FACE RECOGNITION USING RECURSIVE FISHER LINEAR DISCRIMINANT WITH GABOR WAVELET CODING

C. Xiang, X. A. Fan, and T. H. Lee

Department of Electrical and Computer Engineering,
National University of Singapore
Email: elexc@nus.edu.sg

ABSTRACT

The constraint on the total number of features available from Fisher Linear Discriminant (FLD) has seriously limited its application to a large class of problems. In order to overcome this disadvantage of FLD, a recursive procedure of calculating the discriminant features is suggested in this paper. Extensive experiments of comparing the new algorithm with the traditional PCA and FLD approaches have been carried out on face recognition problem, in which the resulting improvement of the performances by the new feature extraction scheme is significant.

1. INTRODUCTION

Extracting proper features is crucial for satisfactory design of any pattern classifier. Traditionally, principal component analysis (PCA) has been the standard approach to reduce the high-dimensional original pattern vector space into low-dimensional feature vector space. An alternative approach using Fisher linear discriminant (FLD) [1,2] has gained popularity recently following a number of successful applications of FLD to face recognition problem in 1990's [3-9].

Although FLD has proven to be more efficient than PCA in many of the applications mentioned above, there is a serious limitation which is that the total number of the features available from FLD is limited to $c-1$, where c is the number of classes. Therefore it is essential to eliminate this constraint completely if possible such that FLD can be applied to a much wider class of pattern classification problems. It is for this purpose that we wish to suggest a recursive procedure for extracting FLD features (which will be abbreviated as RFLD), which constitutes the main contribution of this paper.

In order to verify whether this new approach would bring any advantages over FLD as well as PCA, we choose to carry out experiments on face recognition problem. All of the experimental results have unanimously demonstrated that the performance of the classifier can be improved significantly by RFLD compared to both FLD and PCA.

In the following section, a brief introduction of FLD and the detailed algorithm for RFLD will be presented. The experimental results on face recognition problem will be discussed in section three.

2. FLD VS. RFLD

2.1. Fisher Linear Discriminant (FLD)

Suppose that we have a set of n d-dimensional samples $x_1,\cdots,x_n$ belonging to c different classes with n_i samples in the subset D_i labeled ω_i, $i=1,\cdots,c$. Then the objective of Fisher linear discriminant is to seek the direction w not only maximizing the between-class scatter of the projected samples, but also minimizing the within-class scatter, such that the following criterion function:

$$J(w)=\frac{w^T S_B w}{w^T S_W w}, \tag{1}$$

is maximum, where the between-class scatter matrix, S_B, is defined by

$$S_B=\sum_{i=1}^{c} n_i (m_i-m)(m_i-m)^T, \tag{2}$$

in which m is the sample mean of the whole set,

$$m=\frac{1}{n}\sum_{k=1}^{n} x_k, \tag{3}$$

and m_i is the sample mean for class labeled ω_i given by

$$m_i=\frac{1}{n_i}\sum_{x\in D_i} x, \tag{4}$$

and the within-class scatter matrix, S_W, is defined by

$$S_W=\sum_{i=1}^{c} S_i, \tag{5}$$

where the scatter matrix, S_i, corresponding to class ω_i is defined by

$$S_i=\sum_{x\in D_i}(x-m_i)(x-m_i)^T, \quad i=1,\cdots,c. \tag{6}$$

It is easy to show that a vector w that maximizes $J(w)$ must satisfy

$$S_B w=\lambda S_W w. \tag{7}$$

If S_W is nonsingular we can obtain a conventional eigenvalue problem by writing

$$S_W^{-1} S_B w=\lambda w. \tag{8}$$

It is obvious that the at most $c-1$ features may be extracted from above procedure simply because the rank of S_B is at most $c-1$. In order to eliminate this upper bound on the total number of discriminant features, a recursive procedure applying essentially the same basic idea of FLD is proposed and will be described in detail in the following section.

2.2. Recursive Fisher Linear Discriminant (RFLD)

Instead of extracting feature vectors from an eigenvalue problem of $S_W^{-1}S_B$ once and for all, the feature vectors will be obtained

recursively, step by step. At every step, the calculation of a new feature vector will be based upon all the feature vectors obtained previously. More specifically, at each step when a new feature vector is calculated, the training samples have to be pre-processed such that all the information represented by those "old" features will be discarded, i.e., the projections of the sampled vectors on those "old" features will be eliminated. And then the problem of extracting the new feature most efficient for classification based upon the pre-processed database will be formulated in the same fashion as that of FLD.

Let's consider a set of n d-dimensional samples $x_1,\cdots,x_n$ belonging to c classes, as discussed in section 2.1. The first RFLD feature vector w_1 will be the same as that of FLD, which is the normalized eigenvector associated with the largest eigenvalue of matrix $S_W^{-1}S_B$, where the between-class scatter matrix S_B and within-class scatter matrix S_W are defined by (2) and (5) respectively.

Comment 1: If the number of samples n is smaller than the dimension d as is the case of face recognition problem, the within-class scatter matrix S_W may be singular, which makes the maximum value defined by (1) to be infinity. PCA is then usually employed to reduce the dimension first such that S_W is nonsingular as suggested by Belhumeur *et al.* [3]. From now on, we will always assume that S_W is non-singular.

Before the second feature w_2 is computed, the information represented by the first feature vector w_1 is first discarded from all the sampled vectors x_i as follows,

$$x_i^{(2)} = x_i - (w_1^T x_i)w_1, \quad i = 1,2,.....,n. \tag{9}$$

Based on this new set of sampled vectors $x_1^{(2)},\cdots,x_n^{(2)}$, the sample means for the whole set, $m^{(2)}$, as well as for individual classes, $m_i^{(2)}$, are calculated as follows according to the standard definitions in section 2.1.

$$m^{(2)} = m - (w_1^T m)w_1, \tag{10}$$

and

$$m_i^{(2)} = m_i - (w_1^T m_i)w_1, \quad i = 1,2,.....,c. \tag{11}$$

The new between-class scatter matrix $S_B^{(2)}$ and within-class scatter matrix $S_W^{(2)}$ may then be computed by

$$S_B^{(2)} = \sum_{i=1}^{c} n_i (m_i^{(2)} - m^{(2)})(m_i^{(2)} - m^{(2)})^T, \tag{12}$$

and

$$S_W^{(2)} = \sum_{i=1}^{c} S_i^{(2)}, \tag{13}$$

where as before, the scatter matrix $S_i^{(2)}$ corresponding to class ω_i is defined by

$$S_i^{(2)} = \sum_{x^{(2)} \in D_i} (x^{(2)} - m_i^{(2)})(x^{(2)} - m_i^{(2)})^T. \tag{14}$$

The objective is to seek direction w_2 that maximize the same criterion function defined by (1), just replacing S_B and S_W with $S_B^{(2)}$ and $S_w^{(2)}$ respectively,

$$J^{(2)}(w) = \frac{w^T S_B^{(2)} w}{w^T S_W^{(2)} w}. \tag{15}$$

Similarly, it is easy to show that the optimal solution w_2 has to satisfy

$$S_B^{(2)} w_2 = \lambda S_W^{(2)} w_2. \tag{16}$$

Since $S_W^{(2)}$ is not of full rank but of rank $d-1$, the above equation cannot be directly reduced to a conventional eigenvalue problem as before. Because the information represented by previous feature w_1 has already been dismissed from the samples, it is natural to impose the following condition that the new feature w_2 is orthogonal to w_1, i.e.,

$$w_1^T w_2 = 0, \tag{17}$$

such that the old feature w_1 is excluded from the possible solutions to equation (16).

Combining (16) and (17) results in

$$B_2 w_2 = \lambda W_2 w_2, \tag{18}$$

where the $(d+1)\times d$ matrices

$$W_2 = \begin{bmatrix} S_W^{(2)} \\ w_1^T \end{bmatrix} \quad \text{and} \quad B_2 = \begin{bmatrix} S_B^{(2)} \\ 0 \end{bmatrix}. \tag{19}$$

Using the fact that W_2 is of full rank, equation (18) can be reduced to

$$(W_2^T W_2)^{-1} W_2^T B_2 w_2 = \lambda w_2, \tag{20}$$

which becomes a conventional eigenvalue problem, and w_2 can be obtained as the normalized eigenvector with the largest eigenvelue of square matrix $(W_2^T W_2)^{-1} W_2^T B_2$.

Similarly it can be readily shown that, the k^{th} feature vector w_k may be computed as the normalized eigenvector with the largest eigenvelue from following eigenvalue problem,

$$(W_k^T W_k)^{-1} W_k^T B_k w_k = \lambda w_k, \tag{21}$$

where the $(d+k-1)\times d$ matrices

$$W_k = \begin{bmatrix} S_W^{(k)} \\ w_1^T \\ \vdots \\ w_{k-1}^T \end{bmatrix} \quad \text{and} \quad B_k = \begin{bmatrix} S_B^{(k)} \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{22}$$

where the within-class scatter matrix $S_w^{(k)}$ and the between-class scatter matrix $S_B^{(k)}$ are calculated from the pre-processed samples $x_i^{(k)}$ in which all the information represented by previous features are eliminated in the way of

$$x_i^{(k)} = x_i^{(k-1)} - (w_{k-1}^T x_i^{(k-1)})w_{k-1}, \quad i = 1,2,.....,n. \tag{23}$$

This recursive process may continue as long as matrix B_k is not a zero matrix, i.e., the between-class scatter $S_B^{(k)}$ is not a zero matrix. When $S_B^{(k)}$ is a zero matrix, the process naturally stops because the between-class scatter is zero and cannot be maximized by projection any more. It is obvious from above procedure that not only has the constraint of at most $c-1$ features been totally eliminated, but also is each feature an optimal direction to maximize the between-class scatter relative to the within-class scatter for the sampled data under orthogonal conditions, which in turn provides a sound basis for exploiting them as features for classification purpose.

Comment 2: RFLD bears the similar idea as suggested by Iterative Linear Classification (ILC) [10], to generate new sample set by projecting the samples into a subspace that is orthogonal to selected features. However, different from support vector

machine (SVM) applied in [10], the similar recursive procedure is applied to FLD in this paper.

It is observed that RFLD is more computational intensive than FLD because a new pair of scatter matrices has to be generated for each new feature. In order to reduce the computation load, a new iterative method of computing the scatter matrices can be derived such that the new scatter matrix can be directly calculated from the old scatter matrix as follows.

$$S_W^{(k)} = S_W^{(k-1)} - w_{k-1}w_{k-1}^T S_W^{(k-1)} - S_W^{(k-1)} w_{k-1}w_{k-1}^T, + (w_{k-1}^T S_W^{(k-1)} w_{k-1}) w_{k-1}w_{k-1}^T \quad (24)$$

and

$$S_B^{(k)} = S_B^{(k-1)} - w_{k-1}w_{k-1}^T S_B^{(k-1)} - S_B^{(k-1)} w_{k-1}w_{k-1}^T. + (w_{k-1}^T S_B^{(k-1)} w_{k-1}) w_{k-1}w_{k-1}^T \quad (25)$$

The detail derivation of (24) and (25) is omitted due to space limitation.

Comment 3: After RFLD was developed, it was recognized that the extracted features are mathematically equivalent to those obtained by Orthonormal FLD [11]. However, the motivation and interpretation for Orthonormal FLD as well as the calculation process are significantly different from RFLD. There is always a pleasure in recognizing old things from a new point of view. Also, there are problems for which the new point of view offers a distinct advantage. For instance, the fundamental idea underlying RFLD of recursively deriving new features by discarding all the information represented by the old features can be readily applied to other techniques such as PCA and SVM [10]. Furthermore, only one synthetic example was considered in [11] to demonstrate its efficiency over FLD, while in this paper RFLD has been applied to a real-world application problem, i.e., the face recognition problem, which will be discussed in the following section.

3. EXPERIMENTS ON FACE RECOGNITION PROBLEM

Extensive experiments have been carried out to test the effectiveness of the suggested RFLD against PCA and FLD. In all the experiments, RFLD outperforms both PCA and FLD significantly. Due to space limitation, only part of our experimental results will be reported in detail in this section.

For identity recognition, both Yale database and ORL (Olivetti Research Laboratory) database were explored while for facial expression recognition problems, only Yale database was utilized since its images are labeled clearly with different expressions.

Yale database consists of 15 persons' frontal face images. We cropped those images by eliminating most of the background and some part of hair and chin. These gray level images were further preprocessed by 5-level, 8-direction, 64 down sampling Gabor wavelet with 12160 components as suggested in [9]

ORL database contains 40 persons' face images, which were also cropped and preprocessed by wavelet expansion with 3960 coefficients.

Four different approaches were tested and compared. While the same nearest neighbor rule with Euclidian distance is applied for all of them, they differ in the feature extraction processes. The first one is called the direct N-N method, in which no feature extraction is processed at all, and the nearest neighbor rule is directly applied to the high-dimensional images. In the second approach, PCA is used to reduce the high-dimensional images into lower-dimensional ones, but no discriminant analysis is performed. For the other two approaches, PCA is applied first to reduce the dimension of the images such that the within-class scatter matrix S_W is nonsingular, then either FLD or RFLD is used to extract features for classification.

Three recognition experiments will be discussed in the following part. They are identity recognition with Yale database and ORL database, and six-expression recognition with Yale database. The lowest recognition error rates achieved by the four classifiers are tabulated in Table 1, while the comparative performances of PCA, FLD and RFLD with varying number of features are plotted in Figure 1.

3.1. Identity Recognition

The identity recognition error rate is determined by "leaving-one-out" strategy [2,3]: to classify one particular image, all the rest of the images are pooled to form the training data set which are used to compute the projecting directions by PCA, FLD and RFLD.

The total number of features available by FLD is limited to 14 (Yale database) and 39 (ORL database), since there are 15 (Yale) and 40 (ORL) different persons to recognize, while no such limitation exists for either PCA or RFLD. From Table 1, it is evident that the classifier using RFLD achieves the lowest error rate and the improvement over all other approaches is substantial. In particular, perfect recognition has been achieved for Yale Database.

3.2. Facial Expression Recognition

It is expected that RFLD would have greater advantage over FLD when the number of classes is small, which is the main motivation behind the experiments on facial expression recognition problem. There are 11 images for each person in Yale database. They are labeled by facial expressions, lighting conditions or whether wearing glasses or not: *"normal", "happy", "sad", "sleepy", "surprise", "wink", "left light", " center light", "right light", " without glasses"* and *" with glasses"*. Images not labeled by expressions were treated as *"normal"* in our experiment.

In this experiment, the recognition error rates are determined by cross validation [2,3] rather than "leaving one out", i.e., all the images belonging to one particular person will be used as test images while the rest of the images are all included in the training data set.

Again the performance of RFLD obviously shines among others. It is interesting to note that the overall performances here are much worse than those for identity recognition, which, as shown in Figure 1 and Table 1, implies that the expression recognition problem is a much tougher one. This is not a surprise at all considering the fact that it is much harder for human beings to recognize different facial expressions rather than different persons.

4. CONCLUSION

This paper deals with the important problem of extracting discriminant features for pattern classification. A novel recursive algorithm (termed RFLD) incorporating the basic idea of classical FLD was suggested to eliminate the constraint on the

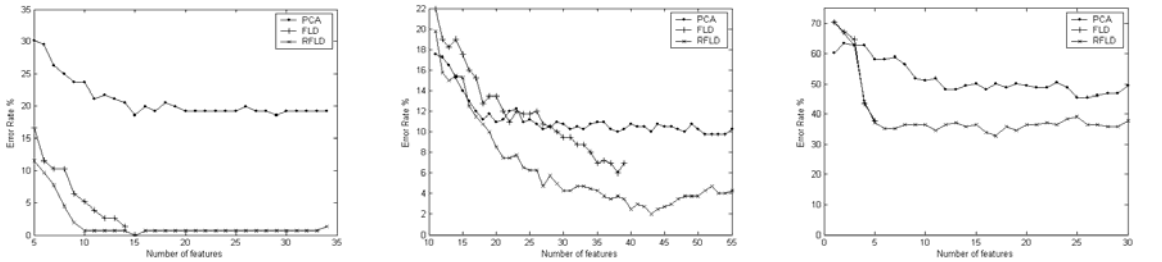
total numbr of features available, which is the main contribution of this paper.

The proposed RFLD is theoretically more appealing than the classical FLD based upon following two considerations. First of all, the total number of available features from RFLD is independent of the number of classes while that of FLD is limited to $c-1$, which is the major disadvantage of FLD over other popular approaches such as PCA. Further, the mathematical interpretations of the features from RFLD are more convincing for using them as features for classifications. While the k-th feature extracted from RFLD can be interpreted as the k-th best direction for separation by the nature of its optimization process involved, the k-th feature (except the first one) from FLD is merely an eigenvector associated with certain matrix.

It is evident from experiment results on face recognition problem that RFLD is indeed superior to other traditional approaches. We are well aware that lots of other factors may affect the performance, and there is still large room to improve the accuracy. For instance, other distance measures instead of Euclidian distance should be experimented with and even the choice of other classifiers such as neural networks may be also worth investigating. Work is currently under progress to study the various design issues for face recognition, and the objective is to achieve 99% accuracy rate [12] for identity recognition for all the widely used databases, and at least 80% accuracy for facial expression recognition for Yale database.

Table 1 Comparative Recognition Error Rates

Method	ID recognition (Yale database)		ID recognition (ORL database)		Six-expression recognition (Yale database)	
	Error rate (%)	Number of features	Error rate (%)	Number of features	Error rate (%)	Number of features
Direct N-N	17.31	12160	6.50	3960	51.28	12160
PCA	18.59	15	8.75	63	45.51	25
FLD	1.28	14	6.00	38	37.82	5
RFLD	0	15	2.00	43	32.69	17



Identity recognition (Yale database) Identity recognition (ORL database) Six-expression recognition (Yale database)

Fig. 1. Comparative recognition performances of PCA, FLD and RFLD

ACKNOWLEDGEMENT

The research reported here was supported by NUS Academic Research Fund R-263-000-224-112.

REFERENCE

[1] R. A. Fisher, "The use of multiple measures in taxonomic problems," *Ann. Eugenics, vol.7,* pp. 179-188, 1936.

[2] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*. Wiley, New York, second edition, 2000.

[3] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs.Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol 19, pp. 711-720, July 1997.

[4] K. Etemad and R. Chellappa, "Discriminant analysis for recognition of human face images," *J. Opt. Soc. Amer. A,* vol. 14, pp. 1724-1733, 1997.

[5] D. L. Swets and J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 18, pp. 831-836, Aug. 1996.

[6] Y. Cheng, K. Liu, J. Yang, Y. Zhuang, and N. Gu, "Human face recognition method based on the statistical model of small sample size," *SPIE Proc. Intelligent Robots and Computer Vision X: Algorithms and Technology,* pp. 85-95, 1991.

[7] Y. Cui, D. Swets, and J. Weng, "Learning-based hand sign recognition using SHOSLIF-M," *Int'l Conf. on Computer Vision*, pp. 631-636, 1995.

[8] S. Baker and S.K. Nayar, "Pattern rejection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 544-549, 1996.

[9] C. Liu and H. Wechsler, "Gabor feature based classification using the Enhanced Fisher Discriminant Model for face recognition," *IEEE Trans. Image Processing*, vol. 11, pp. 467-476, Apr. 2002.

[10] B. Heisele, T. Poggio and M. Pontil , "Face detecton in still gray images", *publications.ai.mit.edu,ai-publications/1500-1999/AIM-1687.ps.Z*

[11] T. Okada and S Tomita, "An optimal orthonormal system for discriminant analysis", *Pattern Recognition,* Vol.18, pp. 139-144, 1985.

[12] J. Daugman, "Face and gesture recognition: Overview," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, no. 7, pp. 675-676, 1997.