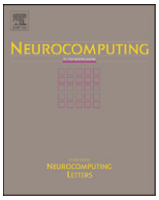
Neurocomputing 74 (2011) 3690–3694

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Modified orthogonal discriminant projection for classification

Shanwen Zhang [a], Ying-Ke Lei [a,b,c,*], Yan-Hua Wu [c], Jun-An Yang [c]

[a] Institute of Intelligent Machines, Chinese Academy of Sciences, P.O. Box 1130, Hefei, Anhui 230031, China
[b] Department of Automation, University of Science and Technology of China, Hefei, Anhui 230027, China
[c] Electronic Engineering Institute, Hefei, Anhui 230027, China

ARTICLE INFO

Article history:
Received 15 March 2010
Received in revised form
30 July 2011
Accepted 31 July 2011
Communicated by T. Heskes
Available online 16 August 2011

Keywords:
Manifold learning
Weight
Orthogonal discriminant projection (ODP)
Modified ODP

ABSTRACT

From the perspective of manifold learning, the weight between two nodes of graph plays an indispensable role, which provides the similarity between pairwise nodes, and can effectively reveal the intrinsic relationship between data classes. In the original Locality Preserving Projections (LPP), Unsupervised Discriminant Projection (UDP), Orthogonal LPP (OLPP), and other spectral mapping methods, the weight between two points is usually defined as a heat kernel or simply 0–1 weight, which cannot effectively reflect the sample class information. In Orthogonal Discriminant Projection (ODP), the weight between two points was defined based on their local information and class information, but it is not a monotonically decreasing with the increase of the distance between two nodes, so it is not very sound. In this paper, we first analyze the defect of the weight in ODP, then propose a novel weight measure between two nodes of a graph by combining their label information and local information, finally present a modified ODP algorithm following the ODP technique. The modified ODP algorithm can explore the intrinsic structure of original data and enhance the classification ability. The experimental results show that the modified ODP algorithm is effective and feasible.

1. Introduction

Manifold learning based methods are becoming the most promising dimensional reduction approaches. Among them, Laplacian Eigenmap (LE) [1] and Locally Linear Embedding (LLE) [2] are two representative spectral mapping methods, which can implicitly find the optimal feature subspace by solving a generalized eigenvalue problem. However, LE and LLE are nonlinear dimensional reduction techniques, whose generalization ability is much weak. They are defined only on the training data points and it is unclear how to evaluate the projection of new test points. That is to say, a sample of the test set in the low dimensional space cannot be easily obtained with the projection results of the training set. Linearization, kernelization, tensorization and some other tricks have been introduced to avoid this problem [3–8]. Locality Preserving Projections (LPP) [5,6] is a linear approximation to LE. Different from the nonlinear dimensional reduction techniques such as LE and LLE, LPP is linear. Particularly, it is defined everywhere in ambient space rather than just on the training data points. LPP shares many of the data representation properties of nonlinear techniques such as LE or LLE. But in the original LPP, the linear transformation matrix is not under the orthogonal constraint. In order to solve the problem, Cai, et al. [9] proposed an orthogonal LPP (OLPP) algorithm, which shows more locality preserving power than LPP. However both LPP and OLPP are unsupervised dimension reduction methods. They ignore the class information and often cause small sample size (SSS) problem when the sample number is less than the dimension of the samples. Some techniques are introduced to solve the problem at the cost of discarding some useful information [10–12]. Recently, Li et al. [13] proposed an orthogonal discriminant projection (ODP) algorithm. ODP maximizes the weighted difference between the non-local scatter and the local scatter. In ODP, the weights between two nodes of a graph are adjusted according to their class information and local information. Although ODP can offer higher recognition rate than some other feature extraction methods, it is found that the weight definition is not sound. In this paper, we first analyze the defect of the weight in ODP, and propose a modified weight measure, then present a modified ODP.

The rest of this paper is organized as follows: Section 2 analyses the shortage of the weight in ODP. Section 3 introduces the modified ODP algorithm. Experimental results are given in Section 4. Finally, some concluding remarks are provided in Section 5.

2. Motivation

In this section, we firstly introduce the definition and plot illustrate of the weight in ODP [13], and then analyze the defect of

[*] Corresponding author at: Institute of Intelligent Machines, Chinese Academy of Sciences, P.O. Box 1130, Hefei, Anhui 230031, China. Tel.: +860 551 5591108.
E-mail address: leiyingke@gmail.com (Y.-K. Lei).

0925-2312/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2011.07.015

the weight. Given n data points $x_1,x_2,\ldots,x_n$, $x_i=\{x_i^m|m=1,2,\ldots,M\}$, let c_i and $N(x_i)$ be the label and k nearest neighbors of the point x_i, respectively. The weight W_{ij} between two nodes is defined as follows:

$$W_{ij}=\begin{cases}\exp(-\frac{d^2(x_i,x_j)}{\beta}), & \text{If } x_i\in N(x_j) \text{ and} x_j\in N(x_i) \text{ and } c_i=c_j\\ \exp(-\frac{d^2(x_i,x_j)}{\beta})\left(1-\exp(-\frac{d^2(x_i,x_j)}{\beta})\right) & \text{If } x_i\in N(x_j) \text{ and} x_j\in N(x_i) \text{ and } c_i\neq c_j\\ 0, & \text{otherwise}\end{cases} \tag{1}$$

where $d(x_i,x_j)$ is the Euclidean distance between x_i and x_j, and β is a control parameter.

Fig. 1 shows that the typical plot of W_{ij} is a function of $d^2(x_i,x_j)/\beta$, where S1 denotes the case that both x_i and x_j are the k nearest neighbors of each other sharing the same label; S2 denotes the case that both x_i and x_j are k nearest neighbors of each other with different labels and S3 denotes the other cases.

From Eq. (1) and Fig. 1, we find that the weight W_{ij} is not a monotonically decreasing function of $d^2(x_i,x_j)/\beta$. More specifically, for the case of two points with the different labels, when $0<d^2(x_i,x_j)/\beta<0.71$, W_{ij} is monotonically increasing. But, in actual applications, W_{ij} should decrease with the increase of $d^2(x_i,x_j)/\beta$. Moreover the authors [13] pointed that the parameter β can be set to $\beta\to+\infty$. That is to say, there are many pairwise points satisfying $0<d^2(x_i,x_j)/\beta<0.71$. According to the physical meaning of the weight, the definition of W_{ij} in ODP is not entirely correct.

3. Modified ODP

In this section, combining the class information and the local information of the original data, a new weight measure between two nodes of the neighbor graph is proposed. Based on the weight, a modified ODP is presented. The procedure of the modified ODP algorithm is similar to that of the ODP algorithm except W_{ij} is replaced by H_{ij}.

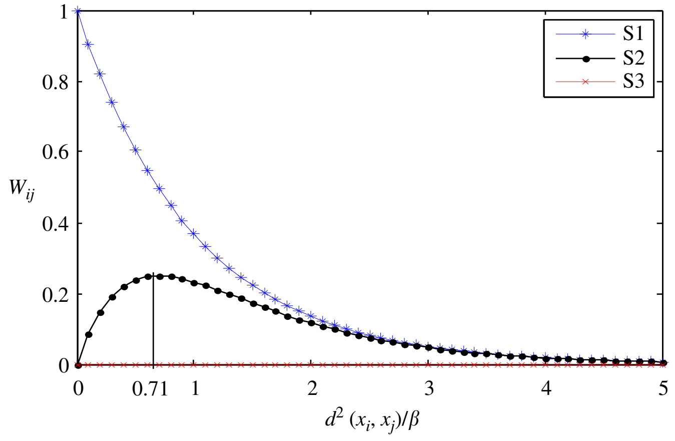


Fig. 1. Typical plot of W_{ij} as a function of $d^2(x_i,x_j)/\beta$.

Given n data points $x_1,x_2,\ldots,x_n$ in R^D, like ODP, the aim of the modified ODP is to find a linear transformation matrix A and to map the n points to a set of points $y_1,y_2,\ldots,y_n$ in R^d, such that y_i provides the most faithful representation of x_i in the lower dimensional space, where $y_i=A^Tx_i$ and $d\ll D$.

3.1. Weight definition

To obtain a suitable weight, we modify Eq. (1) as follows:

$$H_{ij}=\begin{cases}\exp(-\frac{d^2(x_i,x_j)}{\beta}), & \text{If } x_i\in N(x_j) \text{ and } x_j\in N(x_i) \text{ and } c_i=c_j\\ \exp(-\frac{d^2(x_i,x_j)}{\beta})S(x_i,x_j) & \text{If } x_i\in N(x_j) \text{ and } x_j\in N(x_i) \text{ and } c_i\neq c_j\\ 0, & \text{otherwise}\end{cases} \tag{2}$$

where $S(x_i,x_j)$ is the correlation coefficient between x_i and x_j, which is expressed as

$$S(x_i,x_j)=\frac{\|(x_i-\overline{x_i})(x_j-\overline{x_j})^T\|}{\|(x_i-\overline{x_i})\|\cdot\|(x_j-\overline{x_j})\|}=\frac{\sum_m(x_i^m-\overline{x_i})(x_j^m-\overline{x_j})^T}{\sqrt{\sum_m(x_i^m-\overline{x_i})^2\sum_m(x_j^m-x_j)^2}} \tag{3}$$

where

$$\overline{x_i}=\frac{1}{M}\sum_{m=1}^{M}x_i^m,\quad \overline{x_j}=\frac{1}{M}\sum_{m=1}^{M}x_j^m$$

Like W_{ij}, H_{ij} integrates the local neighbor structure and the class information of the original data, and displays the discriminant similarity between x_i and x_j. The primary difference between H_{ij} and W_{ij} is that H_{ij} is strictly monotonically decreasing in the k nearest neighbors of x_i and x_j. H_{ij} not only shares the properties of W_{ij}, but also has following properties:

1. Due to introducing the correlation coefficient $S(x_i,x_j)$ into H_{ij}, H_{ij} can reflect the data similarity relationship.
2. Since $0<S(x_i,x_j)<1$, H_{ij} incurs a heavy penalty if the neighboring points x_i and x_j belong to the different labels.
3. Note that $\exp(-d^2(x_i,x_j)/\beta)$ and $\exp(-d^2(x_i,x_j)/\beta)S(x_i,x_j)$ always decrease when x_i and x_j are far apart and they increase when x_i and x_j are close. So if we replace W_{ij} by H_{ij} in ODP, the modified ODP does not impose far apart points to be close, by which the neighborhood structure of the original data tends to be preserved.

3.2. Computing the local and non-local scatter matrix

Due to introducing the weight matrix H, by simple algebraic formulation, the local scatter matrix $J_L(A)$ can be expressed as follows:

$$\begin{aligned}J_L(A)&=\sum_{i=1}^{n}\sum_{j=1}^{n}H_{ij}(y_i-y_j)(y_i-y_j)^T\\&=\sum_{i=1}^{n}\sum_{j=1}^{n}H_{ij}(A^Tx_i-A^Tx_j)(A^Tx_i-A^Tx_j)^T\\&=A^T\sum_{i=1}^{n}\sum_{j=1}^{n}H_{ij}(x_i-x_j)(x_i-x_j)^TA\\&=A^TS_LA\end{aligned} \tag{4}$$

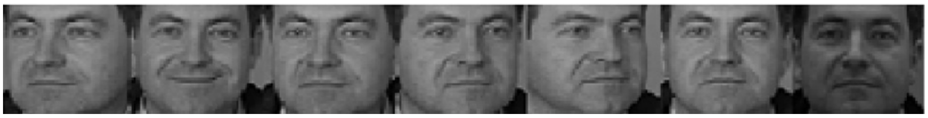


Fig. 2. Seven images of one person in the FERET face database.

where $S_L=\sum_{i=1}^{n}\sum_{j=1}^{n}H_{ij}(x_i-x_j)(x_i-x_j)^T=XLX^T$ and $X=[x_1,x_2,\ldots,x_n]$, L is the Laplacian matrix with the definition of $L=M-H$; M is a diagonal matrix, its entries are column (or row, since H is symmetric) sum of H, i.e. $M_{ii}=\sum_j H_{ij}$.

The matrix M provides a natural important measure of the data points. If M_{ii} is large, it implies that the class containing x_i has a high density around x_i. Therefore, the bigger the value M_{ii} is, the more important the data point x_i is.

The non-local scatter matrix $J_N(A)$ is characterized as follows:

$$J_N(A)=\sum_{i=1}^{n}\sum_{j=1}^{n}(1-H_{ij})(y_i-y_j)(y_i-y_j)^T \tag{5}$$

Similarly, J_N can be derived as

$$J_N(A)=A^T(S_T-S_L)A=A^TS_NA \tag{6}$$

where

$$S_T=\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i-x_j)(x_i-x_j)^T \text{ and}$$

$$S_N=S_T-S_L=\sum_{i=1}^{n}\sum_{j=1}^{n}(1-H_{ij})(x_i-x_j)(x_i-x_j)^T$$

3.3. Extracting classification feature

Like ODP, with the constraint $A^TA=I$, the objective function of the modified ODP is as follows:

$$\underset{A^TA=I}{\arg\max}J(A)=\underset{A^TA=I}{\arg\max}A^T((1-\gamma)S_T-\gamma S_L)A \tag{7}$$

where γ is an adjustable parameter.

So we can find that A consists of the eigenvectors associated with d top eigenvalues of the following eigen-equation:

$$((1-\gamma)S_T-\gamma S_L)a=\lambda a \tag{8}$$

Let $a_1,a_2,\ldots,a_d$ be the first d solution of Eq. (8), which are selected according to their top d eigenvalues $\lambda_1,\lambda_2,\ldots,\lambda_d$, where $\lambda_1>\lambda_2>\cdots>\lambda_d$. The optimal projection matrix A_{opt} is obtained by $A_{opt}=[a_1,a_2,\ldots,a_d]$. Then the optimal linear feature y_{new} of any new test point x_{new} is obtained by the following linear transformation:

$$y_{new}=A^Tx_{new} \tag{9}$$

where $y_{new}\in R^d$.

3.4. Modified ODP for classification

The algorithmic procedure for data classification of the modified ODP algorithm is formally summarized as follows:

Step 1: Construct the adjacency graph using the training data;
Step 2: Calculate the weight of any two data nodes by Eq. (2);
Step 3: Compute the top d eigenvalues and its corresponding eigenvectors of the generalized eigenvalue problem in Eq. (8), and obtain the final linear projection matrix A;
Step 4: Project the test data into low-dimensionality feature representation by Eq. (9);
Step 5: Predict the corresponding class labels using the optimal classifier.

4. Experiments

To evaluate the performance of the modified ODP algorithm, in this section, we conduct a series of experiments on the FERET face, Extended Yale B face, and the plant leaf datasets, and compare with ODP. Since the main purpose of the experiments is to compare the performances of ODP and the modified ODP, the 1-NN classifier is used in during all the experiments for its simplicity.

4.1. Experiments on FERET face database

In this experiment, a subset is selected from the original FERET face database. It contains 100 individuals with seven images for each person. It is composed of images whose names are marked with two-character strings, i.e., "bd", "bj", "bf", "be", "bc", "ba", and "bk", which denotes two facial expression images, two left pose images, two right pose images and an illumination image, respectively. All images in this subset are cropped to be the size of 32 × 32. Fig. 2 shows the cropped images of one person in FERET face database.

We selected the l=3, 4, 5, and 6 images from each class as training set, and the rest of each class for test set. For each value of l, 20 runs are performed with different random partitions between training set and test one. The experimental conditions,

Table 1
Recognition rates (percent) of the modified ODP and ODP on FERET database and their corresponding dimensions (shown in parentheses).

Method	l=3	l=4	l=5	l=6
ODP	80.73 (100)	82.58 (100)	83.24 (100)	83.95 (100)
Modified ODP	**81.79 (100)**	**83.31 (98)**	**84.05 (98)**	**84.86 (98)**

Table 2
Recognition rates (percent) of the modified ODP and ODP on Extended Yale B database and their corresponding dimensions (shown in parentheses).

Method	l=10	l=50
ODP	86.31 (65)	97.39 (67)
Modified ODP	**88.54 (71)**	**97.95 (60)**

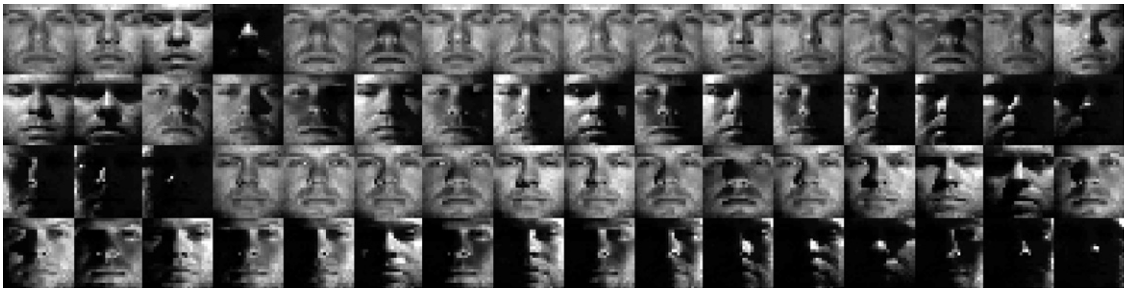
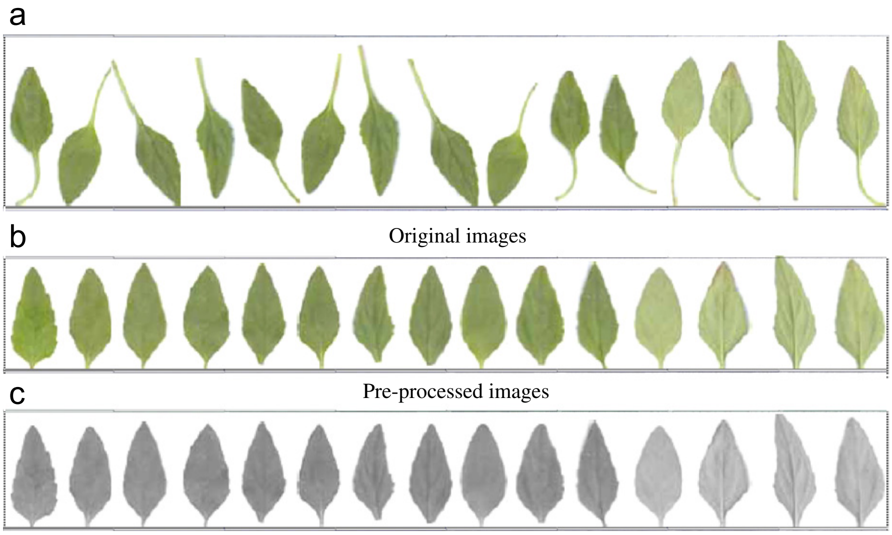


Fig. 3. Sample images from one person in the extended Yale B database.



Original images

Pre-processed images

Pre-processed gray images

Fig. 4. Fifteen leaf images of Spica Prunellae under three periods and five illuminations: (A) Original images; (B) Pre-processed images; (C) Pre-processed gray images.

Table 3
Recognition rates (percent) of the modified ODP and ODP on the plant leaf image database and their corresponding dimensions (shown in parentheses).

Method	$l=4$	$l=6$	$l=8$	$l=10$	$l=12$	$l=14$
ODP	76.81 (35)	82.04 (35)	87.72 (35)	89.55 (35)	90.36 (35)	91.04 (35)
Modified ODP	**81.07 (35)**	**85.69 (35)**	**90.84 (35)**	**91.73 (30)**	**92.04 (30)**	**92.87 (30)**

parameters, and procedures of the ODP and modified ODP are basically same. When constructing the k nearest neighborhood graph, k is set to be $l-1$. After the discriminant features have been extracted by performing the two algorithms, the 1-NN classifier is adopted to predict the labels of the test data, where the control parameter $\beta=800$ and the adjustable parameter $\gamma=0.8$. Table 1 illustrates the optimal recognition rates and their corresponding dimensions from the 20 runs for the ODP and modified ODP, From Table 1, it can be found that the modified ODP consistently outperforms ODP among all the cases.

4.2. Experiments on Extended Yale B face database

The Extended Yale B database consists of 2414 frontal-face images of 38 individuals, where each subject has about 64 images captured under various laboratory-controlled lighting conditions. We simply used the cropped images and resized them to 32 × 32 pixels. Sixty four sample images of one individual are displayed in Fig. 3. We set the parameters k, β, and γ to be the same as before. For the ODP and modified ODP methods, two random subsets with ten and fifty images per individual were selected for training. The rest of the database was used for testing. Such a trial was independently performed 20 times, and then the average recognition results were calculated. The maximal recognition rate of each method and the corresponding reduced dimension are given in Table 2. As can be seen, the modified ODP performs better than ODP regardless of whether the training sample size is 10 or 50.

4.3. Experiments on ICL plant leaf database

To more comprehensively verify the proposed method, we apply it to the plant leaf classification. The ICL-PlantLeaf[1] database was constructed at the Intelligent Computing Laboratory (ICL) of Institute of Intelligent Machines, Chinese Academy of Sciences. It contains more than 30,000 leaf images of 362 plant species. The images were captured at different periods, and have different locations and natural illuminations. In this experiment, we employed a subset containing 750 leaf images of 50 classes. Each kind of plant leaf was sampled and imaged from 3 different periods under 5 different nature illuminations (or locations). Fig. 4 shows the 15 leaf images of Spica Prunellae. In this experiment, each original leaf image is cropped and normalized (in scale and orientation), as shown in Fig. 4(B). The size of each cropped image is 32 × 32 pixels by histogram equilibrium in the experiment with 256 Gy levels per pixel and with the white background, as shown in Fig. 4(C). Thus, each image is represented by a 1024-dimensional vector in image space.

All images are randomly divided into the training subset and test subset with different numbers, i.e., we randomly choose l images from each class as training subset, the rest as test subset. We set $k=l-1$, $\beta=500$ and $\gamma=0.8$. For each given l, we perform 20 splits to randomly choose the training set. Table 3 shows the average recognition rates with their corresponding standard deviations and dimensions over 20 randomly splits. As can be seen, the modified ODP significantly outperforms the ODP, irrespective of the variations in training sample size.

The proposed modified ODP consistently achieves the best recognition rate in all the experimental cases. The datasets used in this study are FERET face, Extended Yale B face, and ICL plant leaf. The multifaceted nature of the datasets enables us to perform a more objective comparison of the tested algorithms. Compared to the ODP, the modified ODP encodes more discriminant information in the reduced feature subspace by more faithfully preserving local geometry and incorporating the data similarity information.

[1] http://www.intelengine.cn/ dataset/index.html.

5. Conclusions

In pattern recognition and classification, dimensional reduction algorithms are widely employed to reduce the dimensionality of the original data and enhance the discriminant information. The orthogonal discriminant projection (ODP) algorithm makes use of the local information and the non-local information as well as the sample class information to model the manifold data, in which the weight between two nodes of the graph is adjusted according to their class information and local information. Although ODP is suitable for the task of classification, the weight definition in ODP is not very sound. In this paper, we analyzed the defect of the weight in ODP, and proposed a novel weight measure of any two points by combining the label information and local and non-local information, which can be introduced to improve the discriminant ability and preserve the local neighborhood structure of the original data. Based on the weight, we presented a modified ODP algorithm. We have applied the modified ODP algorithm to face and plant leaf recognition. The experiments on the FERET face, Extended Yale B face, and ICL plant leaf datasets demonstrate that the modified ODP algorithm is effective and feasible.

Acknowledgments

The authors would like to thank the associate editor and the anonymous reviewers for their careful reading and constructive suggestions. This work was supported by the Grants of the National Science Foundation of China nos. 60975005, 60705007, 60872113, and 61005010, the Grant of China Postdoctoral Science Foundation no. 20100480708, and the Grant of the Key Scientific Research Foundation of Education Department of Anhui Province no. KJ2010A289.

References

[1] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Computation 15 (6) (2003) 1373–1396.

[2] S.T. Roweis, L.K. Saul, Nonlinear dimensional reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.

[3] L.K. Saul, S.T. Roweis, Think globally, fit locally: unsupervised learning of low dimensional manifolds, Journal of Machine Learning Research 4 (2003) 119–155.

[4] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Graph embedding: a general framework for dimensionality reduction, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (1) (2007) 40–51.

[5] X. He, S. Yang, Y. Hu, P. Niyogi, H.J. Zhang, Face recognition using laplacianfaces, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (3) (2005) 328–340.

[6] H. Zhao, S. Sun, Z. Jing, J. Yang, Local structure based supervised feature extraction, Pattern Recognition 39 (2006) 1546–1550.

[7] J. Yang, J.Y. Yang, Why can LDA be performed in PCA transformed space? Pattern Recognition 36 (2) (2003) 563–566.

[8] J. Yang, A.F. Frangi, D. Zhang, J.-y. Yang, J. Zhong, KPCA plus LDA: a complete kernel Fisher discriminant framework for feature extraction and recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (2) (2005) 230–244.

[9] D. Cai, X. He, J. Han, H. Zhang, Orthogonal laplacianfaces for face recognition, IEEE Transactions on Image Processing 15 (11) (2006) 3609–3614.

[10] P. Howlanda, J. Wangb, H. Parkc, Solving the small sample size problem in face recognition using generalized discriminant analysis, Pattern Recognition 39 (2006) 277–287.

[11] W. Zheng, L. Zhao, C. Zou, Foley–Sammon optimal discriminant vectors using kernel approach, IEEE Transactions on Neural Networks 16 (1) (2005) 1–9.

[12] D.L. Swets, J. Weng, Using discriminant eigenfeatures for image retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (8) (1996) 831–836.

[13] B. Li, C. Wang, D.S. Huang, Supervised feature extraction based on orthogonal discriminant projection, Neurocomputing 73 (2009) 191–196.

Shanwen Zhang received the B.Sc., M.Sc. and Ph.D. degrees in Mathematics from Northwest University Xi'an, China, Computer Science and Technology from Northwest Polytechnic University Xi'an, China, Electromagnetic Field and Microwave Technology from Air Force Engineering University Xi'an, China in 1988, 1995, and 2001, respectively. From July 1988 to September 2007, he worked at Missile Institute at Air Force Engineering University, Xi'an, China. In October 2007, he joined the Institute of Intelligent Machines, Chinese Academy of Sciences for postdoctoral studies. Research interests: Wavelet Transforms, Rough Sets, Genetic Algorithm, Manifold Learning.

Ying-Ke Lei received the B.E. degree in communication engineering from the Electronic Engineering Institute, China, in 1998, and the Ph.D. degree in Pattern Recognition and Intelligent System from the University of Science and Technology of China (USTC), China in 2010. Research Interests: Manifold Learning, Pattern Recognition, Compressed Sensing, Bioinformatics, Signal Processing.

Yan-Hua Wu received the B.E. degree in Communication Engineering in 1995 and the Ph.D. degree in 2005 from the Electronic Engineering Institute, Hefei, China. He is currently an associate professor in the Department of Information, Electronic Engineering Institute, China. Research Interests: Pattern Recognition, Data Mining, Signal Processing.

Jun-An Yang received the B.E. degree in Wireless Engineering from Southeast University in 1986, and the Ph.D. degree in Signal and Information Processing from University of Science and Technology of China in 2003. He is currently a professor in the Department of Information, Electronic Engineering Institute, China.