likely to be full-rank after the removing operation.

Huang et al. [7] introduced a more efficient null space approach. The basic notion is that the null space of S_w is particularly useful in discriminating ability, whereas, that of S_b is useless. They proved that the null space of the total scatter matrix S_t is the common null space of both S_w and S_b. Hence the algorithm firstly removes the null space of S_t and projects the samples onto the null space of S_w. Then it removes the null space of the between-class scatter in the subspace to get the optimal discriminant vectors.

Wang and Tang [8] gave a random sampling LDA for face recognition with small training sample. This paper analyzes that both Fisherface and N-LDA encounter respective over-fitting problem for different reasons. To solve it, in Fisherface, they apply random subspace to reduce the feature vector dimension to reduce the discrepancy. In N-LDA, this problem can be alleviated by bagging, since each replicate has a smaller number of training samples. A fusion rule is adopted to combine these two kinds of random sampling based classifiers. Recently, a Two Dimensional Princi-

pal Component Analysis (2D-PCA)[9] method is proposed. As opposed to traditional PCA-based approaches, 2D-PCA is based on 2D image matrices rather than 1D image vectors. However, like PCA, 2D-PCA is only good at image representation rather than discrimination. When there are large pose and illumination variations in face images, the top eigenvectors in 2D-PCA does not models identity information but these external variations.

In this paper, inspired by 2D-PCA, a 2D-FDA algorithm is proposed for face recognition with small number of training samples. In 2D-FDA, the between-class scatter and the within-class scatter is constructed using the image matrices. In contrast to the between-class and within-class scatter of FDA, the within-class scatter obtained by 2D-FDA is not singular generally. As a result, the 2D-FDA has three important advantages over the 2D-PCA, original FDA and N-LDA. Firstly, the features are extracted using Fisher discriminant analysis, not the PCA, thus the discriminating ability is better than 2D-PCA. Secondly, it does not encounter SSS problem any more when the training sample size is small. Thirdly, it takes full advantage of the discriminative information in the face space, and does not discard any subspace which may be valuable for recognition.

2. TWO DIMENSIONAL FISHER DISCRIMINANT ANALYSIS

Let $\mathbf{x}$ denote an n-dimensional unitary column vector. The idea is to project image $\mathbf{A}$, an $m \times n$ matrix, onto $\mathbf{x}$ by the following linear transformation: $\mathbf{y} = \mathbf{A}\mathbf{x}$. Thus, we obtain an m-dimensional projected vector $\mathbf{y}$, which is called the projected feature vector of image $\mathbf{A}$. How do we determine the optimal project direction $\mathbf{x}$? In fact, the discriminatory power of the projection vector $\mathbf{x}$ can be measured by the Fisher criterion [1], i.e., maximizing the between-class scatter of the projected samples and meantime minimizing the within-class scatter of the projected samples. It is known that the scatter of the projected samples can be characterized by the trace of the covariance matrix of the projected feature vectors. From this point of view, the Fisher criterion is adopted as follows:

$$J(\mathbf{x}) = \frac{tr(\mathbf{PS}_b)}{tr(\mathbf{PS}_w)} \tag{1}$$

where $\mathbf{PS}_b$ and $\mathbf{PS}_w$ are the between-class covariance and the within-class covariance of the projected samples, $tr(\mathbf{PS}_b)$ denotes the trace of $\mathbf{PS}_b$, $tr(\mathbf{PS}_w)$ denotes the trace of $\mathbf{PS}_w$.

Lemma 1 : Let $\mathbf{S}_b$, $\mathbf{S}_w$ be the between-class and within-class covariance of the original image matrix. The trace of $\mathbf{PS}_b$, $tr(\mathbf{PS}_b) = \mathbf{x}^{'}\mathbf{S}_b\mathbf{x}$ and the trace of $\mathbf{PS}_w$, $tr(\mathbf{PS}_w) = \mathbf{x}^{'}\mathbf{S}_w\mathbf{x}$.

Proof: Let $\overline{\mathbf{M}}$ be the mean of all the training samples, $\overline{\mathbf{M}}_i$ be the mean of each class, $\overline{\mathbf{M}}^p$ be the mean of all the projected samples, $\overline{\mathbf{M}}_i^p$ be the mean of each projected class.

The between-class covariance of the projected samples, $\mathbf{PS}_b = E[(\overline{\mathbf{M}}_i^p - \overline{\mathbf{M}}^p)\,(\overline{\mathbf{M}}_i^p - \overline{\mathbf{M}}^p)^T] = E[(\overline{\mathbf{M}}_i\mathbf{x} - \overline{\mathbf{M}}\mathbf{x})\,(\overline{\mathbf{M}}_i\mathbf{x} - \overline{\mathbf{M}}\mathbf{x})^T] = E[((\overline{\mathbf{M}}_i - \overline{\mathbf{M}})\mathbf{x})\,((\overline{\mathbf{M}}_i - \overline{\mathbf{M}})\mathbf{x})^T]$.

Because the *trace* of a square matrix is the summation of all the leading diagonal elements. For any two matrix, $\mathbf{A}_{r\times s}$ and $\mathbf{B}_{s\times r}$, we have:

$$tr(\mathbf{AB}) = tr(\mathbf{BA}),$$

Since,

$$\sum_{i=1}^{r}\sum_{j=1}^{s} a_{ij}b_{ji} = \sum_{j=1}^{s}\sum_{i=1}^{r} b_{ji}a_{ij}$$

So,

$$\begin{aligned} tr(\mathbf{PS}_b) &= tr(E[((\overline{\mathbf{M}}_i - \overline{\mathbf{M}})\mathbf{x})\,((\overline{\mathbf{M}}_i - \overline{\mathbf{M}})\mathbf{x})^T]) \\ &= tr(E[((\overline{\mathbf{M}}_i - \overline{\mathbf{M}})\mathbf{x})^T\,((\overline{\mathbf{M}}_i - \overline{\mathbf{M}})\mathbf{x})]) \\ &= tr(E[\mathbf{x}^T(\overline{\mathbf{M}}_i - \overline{\mathbf{M}})^T\,(\overline{\mathbf{M}}_i - \overline{\mathbf{M}})\mathbf{x}]) \\ &= E[\mathbf{x}^T(\overline{\mathbf{M}}_i - \overline{\mathbf{M}})^T\,(\overline{\mathbf{M}}_i - \overline{\mathbf{M}})\mathbf{x}] \\ &= \mathbf{x}^T E[(\overline{\mathbf{M}}_i - \overline{\mathbf{M}})^T\,(\overline{\mathbf{M}}_i - \overline{\mathbf{M}})]\mathbf{x} \\ &= \mathbf{x}^T\mathbf{S}_b\mathbf{x} \end{aligned}$$

where $\mathbf{S}_b = E[(\overline{\mathbf{M}}_i - \overline{\mathbf{M}})^T\,(\overline{\mathbf{M}}_i - \overline{\mathbf{M}})]$. We have the same steps for the proof of $tr(\mathbf{PS}_w) = \mathbf{x}^{'}\mathbf{S}_w\mathbf{x}$, where $\mathbf{S}_w = E[(\mathbf{A} - \overline{\mathbf{M}}_i)^T\,(\mathbf{A} - \overline{\mathbf{M}}_i)|\mathbf{A} \in C_i]$, C_i is the i-th class. □

Therefore, the Fisher criterion in Eq.1 can be converted into:

$$J(\mathbf{x}) = \frac{\mathbf{x}^T\mathbf{S}_b\mathbf{x}}{\mathbf{x}^{'}\mathbf{S}_w\mathbf{x}} \tag{2}$$

The unitary vector $\mathbf{x}$ that maximizes Eq.2 is called the optimal discriminating projection axis. In general, it is not enough to have only one optimal projection axis. It is necessary to select a set of projection directions, $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_d]$. Note that Eq.2 has the standard form for 1D Fisher discriminant analysis.

Theorem 1 : The $\mathbf{S}_w$ in 2D-FDA is not singular.

Proof : Since $\mathbf{S}_w = E[(\mathbf{A} - \overline{\mathbf{M}}_i)^T\,(\mathbf{A} - \overline{\mathbf{M}}_i)|\mathbf{A} \in C_i]$, another form of $\mathbf{S}_w$ can be written as $\mathbf{S}_w = \frac{1}{N}\Phi_{Sw}^T\Phi_{Sw}$, where N is the total training number, $\Phi_{Sw}^T = [\phi_{Sw}^1, \phi_{Sw}^2, \cdots, \phi_{Sw}^L]$, L is the total class number, $\phi_{Sw}^i = [(A_i^1 - \overline{M_i})^T, (A_i^2 - \overline{M_i})^T, \cdots, (A_i^{L_i} - \overline{M_i})^T]$, $i = 1, \cdots, L$. L_i is the number of training samples in i-th class. $\overline{M_i}$ is the mean of each class. A_j^i, $j = 1, \cdots, L_i$, is the j-th training sample in the i-th class. The dimension of Φ_{Sw}^T is $n \times (m\sum_{i=1}^{L} L_i)$, where m and n are the image height and width. Since $rank(\Phi_{Sw}^T\Phi_{Sw}) = rank(\Phi_{Sw}\Phi_{Sw}^T) = rank(\Phi_{Sw}^T)$ and $rank(\Phi_{Sw}^T) = n$[1], additionally, the dimension of $\mathbf{S}_w$ is $n \times n$, we can conclude that $\mathbf{S}_w$ is of full rank. □

[1]It is known that $rank(\Phi_{Sw}^T) \le \min(n, m\sum_{i=1}^{L} L_i)$ and $n \ll (m\sum_{i=1}^{L} L_i)$ in the area of visual pattern recognition, there is, $rank(\Phi_{Sw}^T) \le n$. Further, we take an assumption that the rows of Φ_{Sw}^T are independent of each other (the experiment results will demonstrate that this assumption can be well satisfied in the benchmark databases), it can be obtained that $rank(\Phi_{Sw}^T) = n$

The optimal projection vectors $\mathbf{X}_{opt}$ can be obtained by directly solving the following generalized eigen-system.

$$\mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{X}_{opt} = \Lambda\mathbf{X}_{opt} \quad (3)$$

where Λ is the diagonal matrix whose diagonal elements are eigenvalues of $\mathbf{S}_w^{-1}\mathbf{S}_b$.

3. FEATURE EXTRACTION

The optimal projection features of 2D-FDA, $\mathbf{x}_1, \cdots, \mathbf{x}_d$, are used for feature extraction. For a given image sample $\mathbf{A}$, let

$$\mathbf{y}_k = \mathbf{A}\mathbf{x}_k, k = 1, 2, \cdots, d. \quad (4)$$

Then we obtain a family of projected feature vectors, $\mathbf{y}_1, \cdots, \mathbf{y}_d$, which are called *Fisher feature vectors* of the sample image $\mathbf{A}$. These *Fisher feature vectors* are used to form an $m \times d$ *Fisher feature matrix* $\mathbf{B} = [\mathbf{y}_1, \cdots, \mathbf{y}_d]$.

4. CLASSIFICATION METHOD

After a transformation by 2D-FDA, a *Fisher feature matrix* is obtained for each image. Then, a nearest neighbor classifier is used for classification. Here, the distance between two arbitrary *Fisher feature matrices*, $\mathbf{B}_i = [\mathbf{y}_1^i, \cdots, \mathbf{y}_d^i]$ and $\mathbf{B}_j = [\mathbf{y}_1^j, \cdots, \mathbf{y}_d^j]$, is defined by

$$dist(\mathbf{B}_i, \mathbf{B}_j) = \sum_{k=1}^{d} \|\mathbf{y}_k^i - \mathbf{y}_k^j\|_2 \quad (5)$$

where $\|\mathbf{y}_k^i - \mathbf{y}_k^j\|_2$ denotes the Euclidean distance between the two *Fisher feature vectors* $\mathbf{y}_k^i$ and $\mathbf{y}_k^j$.

Suppose that the training samples are $\mathbf{B}_1, \cdots, \mathbf{B}_M$ (where *M* is the total number of training samples), and that each of these samples is assigned a given identity (class) C_k. Given a test sample $\mathbf{B}$, if $dist(\mathbf{B}, \mathbf{B}_l) = \min_{i=1}^{M} dist(\mathbf{B}, \mathbf{B}_i)$, and $\mathbf{B}_l \in C_t$, then the resulting decision is $\mathbf{B} \in C_t$.

5. PRINCIPAL COMPONENT ANALYSIS OF THE FISHER FEATURE MATRIX

From the above classification step, we can see that the computation load for 2D-FDA is much heavier than 1D-LDA. To reduce computation time, the *Fisher feature matrix* are further condensed by PCA. In the classification step, the feature vectors extracted from PCA of the *Fisher feature matrix* are used. The whole procedure for the PCA of the *Fisher feature matrix* and classification is listed as follows:

1. Transform the corresponding *Fisher feature matrix*, $\mathbf{B}_i$, for each image, $\mathbf{A}_i$ into a 1D feature vector.

2. Apply PCA to all the 1D feature vectors and obtain a subspace consisting of a set of eigenvectors.

3. Project all the 1D feature vectors onto the subspace, finally a lower-dimensional feature vector is obtained for each image.

4. Recognition using nearest neighbor classifier on the finally acquired lower-dimensional feature vectors.

6. EXPERIMENTS AND ANALYSIS

The proposed 2D-FDA method is used for face recognition and tested on two well-known face image databases (ORL, Yale face database B[10]). The ORL database is used to evaluate the performance of 2D-FDA under conditions where the pose, face expression, face scale vary. The Yale face database B is used to examine the system performance when illumination varies extremely.

6.1. Experiments on the ORL Database

The ORL database (http://www.cam-orl.co.uk) contains images from 40 individuals, each providing 10 different images. The facial expressions and facial details (glasses or no glasses) also vary. The images were taken with a tolerance for some tilting and rotation of the face of up to 20 degrees. Moreover, there is also some variation in the scale of up to about 10 percent. All images are grayscale and normalized to a resolution of 46×56 pixels. Ten sample images of two persons from the ORL database are shown in Fig.1. We test the recognition performance with different training

Fig. 1. Ten sample images of two subjects in ORL database

numbers. k ($2 \leq k \leq 9$) images of each subject are randomly selected for training and the remaining 10-k images of each subject for testing. For each number k, 50 runs are performed with different random partition between training set and testing set. Fig.2 shows the average recognition rate. The dimension for the *Fisher feature matrix* and *Eigen feature matrix* of 2D-FDA and 2D-PCA is 56×3. From Fig.2, it can be seen that the performance of 2D-FDA is much better than the other linear subspace methods, the superiority is more obvious when the number of training sample is small.

6.2. Experiments on Yale Face Database B

This database contains 5760 images of 10 subjects each seen under 576 viewing conditions (9 poses x 64 illumination conditions). Twenty sample images of two persons

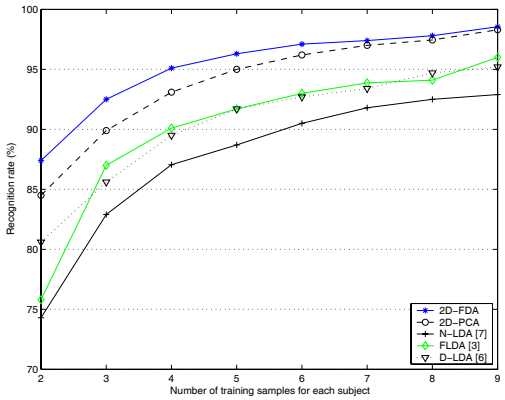

Fig. 2. Recognition rate on the ORL database

Fig. 3. Twenty sample images of two subjects in Yale face database B

from the Yale face database B are shown in Fig.3. In our experiment, altogether 640 images for 10 subjects are used (64 illumination conditions under the same frontal pose). The image size is 50×60. The recognition performance is tested with different training numbers. k ($2 \leq k \leq 12$) images of each subject are randomly selected for training and the remaining 64-k images of each subject for testing. For each number k, 100 runs are performed with different random partition between training set and testing set. Fig.4 shows the average recognition rate. The dimension for the *Fisher feature matrix* of 2D-FDA and *Eigen feature matrix* of 2D-PCA is 60×22. But with the PCA of the *Fisher feature matrix* and the *Eigen feature matrix*, the dimension is reduced to a vector with length of 639. From Fig.4, it can be seen that 2D-FDA outperforms greatly the other linear subspace methods when the training sample number is small.

7. CONCLUSIONS

In this paper, a novel 2D-FDA algorithm for face recognition is proposed. This method has great advantage over the other linear LDA algorithms: Small sample size problem arising from few training samples does not exist any more. In addition, the algorithm is very simple to be implemented. The performance is much better than the current LDA-based algorithms when there exist limited training samples for each class.

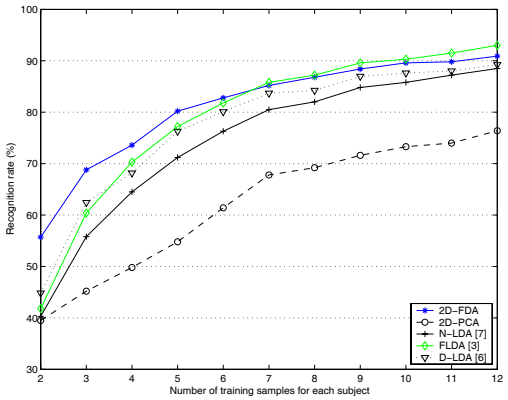

Fig. 4. Recognition rate on the Yale face database B

8. REFERENCES

[1] K. Fukunnaga, "Introduction to Statistical Pattern Recognition," Academic Press, second edition, 1991.

[2] D. L. Swets and J. Weng, "Using Discriminant Eigenfeatures for Image Retrieval," IEEE Trans. on PAMI, 1996.

[3] P. N. Belhumeur and J. Hespanha and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," IEEE Trans. on PAMI, 1997.

[4] M. Turk and A. Pentland, "Eigenfaces for Recognition," Journal of Cognitive Neuroscience, 1991.

[5] L. Chen, H. Liao, M. Ko, J. Liin, and G. Yu, "A New LDA-based Face Recognition System Which can Solve the Samll Sample Size Problem," Pattern Recognition, 2000.

[6] H. Yu and J. Yang, "A direct lda algorithm for high-dimensional data with application to face recognition," Pattern Recognition, 2001.

[7] R. Huang, Q. S. Liu, H. Q. Lu and S. D. Ma, "Solving the Small Sample Size Problem of LDA," Proc. IEEE ICPR, 2002.

[8] X. Wang and X. Tang, "Random sampling LDA for face recognition," Proc. IEEE Conf. CVPR, 2004.

[9] J. Yang, D. Zhang, A. F. Frangi and J. Yang, "Two-Dimensional PCA: A New Approach to Appearance-Based Face Representation and Recognition," IEEE Trans. on PAMI, 2004.

[10] A. S. Georghiades, P. N. Belhumeur, D. J. Kriegman, "From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose," IEEE Trans. on PAMI, 2001.