Face Recognition Using Recursive Fisher Linear Discriminant

C. Xiang, *Member, IEEE*, X. A. Fan, and T. H. Lee, *Member, IEEE*

***Abstract*—Fisher linear discriminant (FLD) has recently emerged as a more efficient approach for extracting features for many pattern classification problems as compared to traditional principal component analysis. However, the constraint on the total number of features available from FLD has seriously limited its application to a large class of problems. In order to overcome this disadvantage, a recursive procedure of calculating the discriminant features is suggested in this paper. The new algorithm incorporates the same fundamental idea behind FLD of seeking the projection that best separates the data corresponding to different classes, while in contrast to FLD the number of features that may be derived is independent of the number of the classes to be recognized. Extensive experiments of comparing the new algorithm with the traditional approaches have been carried out on face recognition problem with the Yale database, in which the resulting improvement of the performances by the new feature extraction scheme is significant.**

***Index Terms*—Face recognition, feature extraction, Fisher Linear Discriminant (FLD), principal component analysis (PCA), recursive Fisher linear discriminant (RFLD).**

I. Introduction

EXTRACTING proper features is crucial for satisfactory design of any pattern classifier, and how to develop a general procedure for effective feature extraction remains an interesting and challenging problem. 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. Around the year of 1997, comparative studies between Fisher linear discriminant (FLD) and PCA on the face recognition problem were reported independently by numerous authors [1]–[3], in which FLD outperformed PCA significantly. These successful applications of FLD have drawn a lot of attention on this subject and the ensuing years have witnessed a burst of research activities on various issues relating to applying subspace methods such as PCA and FLD to pattern recognition problems [4]–[8], with the latest development being an attempt to unify all theses subspace methods under the same framework [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. This cap on the total number of features is rooted in the mathematical treatment of FLD, which may attribute to the fact that although this constraint is well known, it has somehow been accepted as an inherent characteristic of FLD and received little attention in the literature. If the number of classes is large as is the case for identity recognition problems considered in most of the papers, this limitation may not arise as a visible obstacle. However, it may pose as a bottleneck if the number of classes is small. For instance, for the glasses-wearing recognition problem treated in [1], the number of classes is two, and hence the number of features resulting from FLD is only one. Although it was demonstrated there that even one FLD feature could outperform PCA for this particular case, it may not be the case for most of the other two-class classification problems since it is too naive to believe that only one FLD feature would suffice for all. 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, recursive Fisher linear discriminant (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 as well as a number of their variations.

After RFLD was developed, it was recognized that the extracted features are mathematically equivalent to those obtained by Orthonormal FLD [10], [11]. However, the motivation and interpretation for Orthonormal FLD as well as the calculation process are very 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 [12]. Furthermore, only one synthetic example and IRIS database were considered in [10] and [11] respectively 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.

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

Manuscript received May 20, 2004; revised August 2, 2005. This work was supported by the National University of Singapore Academic Research Fund R-263-000-224-112. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gopal Pingali.

C. Xiang and T. H. Lee are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117576 (e-mail: elexc@nus.edu.sg; eleleeth@nus.edu.sg).

X. A. Fan is with Huawei Technologies, Beijing, China (e-mail: xiaoan.fan@gmail.com).

Digital Object Identifier 10.1109/TIP.2006.875225

II. FLD Versus RFLD

A. Fisher Linear Discriminant (FLD)

Suppose that we have a set of n d -dimensional samples $x_1, \ldots, x_n$ belonging to c different classes with n_i samples in the subset D_i labeled $\omega_i, i = 1, \ldots, c$. Then the objective of FLD 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 maximized, 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 d-dimensional sample mean for 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

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

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.

B. 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 preprocessed 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 preprocessed database will be formulated in the same fashion as that of FLD.

Let us consider a set of n d-dimensional samples $x_1, \ldots, x_n$ belonging to c classes, as discussed in Section II-A. 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 is 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 in [1]. From now on, we will always assume that S_W is nonsingular.

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:

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

Based on this new set of sampled vectors $x_1^{(2)}, \ldots, x_n^{(2)}$, 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 II-A

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

and

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

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

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

and

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

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

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

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

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

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

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

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. It is also obvious that w_1 satisfies equation (16) since $S_B^{(2)} w_1 = S_W^{(2)} w_1 = 0$, which would make the ratio, $J^{(2)}(w)$, indefinite. In order to prevent such a situation from occurring, additional constraint has to be imposed on this optimization problem. Considering the fact that the information represented by previous feature w_1 is supposed to be discarded from the samples, it is natural to impose the following condition that the new feature w_2 is orthogonal to w_1, i.e.

$$w_1^T w_2 = 0. \quad (17)$$

Combining (16) and (17) results in

$$B_2 w_2 = \lambda W_2 w_2 \quad (18)$$

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

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

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

$$\left(W_2^T W_2\right)^{-1} W_2^T B_2 w_2 = \lambda w_2 \quad (20)$$

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

Similarly, it can be readily shown that the kth feature vector w_k may be computed as the normalized eigenvector with the largest eigenvalue from following eigenvalue problem:

$$\left(W_k^T W_k\right)^{-1} W_k^T B_k w_k = \lambda w_k \quad (21)$$

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

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

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

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

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 further maximized by projection. It is obvious from the 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 that maximizes 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) [12], which is to generate new sample sets by projecting the samples into a subspace that is orthogonal to selected features. However, different from support vector machine applied in [12], the similar recursive procedure is applied to FLD in this paper.

It is observed that RFLD is more computationally intensive than FLD because a new pair of scatter matrices have to be generated for each new feature. In particular, if the number of samples is large, the process of recalculating the new set of sample vectors at each step as described by (23) may be laborious. In order to reduce the computation load, a new method of computing the scatter matrices is introduced as follows such that the new scatter matrices can be directly calculated from the old scatter matrices rather than relying on the preprocessed sampled vectors.

To simplify the notation, we will write the within-class scatter matrix S_W in the form of

$$S_W = \sum_{i=1}^{n} z_i z_i^T \quad (24)$$

where $z_i \in R^d$, which refers to $x_i - m$, where m is the mean of the class to which x_i belongs to.

Let the $d \times (k-1)$ matrix

$$W^{(k-1)} = [w_1, \ldots, w_{k-1}] \quad (25)$$

where $w_1, \ldots, w_{k-1}$ are the feature vectors obtained from previous $k-1$ steps, and denote the $d \times (d-k+1)$ matrix

$$W_N^{(k-1)} = [v_k, v_{k+1}, \ldots, v_d] \tag{26}$$

where $v_k, v_{k+1}, \ldots, v_d$ is a set of orthonormal basis for the null space of the space spanned by $w_1, \ldots, w_{k-1}$, such that the set of vectors $\{w_1, \ldots, w_{k-1}, v_k, v_{k+1}, \ldots, v_d\}$ constitutes an orthonormal basis for R^d. Then the new set of vectors at the k th step, $z_i^{(k)}$, are formed by discarding the projections along all the selected feature directions

$$z_i^{(k)} = z_i - \sum_{j=1}^{k-1} \left(w_j^T z_i\right) w_j, \quad i = 1, 2, \ldots\ldots, n. \tag{27}$$

Using the fact that $z_i = \sum_{j=1}^{k-1} (w_j^T z_i) w_j + \sum_{j=k}^{d} (v_j^T z_i) v_j$, it follows that

$$z_i^{(k)} = \sum_{j=k}^{d} \left(v_j^T z_i\right) v_j, \quad i = 1, 2, \ldots\ldots, n. \tag{28}$$

Using definition of (26), (28) can be rewritten as

$$z_i^{(k)} = W_N^{(k-1)} \left[\left(W_N^{(k-1)}\right)^T z_i\right], \quad i = 1, 2, \ldots, n. \tag{29}$$

It follows immediately that the within-class scatter matrix at kth step, $S_W^{(k)}$, can be computed from

$$\begin{aligned} S_W^{(k)} &= \sum_{i=1}^{n} z_i^{(k)} z_i^{(k)T} \\ &= \sum_{i=1}^{n} \left(W_N^{(k-1)} \left(W_N^{(k-1)}\right)^T z_i\right) \\ &\quad \times \left[W_N^{(k-1)} \left(W_N^{(k-1)}\right)^T z_i\right]^T \end{aligned} \tag{30}$$

which may be rewritten as

$$S_W^{(k)} = W_N^{(k-1)} \left(W_N^{(k-1)}\right)^T \left(\sum_{i=1}^{n} z_i z_i^T\right) W_N^{(k-1)} \times \left(W_N^{(k-1)}\right)^T. \tag{31}$$

Substituting (24) into (31) yields

$$S_W^{(k)} = W_N^{(k-1)} \left(W_N^{(k-1)}\right)^T S_W W_N^{(k-1)} \left(W_N^{(k-1)}\right)^T. \tag{32}$$

Similarly, the between-class scatter matrix at the kth step, $S_B^{(k)}$, can be calculated by

$$S_B^{(k)} = W_N^{(k-1)} \left(W_N^{(k-1)}\right)^T S_B W_N^{(k-1)} \left(W_N^{(k-1)}\right)^T. \tag{33}$$

Now, we are ready to summarize the RFLD algorithm as follows.

Initialize $S_B^{(1)}$ and $S_W^{(1)}$ with the original scatter matrices S_B, and S_W respectively, and derive the first feature direction w_1 as the normalized eigenvector with the largest eigenvalue of the matrix $S_W^{-1} S_B$, which is the same as that from the classical FLD.

The kth feature vector w_k may be computed as the normalized eigenvector with the largest eigenvalue from following eigenvalue problem:

$$\left(W_k^T W_k\right)^{-1} W_k^T B_k w_k = \lambda w_k, \quad k > 1 \tag{34}$$

where the $(d+k-1) \times d$ matrices W_k and B_k are defined by (22), in which the within-class scatter matrix, $S_W^{(k)}$, and the between-class scatter matrix, $S_B^{(k)}$, are updated by (32) and (33), respectively.

III. Experiments on Face Recognition Problems

Extensive experiments have been carried out to test the effectiveness of the suggested RFLD against other well known methods. Due to space limitation, only part of our experimental results will be reported in detail in this section.

A. *Yale Database*

The Yale database was utilized in our experiments for identity recognition, facial expression recognition and glasses-wearing recognition problems. The Yale database consists of 15 persons' frontal face images, with 11 images for each person. However, it was realized during our experiments that there are some duplicate images, which were then discarded from all the experiments. We cropped those images by eliminating most of the background and some part of hair and chin. The size of images were changed from 320 × 243 to 124 × 147.

B. *Image Coding Methods*

The original cropped gray-level images are of 18228 dimensions. To improve the performance, two types of 2-D wavelet transform were preprocessed separately. They are five-level, eight-direction, 64 downsampling Gabor wavelet with 12160 components as suggested in [8] and 5-layer Bi-orthogonal 1.1 wavelet expansion with 18437 coefficients, which are available in MATLAB toolboxes. Again, due to space limitation, only the recognition results for gray-level and Gabor wavelet representations will be presented in this paper, as the results from bi-orthogonal wavelet expansion are very similar and hence omitted.

C. *Classification Methods for Comparison*

Seven different classification approaches have been 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.

Comment 3: The Mahalanobis distance was also experimented with to improve the recognition accuracy as recommended in [8]. However, the comparison study between the

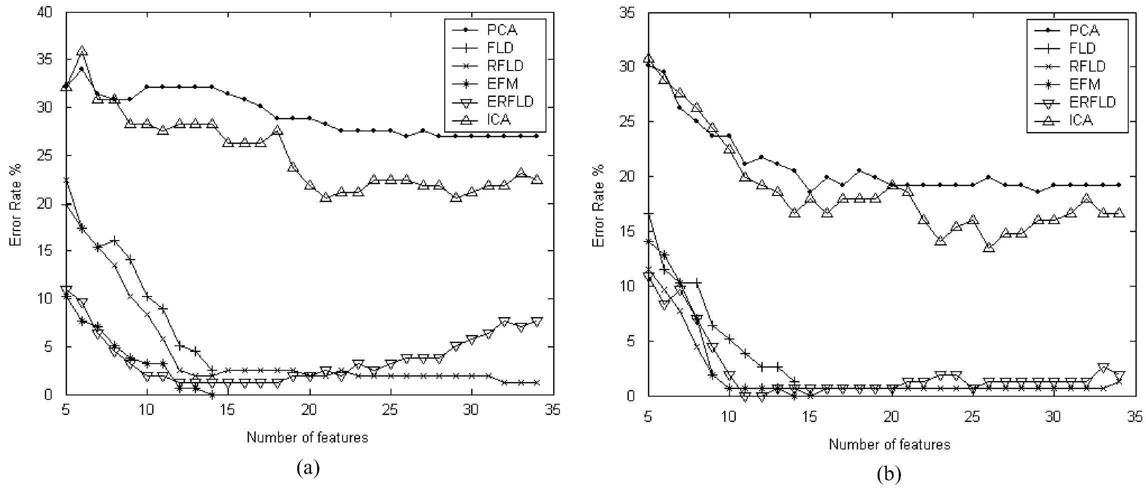


Fig. 1. Comparative identity recognition performances using different methods. (a) Gray level. (b) Gabor wavelet.

Mahalanobis and Euclidean distances is inconclusive from our experiments, and it is decided to report the results by Euclidean distance only due to space limitation.

The first method is the simplest one, called the direct N-N method, in which no feature extraction is processed at all, and the nearest neighbor rule is applied directly to the original 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 afterwards.

The third method utilized is the independent component analysis (ICA), which is known as a statistical method that linearly transforms a multidimensional vector data into components that are as statistically independent from each other as possible. The reader is referred to [13]–[15] for further details on ICA.

The fourth and fifth approaches are FLD and RFLD, for which PCA is applied first to reduce the dimension of the images such that the within-class scatter matrix, S_W, is nonsingular. The question of how to decide which PCA components to retain and which to discard is quite intriguing. For the purpose of maintaining as much information as possible, we choose the maximum number of eigenvectors with largest eigenvalues from PCA that ensure the full rank of S_W, which is at most $n - c$, where n is the number of training samples and c is the number of classes.

Another approach for selecting proper number of PCA components for FLD is the enhanced FLD Model (EFM) proposed in [8], which is also tested in our experiments. EFM aims to seek a proper number of PCA features that balance between the need to keep enough spectral energy of raw data and the requirement that the eigenvalues of within-class scatter in the reduced PCA space are not too small, for the tiny eigenvalues are associated with noise that make FLD over-fitting while exposed to new data. Unfortunately, no quantitative criterion for measuring the adequacy of energy and the smallness of eigenvalues of within-class scatter is currently available and hence the cut-off point for the number of PCA components to retain has to be obtained through trial and error. In our experiments, the optimal number of PCA features is the one leads to the lowest error rate, and is found through simple exhaustive search rather than analyzing the spectrum of the eigenvalues as suggested in [8].

Inspired by the basic idea of EFM, the enhanced RFLD (ERFLD) is also proposed and compared in this paper, in which the dimension of the reduced PCA space is varied such that an optimal number of PCA components are searched out, which leads to the best performance by RFLD.

D. Identity Recognition

The identity recognition error rate is determined by "leaving-one-out" strategy: 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, ICA, FLD, RFLD, EFM, and ERFLD.

The lowest recognition error rates achieved by the seven classifiers and two representation methods are listed in Table I. The recognition error rates depend upon the number of features used for each approach as shown in Fig. 1. The total number of features available by FLD and EFM is limited to 14 since there are 15 different persons to recognize, while no such limitation exists for PCA, ICA, RFLD, and ERFLD.

It is evident from Table I that the classifiers using RFLD, EFM, and ERFLD with Gabor wavelet coding and EFM with gray-level representation have achieved perfect recognition with zero error rates. The classifier using RFLD achieves lower error rate than those by PCA, FLD, and ICA, and the improvement is substantial. The performance of FLD is much better than that of PCA, which is the same conclusion as drawn in [1]. ICA is better than PCA, which is consistent with the result in [14], but it is worse than FLD, RFLD, EFM, and ERFLD, which is reasonable since the objective of ICA is to make the components of projected vectors as independent as possible, which may not necessarily be the best for classification. The improvement by EFM over FLD is significant, which was also observed in [8]. As far as the coding method is concerned, the Gabor wavelet does show distinct advantage for the identity recognition problem, which is also consistent with similar conclusions in [8]. Since EFM

TABLE I
COMPARATIVE LEAST ERROR RATES FOR IDENTITY RECOGNITION

Methods	Gray-level representation		Gabor wavelet representation	
	Error rate %	Number of features	Error rate %	Number of features
N-N	25.64	18228	17.31	12160
PCA	26.28	35	18.59	15
ICA	19.23	36	13.46	26
FLD	2.56	14	1.28	14
RFLD	1.28	32	0	15
EFM	0	14	0	14
ERFLD	1.28	12	0	11

TABLE II
COMPARATIVE LEAST ERROR RATES FOR SIX-EXPRESSION RECOGNITION

Methods	Gray-level representation		Gabor wavelet representation	
	Error rate %	Number of features	Error rate %	Number of features
N-N	47.44	18228	51.28	12160
PCA	44.23	17	45.51	25
ICA	39.10	13	46.79	25
FLD	37.82	5	37.82	5
RFLD	29.49	13	32.69	17
EFM	32.05	5	35.26	5
ERFLD	28.85	10	28.85	26

has already achieved perfect performance, there appears no advantage gained by RFLD as well as ERFLD, which may be attributed to the fact that the number of classes for identity recognition is not small and hence the number of features from EFM is quite sufficient. Therefore, it would be interesting to apply all these methods to recognition problems with small number of classes, which motivates us to carry further experiments on facial expression recognition and glasses-wearing recognition problems as will be discussed later in this section.

Comment 4: While the number of features from RFLD and ERFLD is free of the constraint of $c - 1$ (14 in this case), it is not true that the performances of RFLD and ERFLD always increase monotonically with the number of features as shown in Fig. 1. Usually the performance would improve with the number of features increasing, and subsequently deteriorate as more features are used, sometimes even fluctuating. The optimal value of number of features is problem dependent in general.

Comment 5: The optimal numbers of PCA components retained by EFM and ERFLD are usually different. For example, for Gabor wavelet coded images, EFM accomplishes zero error rate with 50 PCA features while ERFLD achieves the same goal with 100 PCA features. On the other hand, if 100 PCA features are retained for FLD, the lowest error rate is 1.28%, and if 50 PCA components are kept for RFLD, the least error rate is 0.64%, both of which would not be perfect. Such difference also exists in other experiments with different representation methods and different recognition problems. The possible reason may be explained as follows. Both noise and useful information exist together at the directions that are the eigenvectors of total scatter associated with small eigenvalues. However, the responses of EFM and ERFLD to them are different, so the best performances emerge at different choices of the number of PCA vectors. Such phenomenon also indicates that it would be misleading to select the optimal number of PCA features by merely studying the two sets of eigenvalues (corresponding to the total scatter of raw data and the within-class scatter in reduced PCA space) [8] without trial and error experiments.

E. Six-Expression Recognition

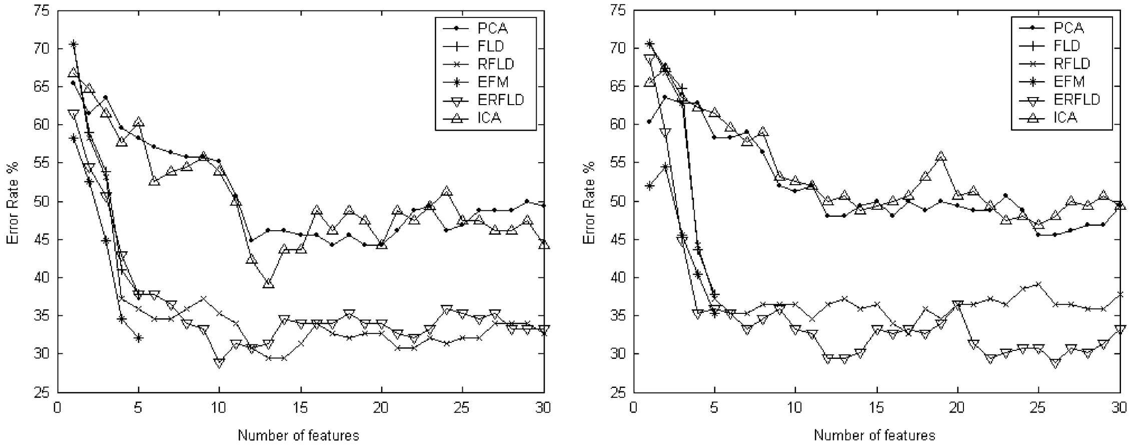
It is expected that RFLD and ERFLD would have greater advantage over FLD and EFM when the number of classes is small, which is the main motivation behind the experiments on facial expression recognition problem and glasses-wearing recognition problem. As previously mentioned, there are 11 images for 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.”* For those images not labeled by expression, their expressions are usually “*normal*.”

In this experiment, the recognition error rates are determined by cross validation 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.

Comment 6: The “leaving one out” strategy was also experimented with, in which the performances of PCA, ICA and direct N-N were much worse, because in this case the nearest neighbor determined by PCA, ICA, or direct N-N is usually the image belonging to the same person which might be labeled with a different expression. In some sense, it is not “fair” for PCA and ICA, and hence cross validation, as described above was adopted.

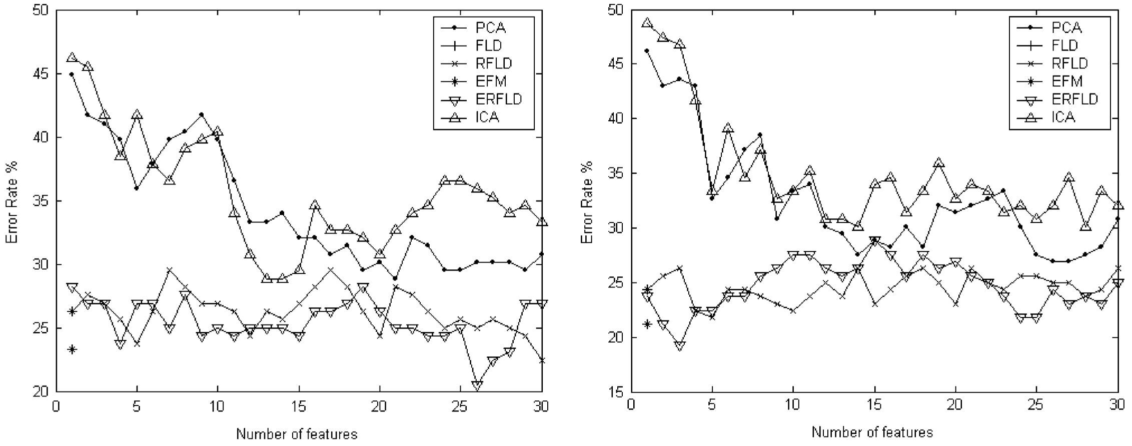
The lowest recognition rates of the seven classification methods corresponding to two representation methods are shown in Table II, while the comparative performances of them with varying number of features are plotted in Fig. 2. It is interesting to note that overall performances are much worse than those for identity recognition, which 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.

As expected, the improvements of RFLD over FLD, and ERFLD over EFM, are quite substantial for this fewer-classes problem as shown in Table II, because RFLD and ERFLD eliminate the constraint on maximal number of features and hence obtains more information than FLD and EFM for classification. The number of features from FLD and EFM is limited to five since there are only six types of expressions to recognize while



(a) (b)

Fig. 2. Comparative six-expression recognition performances using different methods. (a) Gray level. (b) Gabor wavelet.



(a) (b)

Fig. 3. Comparative two-expression recognition performances using different methods. (a) Gray level. (b) Gabor wavelet.

no such constraint exists for RFLD and ERFLD. ERFLD provides the least error rate of 28.85% throughout all the results, while the smallest error rate by EFM is only 32.05%.

Comment 7: It is interesting to note that for the six-expression recognition problem, gray-level coding appears better than Gabor-wavelet representation in terms of the lowest error rates as well as the number of features necessary for good performances. Therefore, it should not be taken for granted that Gabor-wavelet coding is always better than gray-level representation for classification.

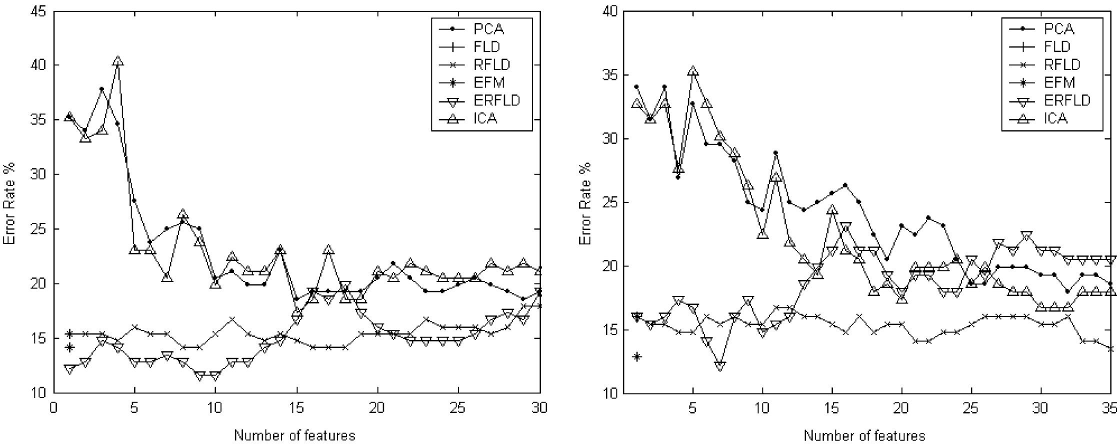
F. Two-Expression Recognition

Now we consider another problem, in which images are classified into only two classes: *"normal"* or *"abnormal"*. All those images labeled by *"happy," "sad," "sleepy," "surprise,"* and *"wink"* are treated as *"abnormal."* For two-class problem, only one feature may be obtained by FLD and EFM, while RFLD and ERFLD are totally free from this constraint.

As shown in Table III and Fig. 3, the performances for this two-expression recognition problem are much better than those of the six-class recognition which is consistent with human experience in some sense since it is easier for human beings to tell whether a person's expression is normal or not rather than to tell the exact type of expression. The comparison results for the seven approaches are also consistent with the previous observations. Again, ERFLD achieves the best result of 19.23% with Gabor wavelet coding.

TABLE III
COMPARATIVE LEAST ERROR RATES FOR TWO-EXPRESSION RECOGNITION

Methods	Gray-level representation		Gabor wavelet representation	
	Error rate %	Number of features	Error rate %	Number of features
N-N	32.69	18228	33.97	12160
PCA	27.56	35	26.92	26
ICA	28.85	13	30.13	14
FLD	26.28	1	24.36	1
RFLD	22.44	30	21.79	5
EFM	23.72	1	21.15	1
ERFLD	20.51	26	19.23	3



(a) (b)

Fig. 4. Comparative glasses-wearing recognition performances using different methods. (a) Gray level. (b) Gabor wavelet.

TABLE IV
COMPARATIVE LEAST ERROR RATES FOR GLASSES-WEARING RECOGNITION

Methods	Gray-level representation		Gabor wavelet representation	
	Error rate %	Number of features	Error rate %	Number of features
N-N	20.51	18228	19.78	12160
PCA	18.59	15	17.95	32
ICA	17.31	15	16.67	30
FLD	15.38	1	16.03	1
RFLD	14.10	8	13.46	35
EFM	14.10	1	12.82	1
ERFLD	11.54	9	12.18	7

G. Glasses-Wearing Recognition

The glasses-wearing recognition problem is another two-class case, where cross validation was used to obtain the recognition error rates. The results are shown in Table IV and Fig. 4. The overall error rate is smaller than that of the facial expression recognition problems, which is also quite natural from human experience. However, while human beings are able to tell whether a person wearing glasses or not at a casual glance, it is not the same case for computer, in our experiments, the error rates were still above 10%.

ERFLD still shines among others with the least error rate of 11.54% for gray-level coded images. FLD, RFLD, EFM, and ERFLD remain better than N-N, PCA and ICA, while RFLD and ERFLD outperform FLD and EFM, respectively.

Comment 8: In this case, it is not easy to distinguish which coding method is better for recognition. Although Gabor wavelet representation appears better for NN, PCA, ICA, RFLD and EFM, it is worse for FLD and ERFLD with the lowest error of 11.54% being achieved by ERFLD with gray-level images. Therefore, it appears that choosing appropriate coding methods for recognition is quite problem dependent and is hard to decide without trial and error experiments.

IV. 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 is suggested, 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 and ICA. By eliminating this bottleneck, RFLD provides the basis for applying the fundamental idea underlying the FLD of seeking features that maximize the separation of different classes to a more general class of pattern classification problems. 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 kth best direction for separation by the nature of its optimization process involved, the kth feature (except the first one) from FLD is merely an eigenvector associated with certain matrix.

It is certainly true that RFLD is more computational intensive than FLD. However, all the computational cost associated with RFLD only occurs in the classifier design process. Once the features are extracted and ready to be used by the final classifier, there will be no extra cost. Furthermore, the superior performance of RFLD as demonstrated by various experiments on face recognition problem is sufficient to convince that it is worth putting up with the computation overhead.

It has been observed that the issue of retaining proper number of PCA components for further discriminant analysis plays an important role in improving the recognition accuracy, which confirms the same conclusion drawn in [8]. Exhaustive search was utilized in this paper to find the optimal number of PCA features, which was far from efficient. Further studies are needed to address this issue.

For identity recognition problems, EFM is capable of achieving perfect recognition, which suggests that for pattern classification problem with large number of classes, EFM may be used first due to its computational efficiency, and ERFLD may be employed only if EFM fails to achieve the design goal. For problems with small number of classes such as the expression and glasses-wearing recognition problems, ERFLD does show distinct advantage and should be utilized for extracting features.

Since perfect recognition has been achieved for both the Yale database in this paper, and the FERET database reported in [8], it is expected that similar performance may be also accomplished for other widely used databases. However, the lowest error rates are 28.85% and 19.23% for six-expression and two-expression recognitions respectively, which are far from satisfactory compared to average recognition accuracy that may be realized by human beings. It is expected that other techniques are needed to further improve the performance of facial expression classifier, and work is currently under progress to try to achieve the error-rate of less than 20% for six-expression recognition, which is a rough estimation of the average expression recognition error rate made by human beings.

Acknowledgment

The authors would like to thank the associate editor and the anonymous reviewers for their invaluable comments to improve the quality of this paper.

References

[1] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

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

[3] D. L. Swets and J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 831–836, Aug. 1996.

[4] C. Liu and H. Wechsler, "Robust coding schemes for indexing and retrieval from large face databases," *IEEE Trans. Image Process.*, vol. 9, no. 1, pp. 132–137, Jan. 2000.

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

[6] A. M. Martínez and A. C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 228–233, Feb. 2001.

[7] H. Yu and J. Yang, "A direct LDA algorithm for high-dimensional data—With application to face recognition," *Pattern Recognit.*, vol. 34, pp. 2067–2070, 2001.

[8] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, Apr. 2002.

[9] X. Wang and X. Tang, "A unified uramework for subspace face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1222–1228, Sep. 2004.

[10] T. Okada and S. Tomita, "An optimal orthonormal system for discriminant analysis," *Pattern Recognit.*, vol. 18, pp. 139–144, 1985.

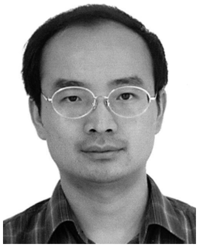
[11] J. Duchene and S. Leclercq, "An optimal transformation for discriminant and principal component analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 10, no. 6, pp. 978–983, Jun. 1988.

[12] B. Heisele, T. Poggio, and M. Pontil*, Face Detecton in Still Gray Images*, 2000 [Online]. Available: publications.ai.mit.edu,ai-publications/1500-1999/AIM-1687.ps.Z

[13] P. Comon, "Independent component analysis, a new concept?," *Signal Process.*, vol. 36, pp. 287–314, 1994.

[14] C. Liu and H. Wechsler, "Independent component analysis of Gabor features for face recognition," *IEEE Trans. Neural Netw.*, vol. 14, no. 4, pp. 919–928, Jul. 2003.

[15] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 626–634, May 1999.



C. Xiang (M'01) received the B.S. degree in mechanical engineering from Fudan University, China, in 1991, the M.S. degree in mechanical engineering from the Institute of Mechanics, Chinese Academy of Sciences, in 1994, and the M.S. and Ph.D. degrees in electrical engineering from Yale University, New Haven, CT, in 1995 and 2000, respectively.

From 2000 to 2001, he was a Financial Engineer at Fannie Mae, Washington, DC. At present, he is an Assistant Professor in the Department of Electrical and Computer Engineering, National University of Singapore. His research interests include computational intelligence, adaptive systems, and pattern recognition.

X. A. Fan received the B.Sc. degree in electronics from Peking University, Beijing, China, in 2002, and the M.Eng degree in electrical and computer engineering from the National University of Singapore in 2004.

She was previously with STMicroelectronics Asia Pacific. Currently, she is with Huawei Technologies, Beijing, China.

T. H. Lee (M'02) received the B.A. degree (with first-class honors) in engineering from Cambridge University, Cambridge, U.K., in 1980, and the Ph.D. degree from Yale University, New Haven, CT, in 1987.

He is a Professor in the Department of Electrical and Computer Engineering, National University of Singapore. His research interests are in the areas of adaptive systems, knowledge-based control, intelligent mechatronics, and computational intelligence. He has co-authored three research monographs, and holds four patents (two of which are in the technology area of adaptive systems, and the other two are in the area of intelligent mechatronics).

Dr. Lee was a recipient of the Cambridge University Charles Baker Prize in Engineering. He currently holds Associate Editor appointments in *Automatica*; the IEEE Transactions on Systems, Man, and Cybernetics; *Control Engineering Practice* (an IFAC journal); the *International Journal of Systems Science*; and *Mechatronics Journal*.