Classifiability-Based Optimal Discriminatory Projection Pursuit

Yu Su[1,2] Shiguang Shan[2,3] Xilin Chen[2,3] Wen Gao[1,4]
[1] School of Computer Science and Technology, Harbin Institute of Technology, China
[2] Digital Media Research Center, Institute of Computing Technology, CAS, China
[3] Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences, China
[4] Institute for Digital Media, Peking University, China
{ysu, sgshan, xlchen, wgao}@jdl.ac.cn

Abstract

Linear Discriminant Analysis (LDA) might be the most widely used linear feature extraction method in pattern recognition. Based on the analysis on the several limitations of traditional LDA, this paper makes an effort to propose a new computational paradigm named Optimal Discriminatory Projection Pursuit (ODPP), which is totally different from the traditional LDA and its variants. Only two simple steps are involved in the proposed ODPP: one is the construction of candidate projection set; the other is the optimal discriminatory projection pursuit. For the former step, candidate projections are generated as the difference vectors between nearest between-class boundary samples with redundancy well-controlled, while the latter is efficiently achieved by classifiability-based AdaBoost learning from the large candidate projection set. We show that the new "projection pursuit" paradigm not only does not suffer from the limitations of the traditional LDA but also inherits good generalizability from the boundary attribute of candidate projections. Extensive experimental comparisons with LDA and its variants on synthetic and real data sets show that the proposed method consistently has better performances.

1. Introduction

Extracting discriminatory features is crucial for the most pattern classification tasks, and how to develop algorithms for effective feature extraction remains an interesting and challenging problem. Among numerous feature extraction methods, Linear Discriminant Analysis (LDA) is probably one of the most well-known approaches. The basic idea of LDA is pursuing a low-dimensional subspace maximizing the between-class scatter while simultaneously minimizing the within-class scatter, which is generally achieved by maximizing the Fisher criterion [1].

In spite of its advantages, however, LDA has some severe problems both in theory and in practice. We discuss these problems in the following and review briefly the corresponding solutions in the literature.

Problem 1. The optimality criterion of LDA, which is essentially based on the *distance* of samples, is not directly related to the classification accuracy [2]. Especially, since the between-class scatter is defined as the sum of all the scatters between the means of any two classes, its maximization does not necessarily guarantee expected separation between any two classes in the output space. Therefore, if some classes overlap heavily in the output space, the Bayes error rate may be very high. To solve this problem, R.Lotilokar et al. proposed the so-called Fractional-step LDA (F-LDA) [2], in which the classes that are closer in the output space and thus likely resulting in more misclassifications are more heavily weighted in the computation of the between-class scatter. This idea is intuitively rational, but the weighting function can only help make the optimality criterion be more representative but yet not directly related to classifiability. Moreover, a good weighting function may be found only by experiments. In [3], M.Rohl et al. presented an approach to find a low-dimensional representation of data by minimizing the actual Bayes error in the reduced space. Although directly related to classification accuracy, it only works well under the assumption that all classes are normally distributed. And, it also needs a very time-consuming optimization.

Problem 2. LDA can only work well under the assumption that all the classes are of Gaussian distribution with the same covariance matrix and different means. Only in this case, LDA coincides with the optimal Bayes classifier. However, if the class distributions are non-Gaussian or share the same mean, LDA will fail to find the discriminatory directions. In [1], by defining nonparametric between-class scatter, Fukunaga proposed the Nonparametric Discriminant Analysis (NDA) to relax the assumption of Gaussian distribution in the two-class case. There are many extensions of NDA from the two-class to the multi-class case such as in [4] [5] [6]. Besides NDA, there are also many other methods aiming to deal with this problem, such as Subclass Discriminant Analysis (SDA) [7] and Heteroscedastic extension of LDA (HLDA) [8].

Problem 3. In practice, for high dimensional data, there are often not enough training samples to guarantee the non-singularity of the within-class scatter matrix (S_W). This problem is also known as the "small sample size" (3S) problem [9]. The traditional solution to this problem is the two-stage PCA+LDA method in which PCA is used for dimension reduction in order to remove the null space of S_W before the application of LDA [10]. However, it has been shown that the null space of S_W also contains a great deal of discriminatory information. Therefore, several null-space

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

based methods are proposed to solve the 3S problem, such as [11] [12] [13] [14]. In addition, in [15], the authors proposed Direct LDA (DLDA) to solve the 3S problem by discarding the null space of S_B and then minimizing the within-class scatter only in the range space of S_B.

Problem 4. The maximal number of the features available from LDA is limited to C-1, where C is the number of classes. This limitation may result in insufficient amount of features for accurate classification, especially when C is small (say C=2). Xiang et al. in [16] proposed to calculate the discriminatory features by a recursive procedure named Recursive Fisher Linear Discriminant (RFLD). In RFLD, the maximal number of features is independent of C. In addition, in NDA [1] and some null-space based methods such as [12], the maximal number of features is not limited to C-1.

In sum, one can notice that numerous variants of LDA have been proposed to solve its inherent problems. Each method, however, fails to cover all the problems. What is more important is that most of the existing methods exploit distance-based criterion, thus they can not solve or completely solve the *Problem 1*.

This paper has made an effort to solve all these problems by proposing a novel paradigm for linear feature extraction named Optimal Discriminatory Projection Pursuit (ODPP). Briefly speaking, totally unlike the traditional LDA, the proposed ODPP formulates linear feature extraction as the selection of the optimal linear projections from a large Candidate Projection Set (CPS) by using AdaBoost with classifiability-based criterion. In short, the contributions of this paper are as follows:

1) Totally unlike the traditional LDA, we formulate linear feature extraction as a "*projection pursuit*" procedure from a large CPS by using AdaBoost with *classifiability-based criterion*.
2) The large amount of candidate projections in the CPS are generated as the difference vectors between the *nearest between-class boundary samples*, which are also quite different from the class-mean-based projections as in the traditional LDA.
3) The above two points safely guarantee that ODPP naturally does not suffer from the above-mentioned four problems of the traditional LDA.
4) The proposed ODPP is evaluated on both synthetic and real data sets. Comparisons with LDA and its variants show the effectiveness of our method.

The remaining part of this paper is organized as follows: in section 2, we present the proposed projection pursuit framework. Then, in section 3, how to construct the candidate projection set is described, followed by the AdaBoost-based projection selection in Section 4. Experiments are presented in section 5. Conclusion and discussion are given in the last section.

2. Projection Pursuit: An Alternative Paradigm of Linear Feature Extraction

In traditional LDA, the within-class scatter matrix (S_W) and the between-class scatter matrix (S_B) are used to measure the class separability. They are defined as,

$$S_W = \sum_{i=1}^{C} \sum_{x \in X_i} (x - m_i)(x - m_i)^T$$

$$S_B = \sum_{i=1}^{C} N_i (m_i - m)(m_i - m)^T$$

where C is the number of classes; n_i is the number of samples of class X_i; m_i is the mean of class X_i; m is the mean of all the samples. The optimality criterion of LDA is then formulated as maximizing the ratio of the determinant of S_B to that of S_W,

$$W_{opt} = \arg\max_W \frac{\left|W^T S_B W\right|}{\left|W^T S_W W\right|}.$$

Mathematically, this ratio is maximized when the column vectors of the projection matrix W are the eigenvectors of $S_W^{-1} S_B$ [1]. The advantage of this distance-based criterion lies in its analytical solution by simple matrix arithmetic.

However, the above LDA suffers from several severe problems as mentioned in Section 1, and none of the existing extension methods can solve all these problems. Therefore, it is necessary to develop novel paradigms for better extracting linear discriminatory features. Aiming at this goal, we have made a primary effort by proposing a "*projection pursuit*" framework as an alternative paradigm for linear feature extraction. The basic idea of the method is illustrated in Fig.1.

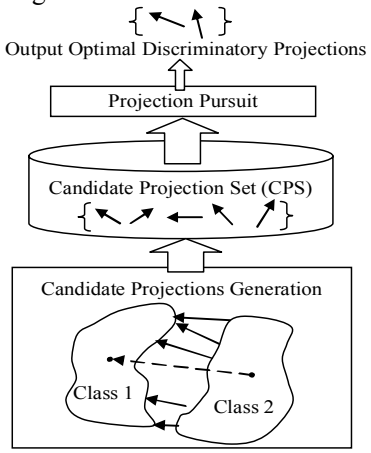

Fig.1. Proposed paradigm for linear feature extraction.

As shown in Fig.1, the proposed "*projection-pursuit*" framework consists of two main steps: 1) generating CPS based on the nearest between-class boundary samples; 2) pursuing a subset of discriminatory projections from the CPS. Though relatively simple, the proposed framework has the following merits:

1) By carefully designing the method of generating

candidate projections, any desirable discriminatory projection can be included in the CPS.

2) No assumption is required on the distributions of the classes, due to that the candidate projections are generated by a non-statistical manner.
3) This framework does not suffer from the "3S" and the "limited maximal number of feature" problems. This is evident since we do not need to solve any eigen-decomposition problem and the amount of possible projections in the CPS can be very large due to its combination nature.
4) The selection of the optimal discriminatory projections can be naturally classifiability-based rather than distance-based. Especially, AdaBoost provides a good choice for this task.

Actually, the above framework is quite general and allows various specific implementations for each step. In this paper, we just present one possibility for each step. For CPS construction, we propose to use as candidates the difference vectors between the nearest between-class boundary samples, as detailed in Section 3. While for projection pursuit, AdaBoost is exploited to select from CPS the most discriminatory projections one by one, as presented in Section 4.

3. Construction of CPS by Boundary Samples

Compared with LDA in which only the class means are taken into account for calculating the between-class scatter matrix, ODPP utilize more boundary information. This is done by generating the CPS based on the nearest between-class boundary samples. More specifically, for a pair of classes, the difference vectors of the nearest boundary samples are calculated and considered as candidate projections. The CPS contains all the candidate projections obtained from any pairs of classes.

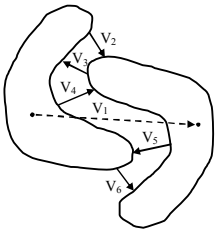

Fig.2. Illustration of different discriminabilities of projections generated by class means (shown by dash line) and the nearest between-class boundary samples (shown by solid line).

The reason of generating the CPS by boundary samples can be intuitively illustrated by Fig.2. Obviously, the difference vectors of boundary samples (shown by solid lines), can provide more discriminatory information than difference vector of class means (shown by dash line). The idea of using boundary samples also has some links with the SVM theory. The SVM algorithm tries to find the decision plane with maximum margin, and this decision plane can be completely determined by the Support Vectors (SVs). An interesting characteristic of the SVs is that they lie geometrically nearby the decision plane, and thus at the boundary of each class. So, the discriminatory information can be extracted effectively by using boundary samples.

If we merely enumerate all the possible difference vectors, one problem may arise: the candidate projections may have similar even the same directions. This may result in high redundancy in the CPS. Although the redundancy can be reduced by the following step of projection selection, large redundancy will make this process more time-consuming. Therefore, it is necessary to generated CPS elaborately to avoid large redundancy.

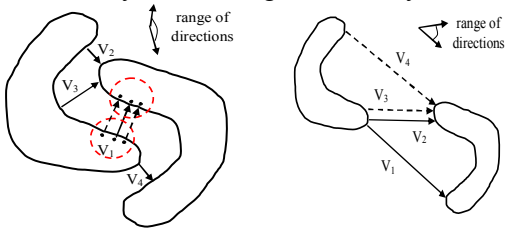

Fig.3. Illustration of the redundancy in candidate projections and its reduction. Only vectors represented by solid lines are chosen as candidate projections. See the text below for more detailed interpretation.

Fig.3 illustrates two cases where redundancy appears. In the first case, as shown in Fig.3a, a lot of candidate projections are generated by the samples located within small areas, thus have very similar directions. In order to reduce this redundancy, we propose to make the candidate projections distribute sparsely. Specifically, if a boundary sample has been used for generating a candidate projection, the samples within its neighborhood (the dotted circle in Fig.3a) will not be used any more in the following process. For example, as in Fig.3a, if V_1 has been chosen as a candidate projection, vectors represented by dotted line are no longer considered in the process of CPS construction. Fig.3b shows another case, in which the two classes have the similar distribution as in Fig.3a but are farther away from each other. In this case, the range of directions of all possible candidate projections is much smaller than the case in Fig.3a, thus making the generated candidate projections more redundant. Intuitively, the number of the candidate projections generated by the two classes in Fig.3b should be less than that in Fig.3a. Considering that, for the multi-class case, a weighting function is introduced into the generation process. Classes that are farther away from each other are associated with small weight. That means, in Fig.3, more candidate projections should be generated from the pair of classes in (a) than in (b).

With the above analysis in mind, we propose the following CPS construction algorithm, as shown in Fig.4.

CPS Construction

- Given N training samples $\{x_1, x_2, ..., x_N\}$ belonging to C classes, and the number of candidate projections generated by boundary samples B.
- Set $\mathrm{CPS} = \varnothing$.
- Calculate the weight of each pair of classes

$$W_{ij} = \frac{1}{\|m_i - m_j\|^2} \ (i, j = 1, 2, ..., C \text{ and } i < j),$$

where m_i is the mean of class i; and normalize W_{ij}:

$$W_{ij} = \frac{W_{ij}}{\sum_{i<j} W_{ij}} \ (i < j)$$

- For $i = 1, ..., C;\ j = 1, ..., C; i < j$
  1. Set $B_{ij} = W_{ij} \cdot B$
  2. Put all the difference vectors between class i and class j into a vector set S_{ij}.
  3. For $k = 1, ..., B_{ij}$
     a) Choose v_k with the smallest length from S_{ij}.
     b) Remove v_k and vectors generated by K nearest neighbors of v_k from S_{ij}.
     c) Normalize v_k to unit length and put it into the CPS.
- Put difference vectors between class means into the CPS.

Fig.4. The process of CPS construction.

4. Projection Selection by AdaBoost

Although each projection in the CPS is useful for discriminating two classes from which the projection is generated, they cannot guarantee high classification accuracy for all the classes. Moreover, evidently, candidate projections in the CPS still have redundancy. Therefore, in this paper, we further exploit AdaBoost as a method of feature selection to draw from the CPS a subset of projections with high classifiability and low redundancy. It is well-known that AdaBoost is a strong tool to solve the two-class classification problem, and have been extended to the multi-class problem by many methods such as AdaBoost.M1 and AdaBoost.M2 [17]. In the proposed ODPP, a variant of AdaBoost.M2 with classifiability-based criterion is used for projection selection. Fig.5 gives the detailed process of projection selection by AdaBoost.M2.

Compared with AdaBoost which is designed for two-class problem, AdaBoost.M2 requires more elaborate communication between the Boosting algorithm and the weak learning algorithm [17]. More specifically, in AdaBoost.M2, the weak hypothesis *h(x,y)* is required to measure the probability that y is the correct label of the sample x other than only give the classification result. By *h(x,y)*, the multi-class problem can be decomposed to several two-class problems. For example: "which is the label of x : y_i or y_j ?" In addition, a label weighting function *q(i,y)* is introduced to attach different degrees of importance to these different two-class problems. With *h(x,y)*and *q(i,y)*, the pseudo-loss used in AdaBoost.M2 is defined as follows:

$$\varepsilon = \frac{1}{2}\sum_{i=1}^{N} D(i)\left(1 - h(x_i, y_i) + \sum_{y \neq y_i} q(i, y) h(x_i, y)\right),$$

where ε denotes the pseudo-loss and D denotes the distribution of samples. So far, it is clear that the pseudo-loss, which can be considered as the criterion of feature selection, is directly related to the classification accuracy of corresponding weak hypothesis/projections.

Projection Selection: AdaBoost.M2

- Given: N samples $\{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$ with labels $y_i \in Y = \{1, ..., C\}$, distribution D over the samples, weak learning algorithm ***WeakLearn***, candidate projection set P, selected projection subset S and T is the size of S.
- Initialize the weight vector: $w_{i,y}^1 = D(i)/(C-1)$ for $i = 1, ..., N$, $y \in Y - \{y_i\}$; set $S = \varnothing$.
- For $t = 1, 2, ..., T$
  1. Set $W_i^t = \sum_{y \neq y_i} w_{i,y}^t$; $q_t(i, y) = \frac{w_{i,y}^t}{W_i^t}$; for $y \neq y_i$; and set $D_t(i) = \frac{W_i^t}{\sum_{i=1}^{N} W_i^t}$;
  2. For each projections p_j in P, get a hypothesis $h_j : X \times Y \to [0,1]$ by ***WeakLearn***, providing the distribution D_t and the label weighting function q_t.
  3. Calculate the pseudo-loss of all h_j:

$$\varepsilon_j = \frac{1}{2}\sum_{i=1}^{N} D_t(i)\left(1 - h_j(x_i, y_i) + \sum_{y \neq y_i} q_t(i, y) h_j(x_i, y)\right)$$

  4. Choose the projection with the least pseudo-loss (suppose to ε), and put it into S. The hypothesis of this projection is supposed to h.
  5. Set $\beta_t = \varepsilon/(1-\varepsilon)$.
  6. For $i = 1, ..., N$, $y \in Y - \{y_i\}$, set the new weights vector

$$w_{i,y}^{t+1} = w_{i,y}^t \beta_t^{(1/2)(1 + h(x_i, y_i) - h(x_i, y))}$$

- Output: a set of selected projections S.

Fig.5. The process of projection selection by AdaBoost.M2.

In AdaBoost.M2, the weak hypothesis has a great influence on the calculation of pseudo-loss as illustrated in Fig.5. Then, we will present the detailed implementation and analysis of the construction of the weak hypothesis.

As above mentioned, the weak hypothesis *h(x,y)* should measure the probability that y is the correct label of the sample x. Considering that the labels of the samples located near x can provide valuable information to determine which class x belongs to, we propose to construct *h(x,y)* by the neighbors of x. Intuitively, around x, if there are more neighboring samples of class j than other classes, x is more likely to be in class j and thus we assign *h(x,j)* a larger value. This idea is formulated as the WeakLearn algorithm shown in Fig.6.

WeakLearn

- Given: N samples $\{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$ with labels $y_i \in Y = \{1, ..., C\}$, distribution D over the samples, K specifying the number of nearest neighbors.
- For $i = 1, 2, ..., N$

  1.Find K nearest neighbors of x_i (KNN_i).

  2.Calculate the weight of each class in KNN_i:

$$W_j = \sum_{x \in KNN_i \cap X^j} D(x),\ j = 1, 2, ..., C,$$

  where X^j is the set of samples of class j.

  3.Set: $h(x_i, j) = \frac{W_j}{\sum_{j=1}^{C} W_j},\ j = 1, 2, ..., C$

Fig.6. The process of constructing weak hypothesis.

Similar to K-Nearest Neighbor algorithm, the number of nearest neighbors K, is an important parameter in our WeakLearn algorithm. Although using more neighbors can increase the robustness of the weak hypotheses to outliers, their discriminability degrades and the algorithm becomes more computationally demanding. Therefore, we should balance the tradeoff between the robustness and the discriminability by determining a proper value of K. In this paper, K is determined by experiments shown in Section 5.

5. Experiments

In this section, we evaluate the proposed ODPP and compare it with other Discriminant Analysis (DA) methods on both synthetic and real-world data sets. DA methods or the proposed ODPP are first used to find a low-dimensional representation of the data, and then the Nearest Neighbor Classifier (NNC) with Euclidean distance is utilized for classification. It is worth noting that the proposed ODPP is a feature extractor rather than a classifier. So, actually, any kind of classifier (e.g. SVM, Decision Tree) can be combined with ODPP for the purpose of classification.

In our experiments, samples in each data set are randomly divided into training set and testing set. In order to reduce the randomness of the experimental results, ten trials with varying randomly sampled training and testing sets are conducted, and the mean and standard deviation of classification accuracies are reported.

5.1. Experiments on Synthetic Data Sets

In this subsection, we compare the first projection found by the ODPP with that by the traditional LDA on four two-dimensional synthetic data sets. Their classification performances on these data sets are also reported.

The distributions of these synthetic data are illustrated in Fig.7. Each data set is randomly divided into training set and testing set with equal size. Data is projected from 2-D to 1-D by the first projection found by LDA or ODPP on the training set, and then the NNC is utilized for classification in the reduced 1D space. TABEL 1 gives the means and standard deviations (in parentheses) of the classification performances on these data sets.

TABEL 1

	LDA	ODPP
Set 1	0.74 (0.045)	0.79 (0.012)
Set 2	0.67 (0.017)	0.96 (0.008)
Set 3	0.97 (0.005)	0.97 (0.004)
Set 4	0.64 (0.010)	0.99 (0.001)

Classification performances of the first projections of LDA and ODPP on four synthetic data sets. Values in bracket are the standard deviations calculated with ten trials.

From TABEL 1, one can clearly find that the proposed ODPP outperforms the traditional LDA very impressively in both the mean and the standard deviation of the classification accuracies, especially on Set 2 and Set 4. The reason behind is intuitively illustrated in Fig.7, which shows the data distribution, as well as the first projection pursued by LDA and ODPP respectively in one of the ten trials. And more analysis is given below.

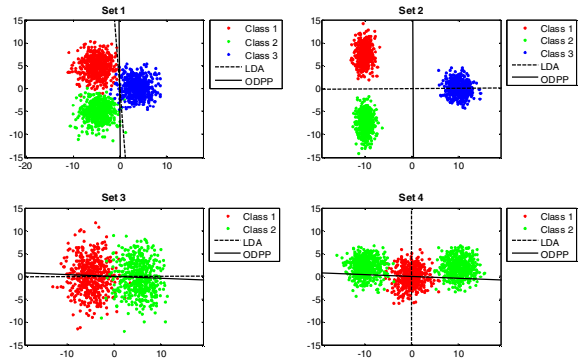

Fig.7. Data distribution of the four synthetic data sets and the first projections of LDA and ODPP on these data sets. Different classes are denoted by different colors.

As shown in Fig.7, for the case of Set 2, the class 1 and the class 2 are very close, but both far away from the class 3. So, in LDA, the Fisher criterion is over-attracted by the large "distance" between class 1/2 and class 3. Therefore, the first projection pursued by LDA cannot separate all the classes very well (actually class 1 and 2 overlap heavily). This problem is naturally solved by ODPP since it is classifiability-based rather than distance-based; so the first projection of ODPP can separate all the classes very well.

In Fig.7, the Set 4 illustrates another case, i.e. non-Gaussian distribution case, in which the class 1 and the class 2 share almost the same mean, but the class 2 contains two clusters. In this case, the traditional LDA fails generally, but the proposed ODPP can still deal with it correctly. The reason behind is that the traditional LDA has the underlying assumption of unimodal Gaussian distribution whereas ODPP do not make any assumption on data distribution.

5.2. Experiments on UCI Data Sets

In this subsection, seven data sets from the UCI databases [18] are used for evaluating the proposed ODPP and comparing it with other DA methods. TABEL 2 gives the detailed information of these data sets.

TABEL 2

Data Set	n	C	N
WDBC	30	2	569
LSD	36	6	6435
MDD-fac	216	10	2000
MDD-fou	76		
MDD-kar	64		
MDD-pix	240		
MDD-zer	47		

Data sets from the UCI database. Information is provided on the initial dimensionality n, the number of class C, and the number of total samples N.

In experiments, samples in each data set except LSD are randomly divided into a training set and a testing set with equal size. For LSD, a training set of 4435 samples and a testing set of 2000 samples are provided by the data creator.

In the implementation of ODPP, two tradeoffs should be balanced. The first is the tradeoff between the amount of candidate projections contained in CPS and the computational cost of the projection selection process. This can be balanced by adjusting the size of the CPS, i.e., $B+ C(C\text{-}1)/2$. Note that C is fixed to the number of classes, so we can only adjust B. Obviously, as B increases, more candidate projections are included in the CPS, thus more possibly one can find better discriminatory projections. However, a large CPS makes the projection selection process more time-consuming. Evidently, the most appropriate B depends on several factors including the number of classes, the feature dimension, and the distribution of the data. In practice, it can be set empirically. For example, we show in Fig.8 that about two hundreds of candidate projections can cover most discriminatory information for the UCI database.

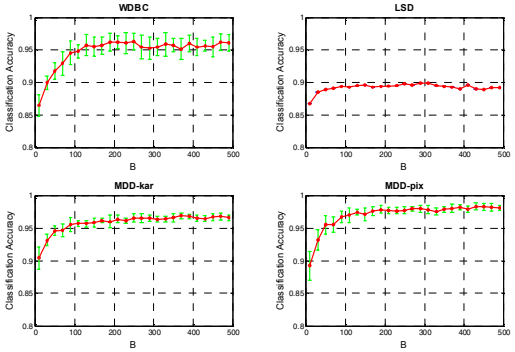

Fig.8. Effect of different parameters B on classification accuracy when K equals to 10. Means (red) and standard deviations (green) of classification accuracies are given.

The second is the tradeoff between robustness and classifiability in the WeakLearn algorithm, which can be balanced by the adjusting the number of nearest neighbors K. We also do some experiments on UCI database to investigate the effect of K on classification accuracy, and the results are shown in Fig.9. Fortunately, one can find that ODPP performs similarly well within a wide range of K value. Therefore, in practice, it can be empirically set safely. In the following experiments, K is fixed to 10.

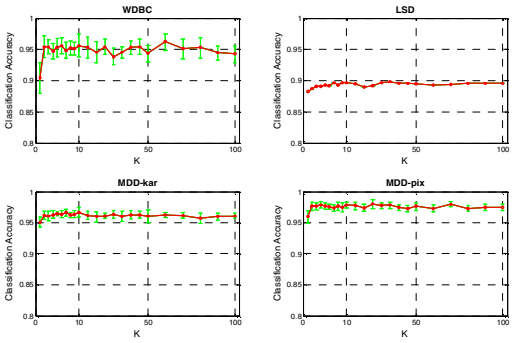

Fig.9. Effect of different numbers of nearest neighbors K on classification accuracy when B equals to 200. Means (red) and standard deviations (green) of classification accuracies are given.

Experiments are then conducted to compare ODPP with other popular DA methods on seven data sets from the UCI database. In order to compare fairly, the experimental setup used in this paper is consistent with that in [7]. The comparison results are shown in TABEL 3, in which the results of NDA, DLDA, HLDA and SDA are cited directly from [7]. From the comparisons, we can see that our method always gives better or comparable performances.

TABEL 3

	LDA	NDA	DLDA	HLDA	SDA	ODPP
WDBC	0.94	0.73	0.87	0.95	0.94	**0.96**
LSD	0.84	0.48	0.86	0.88	0.88	**0.90**
MDD-pix	0.93	0.84	0.95	0.82	0.96	**0.98**
MDD-fou	0.81	0.70	0.80	**0.83**	**0.83**	**0.83**
MDD-fac	0.97	0.79	0.91	0.96	0.96	**0.98**
MDD-kar	0.96	0.90	0.96	**0.97**	**0.97**	**0.97**
MDD-zer	0.79	0.69	0.76	0.79	0.79	**0.82**

Classification performances of ODPP (B=200, K=10) and other DA methods on seven data sets from the UCI database. Bold font in each row denotes the best result on the corresponding data set.

5.3. Experiments on Face Data Sets

Experiments are also conducted on two public face data sets: UMIST [19] and PIE [20]. The UMIST data set contains 575 face images of 20 individuals, which covering a range of poses from profile to frontal views. The PIE data set contains over 40000 face images of 68 individuals, with the variations of 13 different poses, 43 different illuminations and 4 different expressions. Due to its enormous size, only a subset of the PIE data set is used in the following experiments. This subset contains 20 randomly selected individuals with only illumination variations (frontal pose and neutral expression).

For each data set, we randomly divided the face images

into two subsets: 50% for training and 50% for testing. All the face images are cropped and down-sampled to the size of 46 by 56 pixels, and normalized by histogram equalization. Considering the computational feasibility, we firstly reduce the dimensionality to 200 by PCA.

We compare the performances of the proposed ODPP with LDA, DLDA and NDA on these two face data sets. The comparison results are given in Table 4. Note that, since the number of nearest neighbors in NDA is a free parameter, we calculate the results using different choice of this parameter in a typical range and give the average result.

TABEL 4

	LDA	DLDA	NDA	ODPP
UMIST	0.91 (19)	0.95 (19)	0.93 (7)	**0.97** (16)
PIE	0.94 (19)	0.95 (19)	0.93 (17)	**0.96** (30)

Classification accuracies of ODPP (B=200, K=10) and other DA methods on UMIST and PIE face data sets. Values shown in bracket correspond to the optimal dimension of the reduced space obtained by the corresponding methods.

6. Conclusion and Discussion

Taking notice of the limitations of the traditional LDA, this paper proposes Optimal Discriminatory Projection Pursuit (ODPP) as a new computational paradigm for linear feature extraction, which is totally different from LDA and its variants. Two steps are involved in the proposed ODPP: one is the construction of candidate projection set by using the nearest between-class boundary samples; the other is the optimal discriminatory projection pursuit by using AdaBoost with classifiability-based criterion. Both the boundary attribute of the projections and the AdaBoost learning process endow the proposed ODPP impressive generalizability. What is more important is that, in contrast with the traditional LDA, the proposed method almost does not suffer from the four problems of LDA. Extensive comparisons with LDA, NDA, DLDA, HLDA, and SDA on synthetic and real data show that the proposed method always gives better or comparable performances.

The disadvantage of ODPP mainly lies in its relatively slow training stage due to the somewhat time-consuming iterative AdaBoost, especially when the training set is enormous. However, compared with its effectiveness, this limitation becomes trivial especially considering that more and more powerful computers have become available for both academic researchers and engineers.

Acknowledgements

This paper is partially supported by National Natural Science Foundation of China under contract No.60332010, No.60772071, and No. 60728203; Hi-Tech Research and Development Program of China under contract No.2006AA01Z122 and No.2007AA01Z163; 100 Talents Program of CAS; and ISVISION Technology Co. Ltd.

References

[1] K. Fukunaga. Introduction to Statistical Pattern Recognition. Academic Press, Boston, 2nd edition, 1990.

[2] R. Lotlikar and R. Kothari. Fractional-step Dimensionality Reduction. *IEEE Trans. PAMI*, 2000.

[3] M. Röhl and C. Weihs. Optimal vs. Classical Linear Dimension Reduction. In *GfKl*, 1998.

[4] M. Bressan and J.Vitria. Nonparametric Discriminant Analysis and Nearest Neighbor Classification. *Pattern Recognition Letters*, 2003.

[5] Z. Li, W. Liu, D. Lin and X. Tang. Nonparametric Subspace Analysis for Face Recognition. In *CVPR*, 2005.

[6] X. Qiu and L. Wu. Face Recognition by Stepwise Nonparametric Margin Maximum Criterion. In *ICCV*, 2005.

[7] M. Zhu and A. M. Martinez. Subclass Discriminant Analysis, *IEEE Trans. PAMI*, 2006.

[8] M. Loog and R.P.W. Duin. Linear Dimensionality Reduction via a Heteroscedastic Extension of LDA: The Chernoff Criterion, *IEEE Trans. PAMI* , 2004.

[9] S.J. Raudys and A.K. Jain. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Trans. PAMI*, 1991.

[10] D. Swets and J. Weng. Using Discriminant Eigenfeatures for Image Retrieval. *IEEE Trans. PAMI*, 1996.

[11] L. Chen, H. Liao, M. Ko, J. Lin and G. Yu. A New LDA-based Face Recognition System which can solve the small sample size problem, *Pattern Recognition*, 2000.

[12] J. Yang, D. Zhang, and J.Y. Yang. A Generalised K-L Expansion Method Which Can Deal with Small Sample Size and High-Dimensional Problems, *Pattern Analysis & Applications*, 2003.

[13] R. Huang, Q. Liu, H. Lu and S. Ma. Solving the small size problem of LDA. In *ICPR*, 2002.

[14] H. Cevikalp, M. Neamtu, M. Wilkes and A. Barkana. Discriminative Common Vectors for Face Recognition. *IEEE Trans. PAMI*, 2005.

[15] H. Yu and J. Yang. A Direct LDA Algorithm for High-Dimensional Data with Application to Face Recognition, *Pattern Recognition*, 2001.

[16] C. Xiang, X. Fan and T. H. Lee. Face Recognition Using Recursive Fisher Linear Discriminant. *IEEE Trans. Image Processing*, 2006.

[17] Y. Freund and R.E. Schapire. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 1997.

[18] C.L. Blake and C.J. Merz. UCI Repository of Machine Learning Databases. Dept. of Information and Computer Sciences, University of California, Irvine, 1998.

[19] D. B. Graham and N. M. Allinson. Characterizing Virtual Eigensignatures for General Purpose Face Recognition. *Face Recognition: From Theory to Application*, *Computer and Systems Sciences*, 1998.

[20] T. Sim, S. Baker and M. Bsat. The CMU Pose, Illumination, and Expression (PIE) Database, In *AFGR*, 2002.