Neurocomputing 179 (2016) 54–68

Contents lists available at ScienceDirect

Neurocomputing

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

Discriminative orthogonal elastic preserving projections for classification

Tingjin Luo [a], Chenping Hou [a,*], Dongyun Yi [a], Jun Zhang [b]

[a] College of Science, National University of Defense Technology, Changsha 410073, China
[b] College of Information System and Management, National University of Defense Technology, Changsha 410073, China

ARTICLE INFO

Article history:
Received 13 March 2015
Received in revised form
6 July 2015
Accepted 15 November 2015
Communicated by Feiping Nie
Available online 15 December 2015

Keywords:
Machine learning
Manifold learning
Linear dimensionality reduction
Orthogonal analysis
Discriminative orthogonal elastic preserving projections (DOEPP)

ABSTRACT

The traditional manifold learning methods can preserve the local sub-manifold structure or the global geometry effectively, such as elastic preserving projections (EPP). Many experimental results have been shown that EPP, a recently developed linear algorithm, is a strong analyzer for high-dimensional data. However, for classification problems, the traditional methods focused on the geometrical information and ignores discriminative information of different classes. In this paper, we propose a novel discriminative orthogonal elastic preserving projections (DOEPP) by imposing the discriminant information and the orthogonal constraint to improve its classification performance. DOEPP does not only preserve the elasticity of the training set, but also sufficiently utilizes the discriminant information by adding maximum margin criterion and the orthogonality of the projection matrix into its objective function. Extensive experiments on two well-known synthetic manifold data sets and four publicly available databases illustrate the effectiveness of our method.

1. Introduction

In many real applications, with the development of information acquiring technology, the raw data gathered from sensors are often with very large dimensions. Large dimensions lead to difficulty of data analysis and data mining [1–3], even cannot reveal the hidden intrinsic structure of such data, besides many classifiers perform poorly in a high-dimensional space given a small number of training samples. And the direct process of the high dimensional data is also computationally expensive. Dimensionality reduction [4–8] is an effective way to solve these problems.

Over the last two decades, many algorithms have been developed for dimensionality reduction. The most popular conventional methods are principal component analysis (PCA) [9], linear discriminant analysis (LDA) [10] and maximum margin criterion (MMC) [11]. However, PCA, which finds the directions along the maximum variance of samples, is probably the most popular dimensionality reduction method. PCA is unsupervised and not suitable for classification problems since it does not use prior knowledge of class identities. Unlike PCA, LDA is supervised and finds the projection directions by maximizing the trace of the between-class scatter matrix S_B while minimizing the trace of the within-class scatter matrix S_W. But LDA requires S_W which is non-singular during optimizing its objective functions. While in practice, due to the curse of dimensionality, while S_W is always singular in practice. To solve this problem, Li et al. [11] proposed maximum margin criterion (MMC) which eliminates the effect of S_W's singularity for the small size samples (SSS) problem. It is well known that MMC is to maximize the trace of the difference of between-class scatter matrix S_B and within-class scatter matrix S_W which is viewed as a variant of LDA. However, these three algorithms preserved only the global Euclidean structure and cannot discover the nonlinear sub-manifold structure underlying in the high-dimensional data.

Seung and Lee [12] presented that manifolds are fundamental to perception. This implies that the population activity is constrained to lie on a low-dimensional manifold. Recently, a number of manifold learning algorithms [13–18] have been developed to excavate nonlinear sub-manifold structure lying on the observation space and to find out the low-dimensional and compact embeddings of the high dimensional samples. Furthermore, these manifold learning algorithms have been successfully applied into palm recognition [19], face recognition [20], facial expression recognition [21] and human gait recognition [22]. The most popular nonlinear manifold learning methods include ISOMAP [13], locally linear embedding (LLE) [14], Laplacian eigenmaps (LE) [15], Hessian LLE (HLLE) [16], local linear coordination (LLC) [17], maximum variance unfolding (MVU) [18], and local tangent space alignment (LTSA) [23]. ISOMAP, a variant of multidimensional scaling (MDS) [24], preserved the global geodesic distances by

* Corresponding author.
E-mail addresses: luotingjin03@nudt.edu.cn (T. Luo), hcpnudt@hotmail.com (C. Hou), dongyunyi@sina.com (D. Yi), aiya0215@263.net (J. Zhang).

http://dx.doi.org/10.1016/j.neucom.2015.11.037
0925-2312/

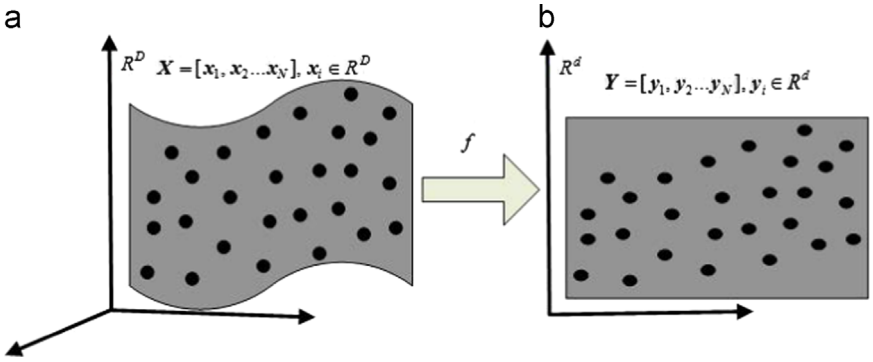

Fig. 1. The process of dimensionality reduction.

Table 1
Important notations used in this paper.

Notation	Description	Notation	Description
X	The given data set	x_i	ith sample of X
Y	The embedding of X	y_i	The embedding of x_i
N	Number of samples	l_i	The label of x_i
D	Dimension of the original samples	U	Projection matrix
d	Reduced dimension	S_B	Between-class scatter matrix
f	Mapping function	S_W	Within-class scatter matrix
Ξ	Label set of α	x_i	Control parameter
C	Number of classes	β	Control parameter
W_l	Weight matrix of local graph	W_g	Weight matrix of global graph
R^D	D-dimensional Euclidean space		

computing shortest paths of all sample pairs. LLE first embedded data points in a low dimensional space by finding the optimal linear reconstruction coefficients in a small neighborhood to represent the local geometry. LE preserved proximity relationships on an undirected weighted graph, which indicates neighbor relations of pairwise samples. LTSA exploited the local tangent information as a representation of the local geometry, and this local tangent information is then aligned to provide a global coordinate. Unfortunately, all of these methods suffer from the out-of-sample problem [25] when the new samples are added.

To solve out-of-sample problem of nonlinear methods, He et al. adopted a linearization procedure to construct explicit maps over new samples and then proposed locality preserving projection (LPP) [26] and neighborhood preserving embedding (NPE) [27], which are the linearization of LLE and LE, respectively. LPP attempts to find the embedding and the projection matrix U by preserving local structural information, while NPE computes U and the corresponding embedding by preserving local neighborhood information. Unfortunately, LPP and NPE are non-orthogonal, this makes them difficult to reconstruct the data. In the recent research [28,29], Kokiopoulou and Saad pointed out that enforcing an orthogonality relationship between the projection directions is more effective for preserving the intrinsic manifold of high dimensional data and then proposed orthogonal neighborhood preserving projection (ONPP) [29] to improve its performance. Based on this idea, Cai et al. [30] developed orthogonal LPP (OLPP) algorithm to produce orthogonal basis. At the same time massive experiments have shown that OLPP and ONPP have more locality preserving and discriminant power than LPP and NPE, respectively.

To incorporate the advantages of both the local sub-manifold structure and the global Euclidean information, Zang et al. [31] proposed elastic preserving projections (EPP). However, they are unsupervised in nature and fail to discover the discriminant information samples from different classes. To improve the discriminative power for classification tasks, Yan et al. [32] presented the unified graph embedding framework and simultaneously proposed marginal fisher analysis (MFA). Zhang et al. [33] provided patch alignment, which is another unified view of popular manifold learning algorithms. Based on this framework, the discriminative locality alignment (DLA) has been developed.

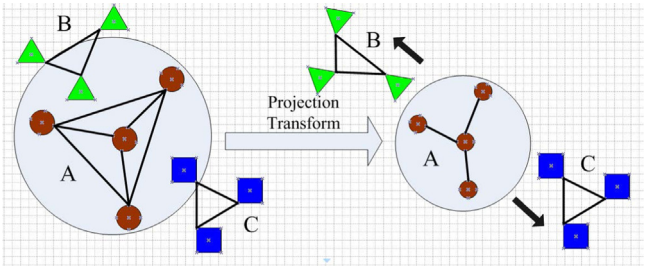

Fig. 2. The process of MMC based on graph viewpoint.

In this paper, we propose a new manifold learning algorithm termed discriminative orthogonal EPP (DOEPP) to incorporate the advantages of both EPP and MMC. Although EPP can explore the local sub-manifold and global structure information, it ignores the discriminative and orthogonality information of high dimensional data for classification tasks. Considering the fact that MMC can preserve the discriminative information sufficiently and the importance of orthogonality of the projection matrix, maximum margin criterion and the orthogonal constraint are introduced into the objective function of DOEPP, which has two advantages: (1) it retains the elastic merits of EPP and MMC; (2) further improves the discriminant power of EPP by adding the discriminative information. We apply our method into two 3D synthetic manifolds and four public datasets (ORL, Yale, Extended YaleB and COIL-20) to valuate its performance. Experimental results show that DOEPP is more suitable for recognition tasks than EPP and MMC.

The rest of the paper is organized as follows: In Section 2, a brief review of MMC and EPP is given. Section 3 presents the basic idea of the proposed DOEPP. Section 4 verifies the effectiveness of our method through a variety of experiments, and Section 5 presents our conclusions.

2. A review of MMC and EPP

In this section, we will first introduce some notations and then take a summary on MMC and EPP briefly. Consider the data set X,

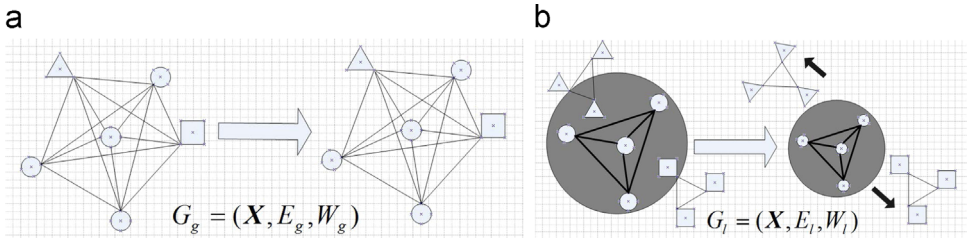

Fig. 3. Graph models of DOEPP. Left: global graph G_g and right: local prior graph G_l.

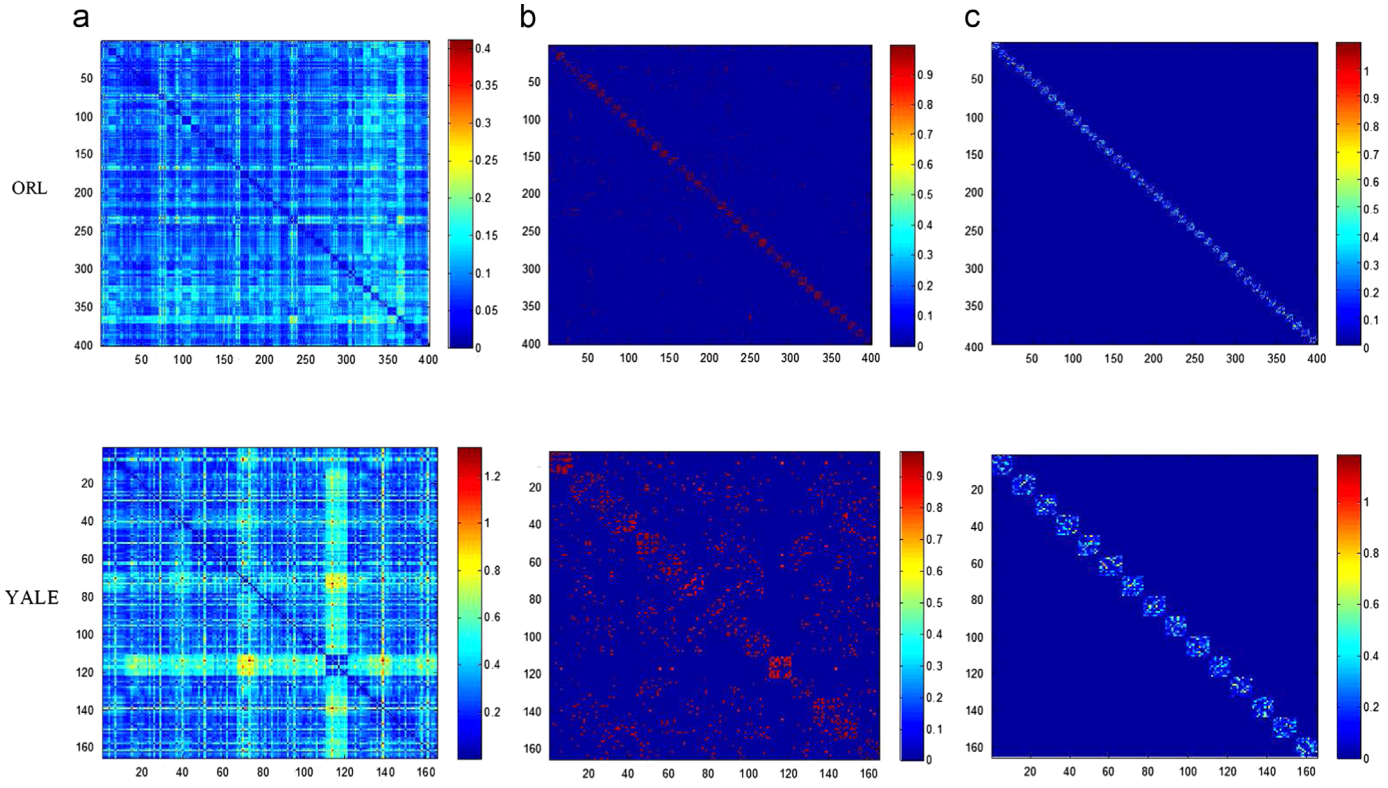

Fig. 4. The visualization cases of weighted matrix on ORL and Yale databases: (a) W_g, (b) W_l of EPP and (c) W_l of DOEPP.

which consists of N samples x_i $(1 \le i \le N)$ from the high-dimensional space R^D. That is $X = [x_1, x_2 \ldots x_N]$ and $\Xi = [l_1, l_2 \ldots l_N]$, where $x_i \in R^D$ and its label $l_i \in \{1, 2 \ldots C\}$. The objectives of a dimensionality reduction are to search the optimal mapping function f and to compute the faithful low-dimensional representations $Y = [y_1, y_2 \ldots y_N] \in R^{d \times N}$, where $d < D$, as shown in Fig. 1. For the linear methods, the optimal mapping function f is equivalent to a projection matrix U. The notations used in this paper are summarized in Table 1.

2.1. MMC

Maximum margin criterion (MMC) [11] aims to maximize the margin between different classes in low dimensional space, as shown in Fig. 2. Extensive experimental results [11] have shown that its discriminative power was better than the power of PCA and LDA. Furthermore, MMC cannot suffer from the small sample size problem, which is known to cause serious stability problems for LDA. In [11], the objective function of MMC is summarized as:

$$J = \max \frac{1}{2} \sum_{i=1}^{C} \sum_{j=1}^{C} p_i p_j d(C_i, C_j) \tag{1}$$

where p_i and p_j are the prior probability of class i and class j, respectively, and C_i and C_j are all samples of class i and class j, respectively; $d(C_i, C_j)$ is defined as the distance between class i and class j.

$$d(C_i, C_j) = d(m_i, m_j) - \phi(i) - \phi(j) \tag{2}$$

Table 2
The main procedure of DOEPP.

Input: Data set $X = \{x_i \mid i = 1, 2, \ldots, n\}$; Balance parameter γ; Neighborhood size k.
Output: The final optimal projection matrix U.

Main steps:
1. Use PCA to preprocess the training samples and to eliminate the useless information of the training
data X. We still denote the projected training data in the subspace by X and the PCA projection matrix by U_{PCA}.
2. Compute W_g and W_l by Eq. (8) after constructing the graph models as Fig. 3.
3. Build the objective functions based on the above graph models described in Eq. (11).
4. Solve the standard eigenvalue problem (12) to obtain the optimal projection matrix $U_{DOEPP} = [u_1, u_2 \cdots u_d]$, whose vectors are the eigenvectors corresponding to the d smallest
eigenvalues. The final optimal projection matrix is $U = U_{PCA} U_{DOEPP}$ and the low dimensional embedding $Y = U^T X$.

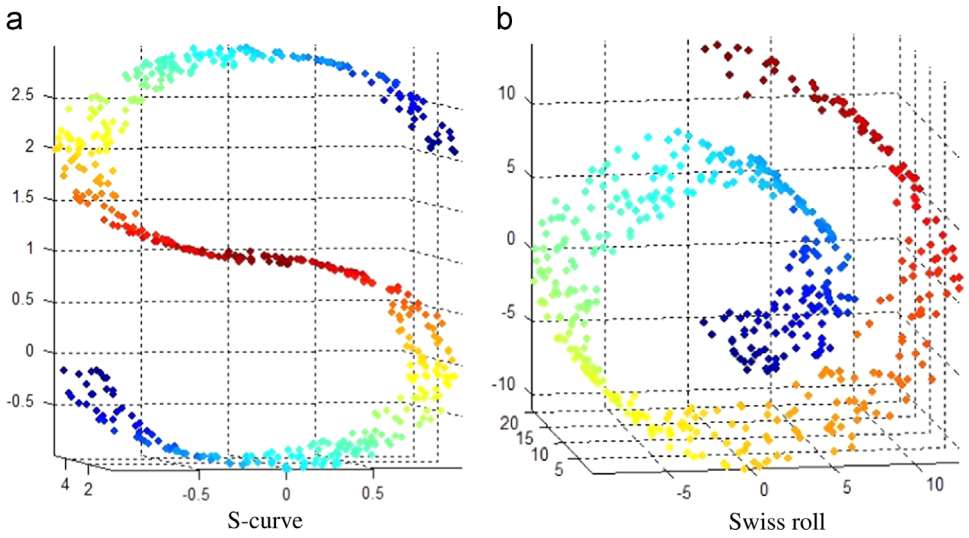

Fig. 5. 500 data points randomly sampled on two well-known synthetic 3D manifolds: (a) S-curve and (b) Swiss roll.

where m_i and m_j are the mean vector of class i and class j, respectively, and $\phi(i)$ and $\phi(j)$ are the covariance matrix of class i and class j, respectively. Here $\phi(i)=\mathrm{tr}(S_i), i=1\ldots C$. Maximum margin in the low dimensional feature space can be transformed as the following matrix form equation:

$$J=\max \mathrm{tr}(S_B^Y-S_W^Y) \tag{3}$$

Furthermore, MMC added the orthogonal constraint into the objective function to reconstruct the data. Thus the objective function of MMC can be deformed as follows:

$$\begin{cases} J=\max \mathrm{tr}(S_B^Y-S_W^Y)=\max \mathrm{tr}\left[U^T(S_B-S_W)U\right] \\ \text{s.t.} \quad U^TU=I \end{cases} \tag{4}$$

Obviously, the solutions of Eq. (4) are obtained by solving a standard eigendecomposition problem, that is $(S_B-S_W)U=\lambda U$. Finally, the optimal projection matrix U is given by $U=[u_1,u_2\ldots u_d]$, where $u_1,u_2\ldots u_d$ are the column vectors ordered according to eigenvalues $\lambda_1<\lambda_2<\cdots<\lambda_d$.

2.2. EPP

Elastic preserving projections (EPP) [31] incorporates the advantages of both the local geometry and global information of the training set. Based on the unified graph embedding framework [34], we analyze elastic preserving projections or elasticfaces. The procedure of EPP is decomposed into three steps as follows:

1. *Graph construction*: EPP [31] constructs two graph models: the undirected neighborhood graph G_l and global graph G_g. The k-nearest neighbor method is adopted to construct the local graph G_l. The graph models are described as $G_g=(X,E_g,W_g)$ and $G_l=(X,E_l,W_l)$, where the edge sets of the two graph models are $E_g=\{(x_i,x_j)|\ \forall i,j=1,\ldots N, i\neq j\}$ and $E_l=\{(x_i,x_j)|\ x_i\in N_e(x_j), x_j\in N_e(x_i)\}$, respectively. The corresponding weights of G_l and G_g are defined as $W_{lij}=\exp\left(-\frac{\|x_i-x_j\|^2}{\sigma^2}\right)$ when $x_j\in N_e(x_i)$ and $W_{gij}=\|x_i-x_j\|^2\exp\left(-\frac{\|x_i-x_j\|^2}{2t^2}\right)$, respectively. Note that σ^2 and t are two suitable kernel parameters.
2. *Targets on the graphs*: After the linear transformation, EPP does not only utilize the global information to discover the Euclidean structure of the high dimensional space, but also exploit the local geometrical structure to seek the nonlinear sub-manifold hidden in the high dimensional space.
3. *Optimization*: The targets of EPP are converted into the following objective functions:

$$\begin{cases} \arg\min\limits_U \sum\limits_{i,j=1}^{n} W_{lij}\|U^Tx_i-U^Tx_j\|^2=\mathrm{tr}[U^TXL_lX^TU] \\ \arg\max\limits_U \sum\limits_{i,j=1}^{n} W_{gij}\|U^Tx_i-U^Tx_j\|^2=\mathrm{tr}[U^TXL_gX^TU] \end{cases} \tag{5}$$

where $L_l=D_l-W_l$ and $L_g=D_g-W_g$, D_g(or D_l) is a diagonal matrix and $D_{gii}=\sum_j W_{gij}$.

To preserve the elasticity of the training set, Zang et al. converted the above two constraints into a constraint which simultaneously exploits the local and global information of the samples. Finally, the goal of EPP was to find the optimal transformation matrix U which maximizes the following optimization problem:

$$\begin{cases} \arg\max\limits_U \mathrm{tr}[U^TX[(1-\alpha)W_l+\alpha L_g]X^TU] \\ \text{s.t.} \quad \mathrm{tr}[U^TX(D_l-D_g)X^TU]=const \end{cases} \tag{6}$$

where $\alpha\in[0,1]$ is a trade-off parameter. Finally, Eq. (6) can be converted into Eq. (7) generalized eigenvalues problem by Lagrangian multiplier method:

$$X[(1-\alpha)W_l+\alpha L_g]X^TU=\lambda X(D_l-D_g)X^TU \tag{7}$$

Although the experimental results show that EPP performed better than the traditional manifold learning methods, it is an unsupervised method which cannot utilize the discriminant information of labeled samples. Therefore EPP did not perform well for the classification tasks.

3. Discriminative orthogonal elastic preserving projections

In many real applications, such as face recognition and image retrieval, many sample points are labeled or annotated by human or systems. However most of traditional manifold learning algorithms are only to find the low dimensional representation Y by preserving the intrinsic structure of high dimensional data X and cannot use the discriminative information directly and explicitly. Therefore, we propose a novel manifold learning method named discriminative orthogonal elastic persevering projections (DOEPP) to utilize all of the class label information and improve the performance of classification tasks.

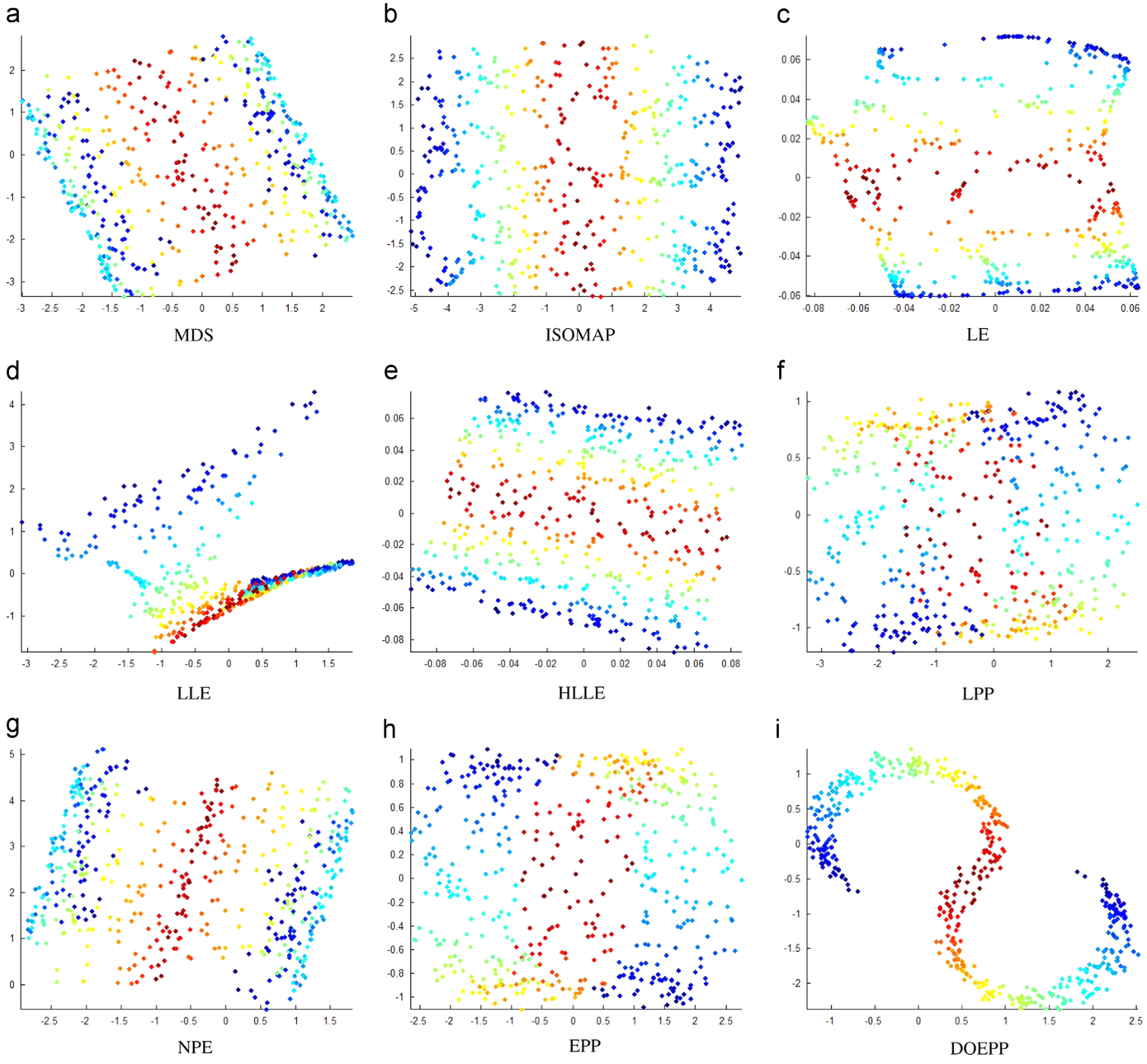

Fig. 6. Projection results of nine manifold learning methods on S-curve data samples.

The main idea of DOEPP is to incorporate the merits of manifold learning and the advantages of supervised learning. Different from the traditional manifold learning methods, like PCA, LPP, NPE and EPP, our algorithm is not only able to preserve the elasticity of the training set, but can effectively enhance the discriminative performance of manifold learning algorithms by extracting label information from the annotated samples. Therefore, to preserve the elasticity of the training set X, DOEPP construct two graph models: global graph G_g and local neighborhood graph G_l. Similar to PCA, DOEPP builds the global graph G_g as shown Fig. 3(a). The traditional manifold methods, like LPP, NPE and EPP, often ignore the labels of the training points and only preserve the local geometrical structure by the graph Laplacian. Different from the traditional methods, DOEPP constructs local prior graph G_l by using labels of training samples as prior knowledge, to make distances of samples in a same class of Y as small as possible, while distances between different classes are as large as possible, as shown in Fig. 3(b). The weights of each edge in G_g and G_l are calculated by Eq. (8). The visualization results of G_g and G_l on ORL and Yale are presented in Fig. 4(a) and (c), respectively. Fig. 4(c) shows that DOEPP can decrease the coefficients or weights of neighborhoods in the different classes and will have better performance than EPP:

$$\begin{cases} W_{lij} = \begin{cases} \exp\left(-\dfrac{\|x_i - x_j\|^2}{2t^2}\right), & l_j = l_i \\ 0, \text{ otherwise} \end{cases} \\ W_{gij} = \begin{cases} \|x_i - x_j\|^2 \exp\left(-\dfrac{\|x_i - x_j\|^2}{2t^2}\right), & i \neq j \\ 0, \quad i = j \end{cases} \end{cases} \tag{8}$$

Therefore, the objective function of the elasticity of the training set X can be represented by two parts: (1) preserving the hidden sub-manifold structure and (2) discovering global Euclidean

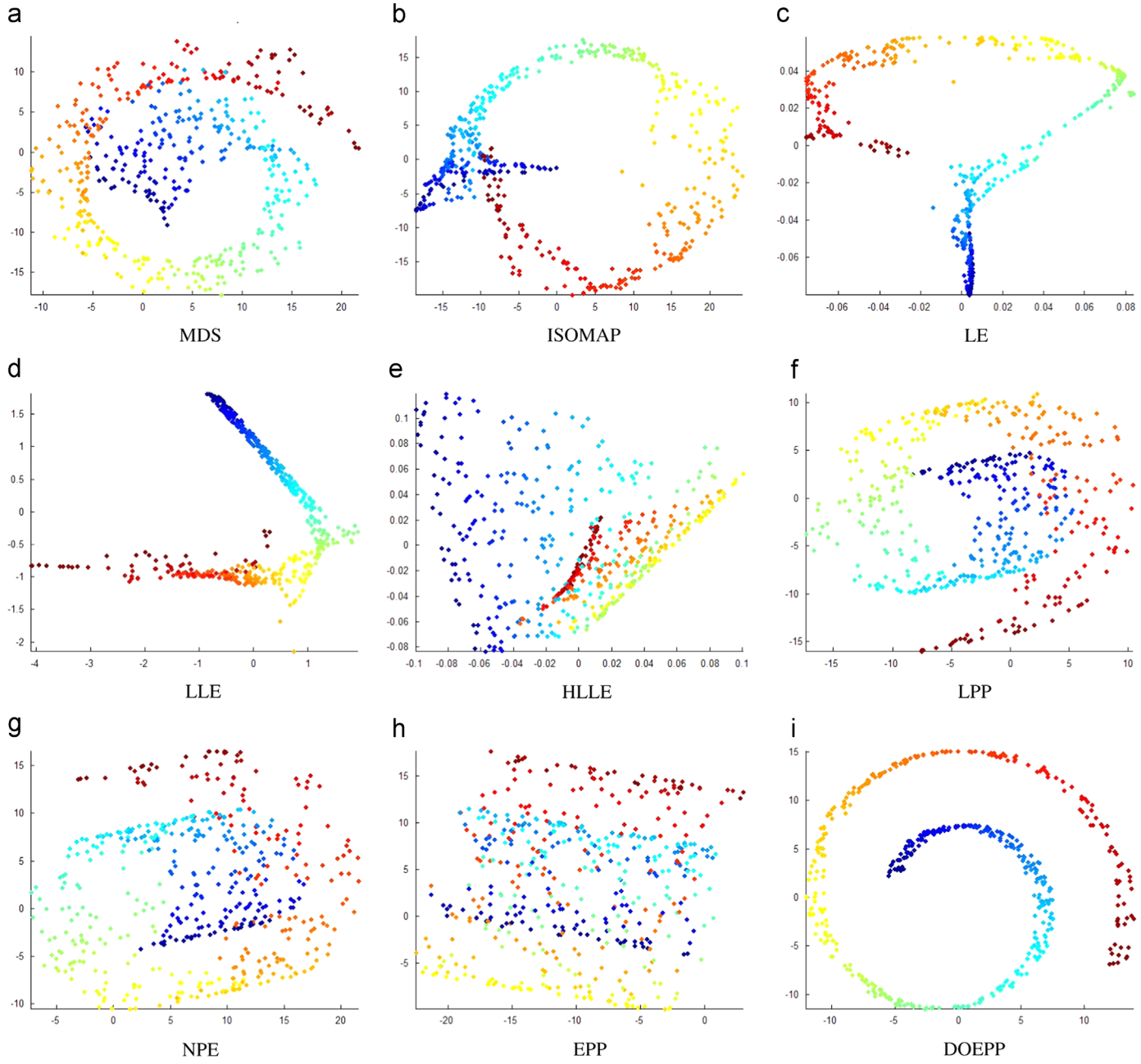

Fig. 7. Projection results of nine manifold learning methods on Swiss Roll data samples.

information. In other words,the objective function of the elasticity is formulated as the following two optimization problems:

$$\begin{cases} \min\limits_{U} \sum\limits_{i,j=1}^{n} W_{lij} \| U^T x_i - U^T x_j \|^2 = \mathrm{tr}[U^T X L_l X^T U] \\ \max\limits_{U} \sum\limits_{i,j=1}^{n} W_{gij} \| U^T x_i - U^T x_j \|^2 = \mathrm{tr}[U^T X L_g X^T U] \end{cases} \tag{9}$$

where the weighted matrices W_g and W_l are calculated by Eq. (8), $L_l = D_l - W_l$ and , $L_g = D_g - W_g$. D_g and D_l are diagonal, $D_{lii} = \sum_{j=1}^{N} W_{lij}$ and $D_{gii} = \sum_{j=1}^{N} W_{gij}$.

EPP only uses k-nearest neighborhoods (kNNs) to build the local graph G_l and does not use the label information. Therefore, it is difficult for EPP to avoid the effect of noise in the training data, that is the k-nearest neighborhoods set of the current sample often consists of different class samples, as shown in Fig. 4(b). To solve this problem, DOEPP does not only add label information into constructing the local graph G_l, but also impose maximum margin constraint between different classes into our objective function. Similar to MMC, the maximum margin constraint can be formulated as

$$J = \max \mathrm{tr}(S_B^Y - S_W^Y) = \max \mathrm{tr}\left[U^T (S_B - S_W) U\right] \tag{10}$$

To further enhance the discriminative power and eliminate the redundant information of a given data set X, DOEPP also imposes the orthogonal constraint of projection matrix U, that is $U^T U = I$. Finally, to preserve the elasticity and discriminant information of the training set, the multi-objective functions of DOEPP can be redefined as the optimization problem:

$$\begin{cases} \max\limits_{U} \mathrm{tr}\left\{U^T [X[(1-\alpha)W^l + \alpha L^g]X^T + \beta(S_B - S_W)]U\right\} \\ \text{s.t.} \quad U^T U = I \end{cases} \tag{11}$$

where α and β are two control parameters. $\alpha \in [0, 1]$ represents the

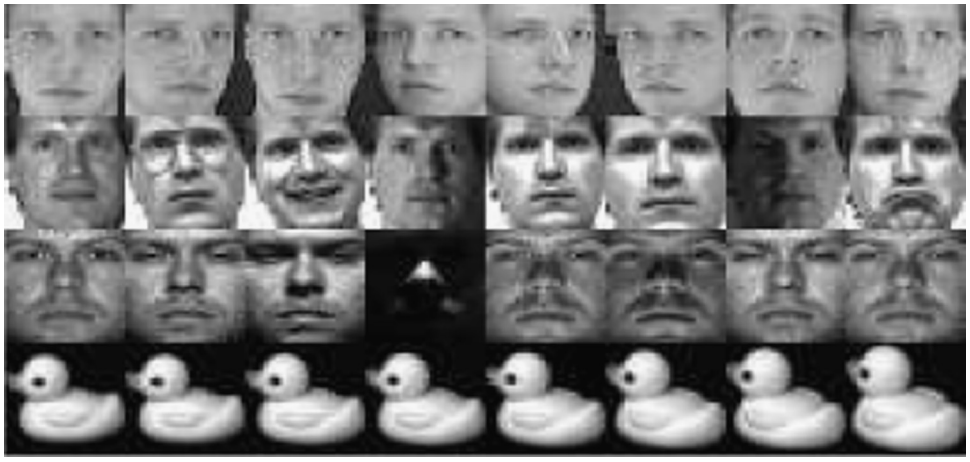

Fig. 8. Sample face images from the four databases. The first row comes from ORL; the second row comes from Yale; the third row comes from Extended YaleB; and the fourth row comes from COIL-20.

Table 3
Average recognition rates (%) ± standard deviation of eight algorithms on ORL.

Mode	MMC	LPP	NPE	DSNPE	EPP	MFA	DLA	DOEPP
2 Train	69.06 ± 3.66	61.08 ± 3.90	60.69 ± 4.03	70.81 ± 3.05	65.20 ± 2.68	76.30 ± 3.49	80.22 ± 3.62	78.45 ± 2.64
3 Train	77.85 ± 2.38	70.01 ± 2.32	71.57 ± 2.47	76.94 ± 2.77	69.72 ± 2.32	86.61 ± 2.17	89.47 ± 1.94	85.88 ± 2.90
4 Train	83.8 ± 1.91	76.13 ± 2.61	78.20 ± 3.22	80.56 ± 2.51	72.41 ± 1.61	91.29 ± 1.94	93.00 ± 1.50	90.49 ± 1.46
5 Train	88.04 ± 2.52	80.07 ± 2.63	82.58 ± 2.53	82.86 ± 2.50	74.03 ± 1.84	93.95 ± 1.56	95.28 ± 1.56	93.44 ± 1.48
6 Train	80.42 ± 3.31	82.65 ± 2.59	85.68 ± 2.52	83.38 ± 3.02	75.3 ± 1.90	95.16 ± 1.57	96.48 ± 1.54	95.15 ± 1.65
7 Train	73.65 ± 4.59	85.7 ± 3.19	88.9 ± 2.96	84.91 ± 3.63	75.65 ± 2.77	96.51 ± 1.80	97.3 ± 1.50	96.4 ± 1.69
8 Train	63.45 ± 4.60	87.93 ± 2.96	90.92 ± 2.73	85.67 ± 3.50	77.4 ± 3.17	97.17 ± 1.85	98.3 ± 1.49	98.45 ± 1.74

weight of global structural information, while $\beta > 0$ indicates the importance of discriminant information for the current task.

To find the optimal transform matrix U, we can use the Lagrange multipliers to transform Eq. (11) into the following eigenvalues problem:

$$[X[(1-\alpha)W^l + \alpha L^g]X^T + \beta(S_B - S_W)]U = \lambda U \tag{12}$$

According to the basic algebra and matrix knowledge, the optimal transformation matrix U is constructed by the eigenvectors $u_1, u_2, \ldots, u_d$ of $X[(1-\alpha)W^l + \alpha L^g]X^T + \beta(S_B - S_W)$ associated with its d smallest eigenvalues that maximize the objective function.

According to the unified framework, the main procedure of DOEPP is listed in Table 2. To some extent, the formulation of our proposed method is a weighted least square regression problem. In other words, it is convex. Therefore, it has a closed and global optimal solution. Compared with other related approaches, the computational complexity of our algorithm is not so high. In fact, the most computational step of DOEPP is to solve the problem in Eqs. (8) and (12). Their computational complexities are $O(DNk^3 + DN^2)$ and $O(dqN^2)$. Thus, the computational complexity of DOEPP is $\max\{O(DNk^3 + DN^2), O(dqN^2)\}$.

4. Experimental results

In this section, several experiments were carried out to examine the effectiveness of our proposed method for classification and dimensionality reduction of synthetic data.

4.1. Synthetic data

In this subsection, we evaluate the performance of DOEPP by comparing against five non-linear and three linear representative dimensionality reduction methods: MDS, ISOMAP, LE, LLE, HLLE, LPP, NPE and EPP on two well-known synthetic data sets from [35]: The S-curve and the Swiss roll, for manifold structure learning. We use an implementation of MDS, ISOMAP, LE, LLE, HLLE, LPP and NPE, which is publicly available.[1] We implemented EPP in MATLAB.

We sample 500 three dimensional data points from the S-curve and Swiss roll manifolds randomly and add Gaussian noise into samples as the training set. The sampled points are shown in Fig. 5. And then we compute the low dimensional embedding of the above nine methods. Experimental results on S-curve and Swiss roll are shown in Figs. 6 and 7, respectively. Fig. 6 illustrates that the performance of DOEPP parallels that of ISOMAP, LE and HLLE and outperforms the other five methods consistently for S-curve data. Fig. 7 shows that the performance of DOEPP parallels that of LE and outperforms that of the other seven methods for Swiss roll data. But LE, ISOMAP and HLLE are the non-linear methods, they cannot process the new data points directly. Therefore, DOEPP outperforms other methods in most of the cases, that is DOEPP can get more faithful projections that others. Because compared with the traditional manifold learning methods, DOEPP does not only preserve global and local geometric structural information of the high dimensional manifolds, but also can maintain the discriminative and orthogonal information.

4.2. Classification on real datasets

In this subsection, we evaluate DOEPP by comparing to the following seven methods on four standard image databases, ORL, Yale, Extended YaleB and COIL-20: (1) MMC [11], (2) LPP [26], (3) NPE [27], (4) EPP [31], (5) Discriminant sparse neighborhood

[1] http://www.cad.zju.edu.cn/home/dengcai/Data/DimensionReduction.html

Table 4
Student *t*-test results between DOEPP and other approaches for the results on ORL. W means DOEPP performs better. F means other method performs better. B means that DOEPP and other methods cannot outperform each other. The value in the bracket is the associated *p*-value. The statistical significance of *t*-test is 5%.

Dataset	Method	2 *Train*	3 *Train*	4 *Train*	5 *Train*	6 *Train*	7 *Train*	8 *Train*
ORL	MMC	W(.00)	W(.00)	W(.00)	W(.00)	W(.01)	W(.01)	W(.01)
	LPP	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)
	NPE	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)
	DSNPE	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)
	EPP	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)
	MFA	W(.00)	W(.01)	W(.02)	W(.01)	W(.03)	W(.02)	W(.02)
	DLA	W(.01)	B(–)	W(.03)	F(.03)	B(–)	B(–)	W(.08)

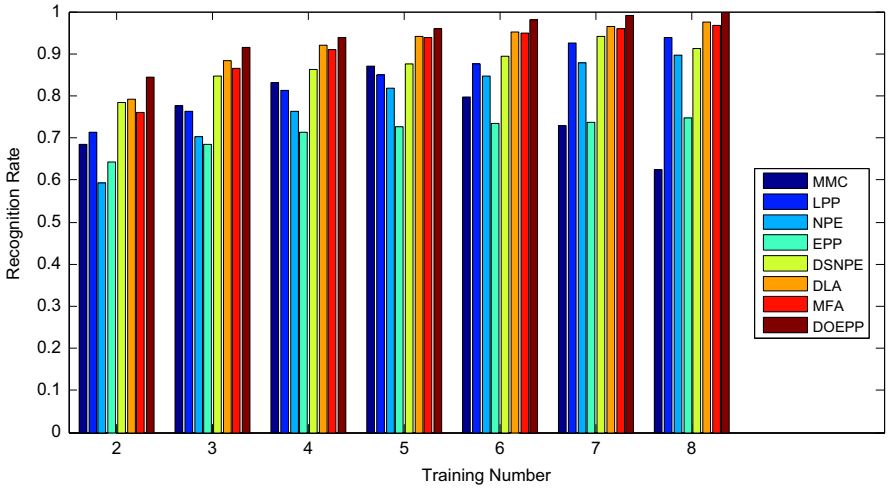

Fig. 9. Best recognition rate versus training numbers with different methods on ORL.

preserving embedding (DSNPE, [36]), (6) MFA [32], (7) DLA [33]. The sample images from these four datasets are shown in Fig. 8.

Among these algorithms, LPP, NPE and EPP are unsupervised algorithms which do not use the class label information, while MMC, DSNPE, MFA and DLA are supervised. DLA and MFA are recently proposed manifold learning algorithms. All images from four databases are cropped and normalized to the 32×32 pixel arrays with 256 gray levels per pixel. Each image was reshaped to one long vector by arranging its pixel values in a fixed order. And all datasets were randomly divided into two separate sets: training set and testing set. Training set was used to learn the low-dimensional subspace along with the projection matrix. Testing set was used to report the final recognition accuracy. The source codes for LPP, NPE[2] and DLA[3] are available online . We implemented other algorithms in MATLAB. Different algorithms follow an equivalent procedure for all experiments on various datasets. The procedures for the classification problems are separated into three steps:

- Each algorithm is applied to training samples to learn the projection matrix *U*.
- Each testing sample is projected onto a low-dimensional space via *U*.
- The testing samples in the projected subspace are identified by the NN classifier (NNC).

4.2.1. ORL

ORL[4] (formerly Olivetti Research Ltd.) contains 40 individuals and 10 different images for each individual, including variations in the lighting, facial expression (smiling/not smiling) and pose. The images were taken with a tolerance for some tilting and rotation of the face up to 20. For each of the 40 peoples, we randomly selected a different number of samples per person for training (2, 3, 4, 5, 6, 7, or 8), and the remainders were used for testing. All the tests were repeated 50 times, and we then calculate the best recognition rate and the recognition results under different reduced dimensions.

Fig. 9 gives comparisons of the best recognition rates of eight algorithms under a different training number and Fig. 9 presents the recognition rates of 8 algorithms under different reduced dimensions. The mean and standard derivation (std) values of classification results are shown in Table 3. Besides, we compare our method with other approaches by Student's *t*-test. The statistical significance with a threshold of 0.05 is listed in Table 4. The smaller *p*-value means the higher assurance of the conclusion.

As seen from Figs. 9, 10, Tables 3 and 4, from the statistical view, we can see that DOEPP achieves significantly better results compared to the other algorithms in most cases. Besides, the experimental results also show consistently that DOEPP can extract features very efficiently.

4.2.2. Yale

Yale[5] contains 165 images of 15 persons. Lighting conditions, gender, facial expressions and configurations are different among these images. For training, we randomly selected a different number of samples per person for training (2, 3, 4, 5, 6, 7, or 8), and use the remainders for testing. All the tests were repeated 50 times, and we then calculate the best recognition rate and the recognition results under different reduced dimensions. Figs. 11

[2] http://www.cad.zju.edu.cn/home/dengcai/Data/DimensionReduction.html
[3] http://www.rad.upenn.edu/sbia/Tianhao.Zhang/DLA.m
[4] http://www.uk.research.att.com/facedatabase.html
[5] http://cvc.yale.edu/projects/yalefaces/yalefaces.html

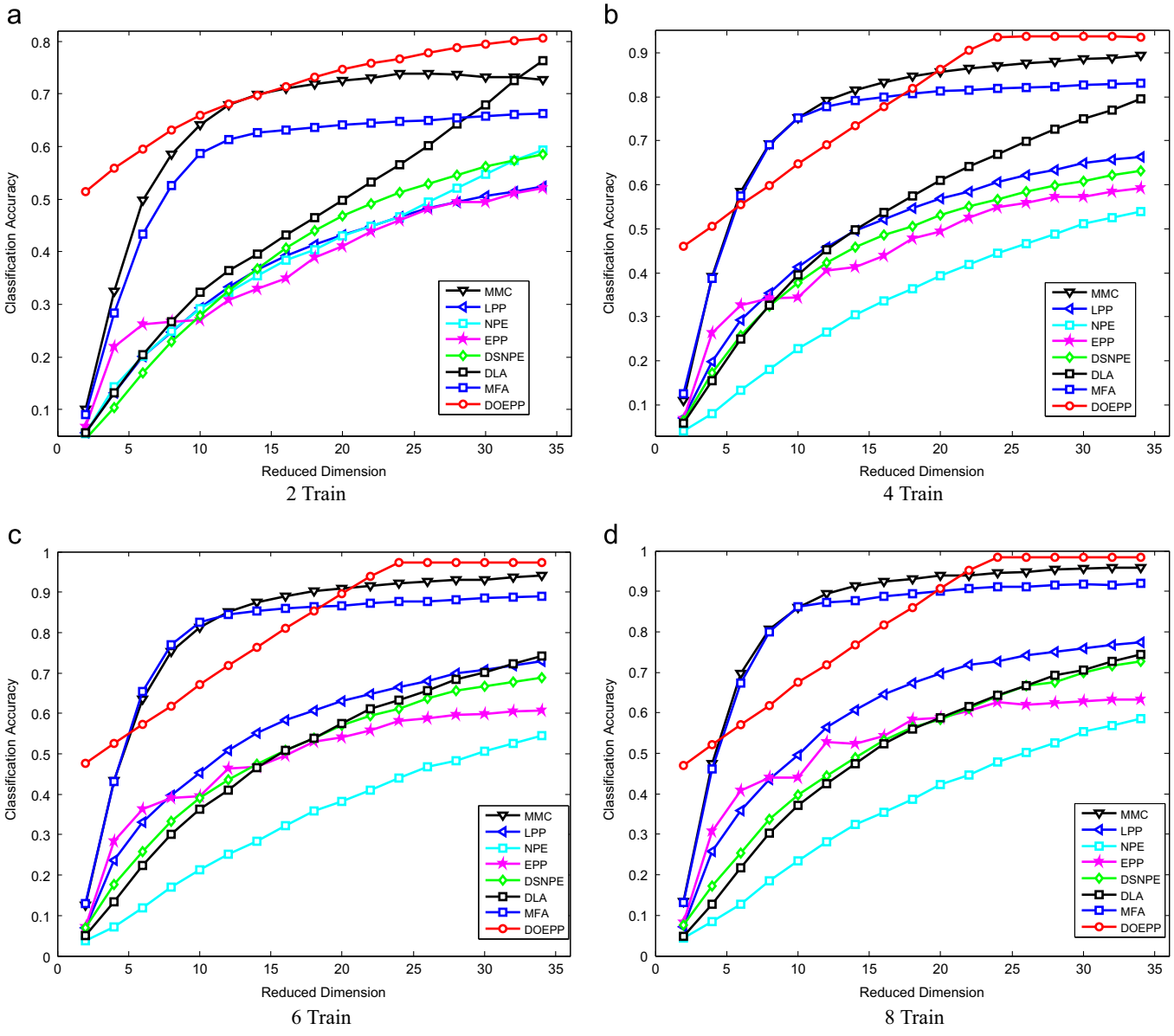

Fig. 10. Recognition rate versus subspace dimension with a different number of training and different methods on ORL.

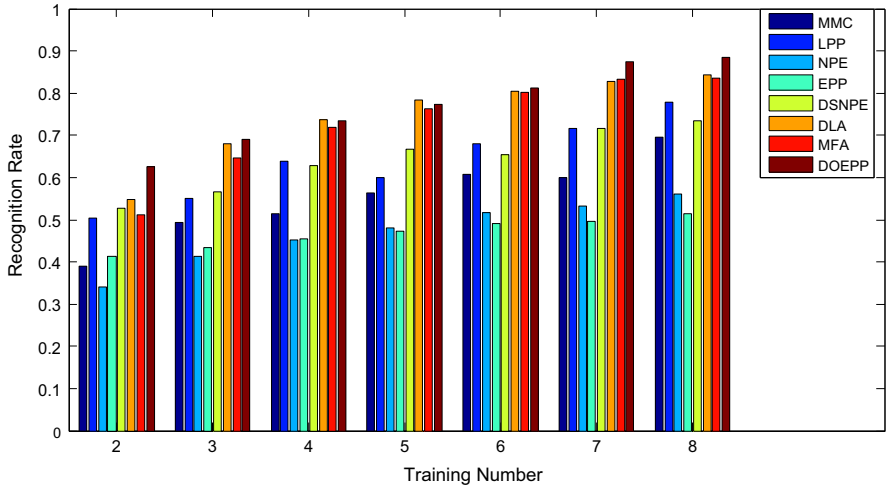

Fig. 11. Best recognition rate versus training numbers with different methods on Yale.

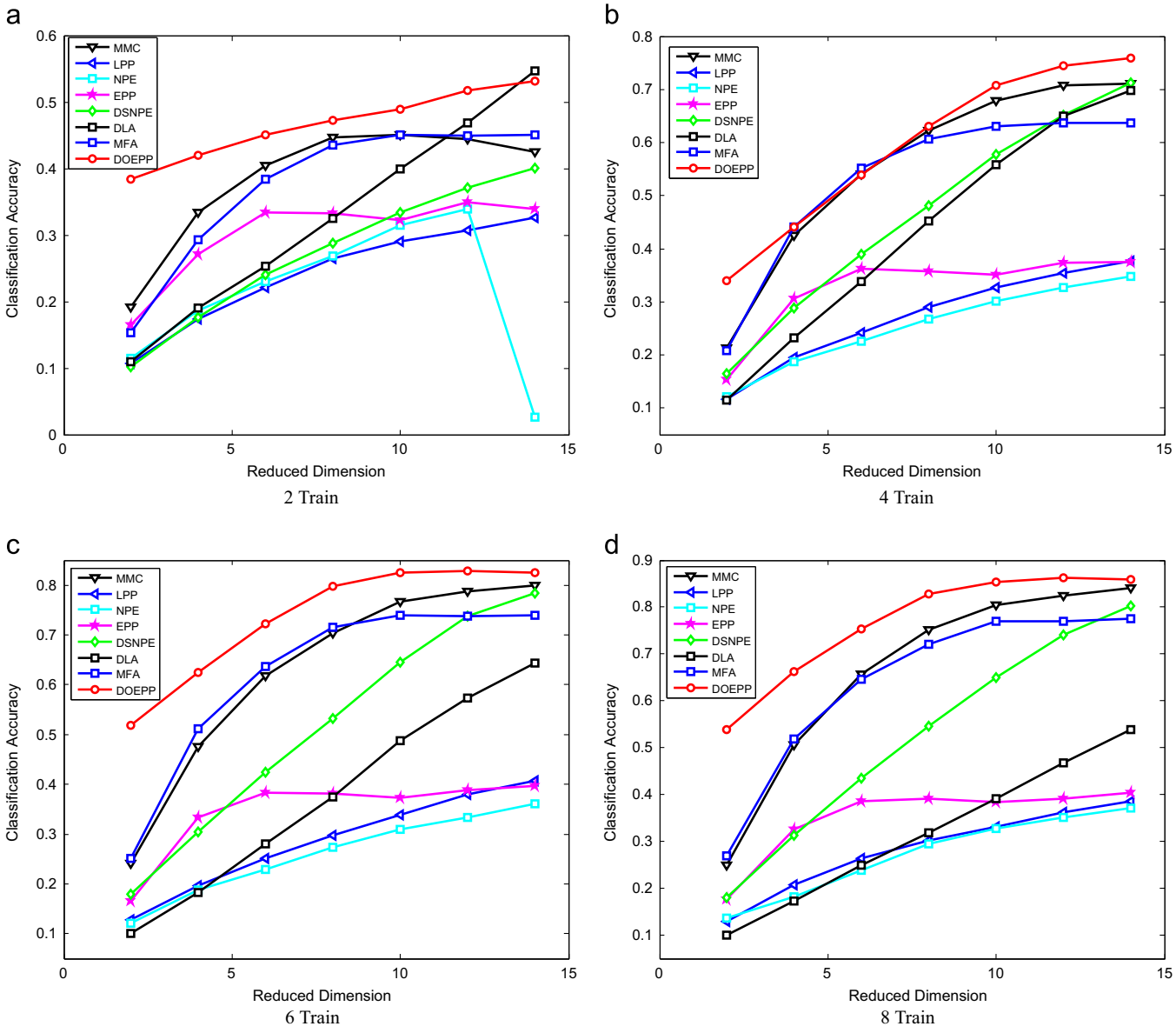

Fig. 12. Recognition rate versus subspace dimension with a different number of training and different methods on Yale.

Table 5
Average recognition rates (%) ± standard deviation of eight algorithms on Yale.

Mode	MMC	LPP	NPE	DSNPE	EPP	MFA	DLA	DOEPP
2 Train	40.69 ± 3.94	42.01 ± 4.09	35.15 ± 4.88	46.19 ± 3.70	43.02 ± 3.42	51.748 ± 4.87	55.92 ± 4.45	59.25 ± 4.10
3 Train	49.96 ± 3.81	48.31 ± 3.44	42.41 ± 4.46	50.8 ± 3.27	45.61 ± 2.64	65.25 ± 4.51	69.81 ± 3.66	67.67 ± 3.37
4 Train	53.69 ± 5.38	51.65 ± 4.42	46.70 ± 4.96	53.27 ± 4.66	47.65 ± 2.82	72.20 ± 4.28	76.28 ± 3.96	70.99 ± 3.16
5 Train	58.04 ± 4.20	54.17 ± 3.56	50.08 ± 3.85	55.86 ± 4.52	49.24 ± 3.11	77.04 ± 4.90	80.95 ± 3.73	75.42 ± 3.65
6 Train	63.17 ± 5.46	57.49 ± 4.60	53.36 ± 4.36	57.09 ± 3.68	50.96 ± 3.83	80.77 ± 4.37	83.62 ± 4.25	78.50 ± 4.44
7 Train	64.50 ± 3.84	59.80 ± 5.98	55.93 ± 4.32	61.00 ± 4.64	52.00 ± 3.48	84.10 ± 3.55	86.33 ± 3.73	81.630 ± 4.73
8 Train	70.84 ± 4.84	62.57 ± 6.07	59.46 ± 5.81	61.02 ± 6.02	54.40 ± 4.40	84.57 ± 4.82	88.93 ± 4.02	84.53 ± 4.42

and 12 give the results for the eight algorithms on Yale. The mean and standard derivation (std) values of classification results are shown in Table 3. Besides, we compare our method with other approaches by Student's *t*-test. The statistical significance with a threshold of 0.05 is listed in Table 4. The smaller *p*-value means the higher assurance of the conclusion.

Fig. 11 presents that the best recognition rate of DOEPP outperforms the other algorithms on training numbers 2, 3, 6, 7 and 8, while the performance of DOEPP is inferior to DLA on training numbers 4 and 5. In addition, it is shown that the recognition rate of DOEPP outperforms the other algorithms obviously with the increase of training number in Fig. 12. As seen from Figs. 11 and 12,

Table 6
Student *t*-test results between DOEPP and other approaches for the results on Yale. W means DOEPP performs better. F means other method performs better. B means that DOEPP and other methods cannot outperform each other. The value in the bracket is the associated *p*-value. The statistical significance of *t*-test is 5%.

Dataset	Method	2 *Train*	3 *Train*	4 *Train*	5 *Train*	6 *Train*	7 *Train*	8 *Train*
Yale	MMC	W(.00)	W(.01)	W(.01)	W(.01)	W(.01)	W(.01)	W(.01)
	LPP	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)
	NPE	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)
	DSNPE	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.01)
	EPP	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)	W(.00)
	MFA	W(.01)	W(.01)	W(.01)	W(.01)	W(.02)	W(.05)	W(.06)
	DLA	W(.03)	B(−)	B(−)	F(.05)	B(−)	W(.05)	B(−)

Table 7
Average recognition rates (%) ± standard deviation of eight algorithms on YaleB.

Mode	**MMC**	**LPP**	**NPE**	**DSNPE**	**EPP**	**MFA**	**DLA**	**DOEPP**
2 Train	51.10 ± 4.27	49.13 ± 3.83	34.28 ± 4.85	49.73 ± 0.07	43.72 ± 4.37	50.30 ± 4.43	54.32 ± 4.03	60.46 ± 4.77
3 Train	62.46 ± 3.81	59.5 ± 4.23	45.71 ± 4.58	55.71 ± 0.65	50.89 ± 3.49	60.43 ± 3.55	66.28 ± 3.45	71.45 ± 4.66
4 Train	69.91 ± 2.66	65.96 ± 2.99	55.77 ± 3.18	60.16 ± 1.54	56.03 ± 2.93	67.63 ± 3.56	73.94 ± 2.55	78.06 ± 3.07
5 Train	76.01 ± 2.23	71.62 ± 2.50	63.19 ± 3.24	63.69 ± 0.99	59.76 ± 3.18	72.39 ± 2.94	78.86 ± 2.68	80.28 ± 3.16
6 Train	79.23 ± 2.13	74.83 ± 2.21	67.65 ± 2.78	65.68 ± 0.66	62.07 ± 2.22	76.09 ± 2.77	82.12 ± 2.16	84.35 ± 2.49
7 Train	80.83 ± 1.42	75.78 ± 1.80	71.21 ± 2.40	59.89 ± 2.77	64.87 ± 1.77	77.66 ± 2.40	84.11 ± 1.42	85.29 ± 2.45
8 Train	81.13 ± 2.03	77.28 ± 1.81	74.98 ± 2.16	65.74 ± 1.52	67.37 ± 2.08	81.11 ± 2.74	86.76 ± 1.69	83.38 ± 2.55

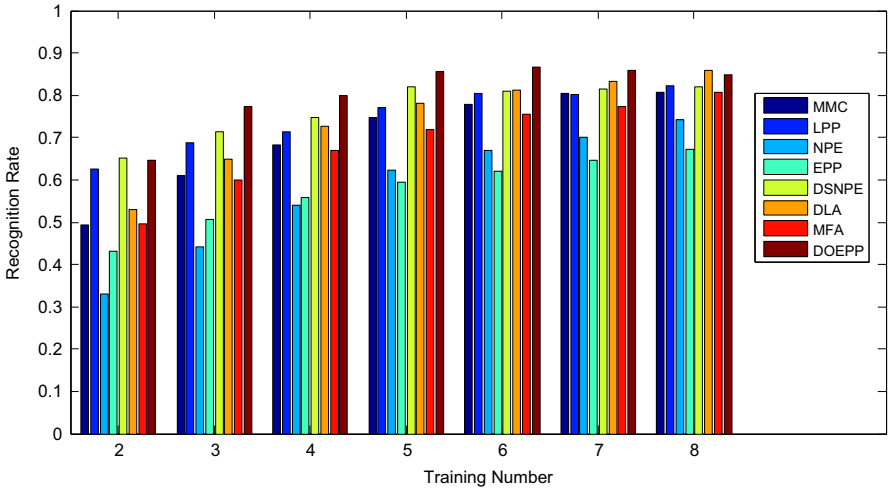

Fig. 13. Best recognition rate of eight methods versus training numbers on Extended YaleB.

Tables 5 and 6, from the statistical view, we can see that the performance of DOEPP is significantly better than other methods on Yale database in most cases.

4.2.3. Extended YaleB

Extended YaleB[6] contains 16,128 images of 38 humans under 9 poses and 64 illumination conditions. In our study, we randomly select 2414 images as a subset from Extended YaleB for performance evaluation. For training, we randomly selected different numbers (2, 3, 4, 5, 6, 7, or 8) of samples per person and use the remainders for testing. All the tests were repeated 50 times, and we then calculate the recognition rates under different reduced dimensions and the average recognition results, and the standard deviation. Figs. 13, 14 and Table 7 give the results for all of selected algorithms on Extended YaleB. Fig. 13 and Table 7 give the results for the eight algorithms on Extended YaleB. According to analyses of experimental results in Fig. 13 and Table 7, we observe that DOEPP outperforms the other methods except on training numbers 2 and 8. From Fig. 14, we have an interesting observation that the performance of DOEPP is much better than others in low dimensional embedding feature space.

4.2.4. COIL-20

COIL-20[7] contains 1440 images of 20 objects rotated on a turntable taken from a fixed camera, resulting in 72 images per object. For each of the 20 objects, we randomly select a different number (2, 3, 4, 5, 6, 7, or 8) of samples per person for training, and the remainders are for testing. All the tests were repeated 50 times, and we then calculate the best and average recognition rates with its corresponding standard deviation. The experimental results on COIL-20 are shown in Figs. 15, 16 and Table 8.

[6] http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html

[7] http://www.cs.columbia.edu/CAVE/coil-20.html

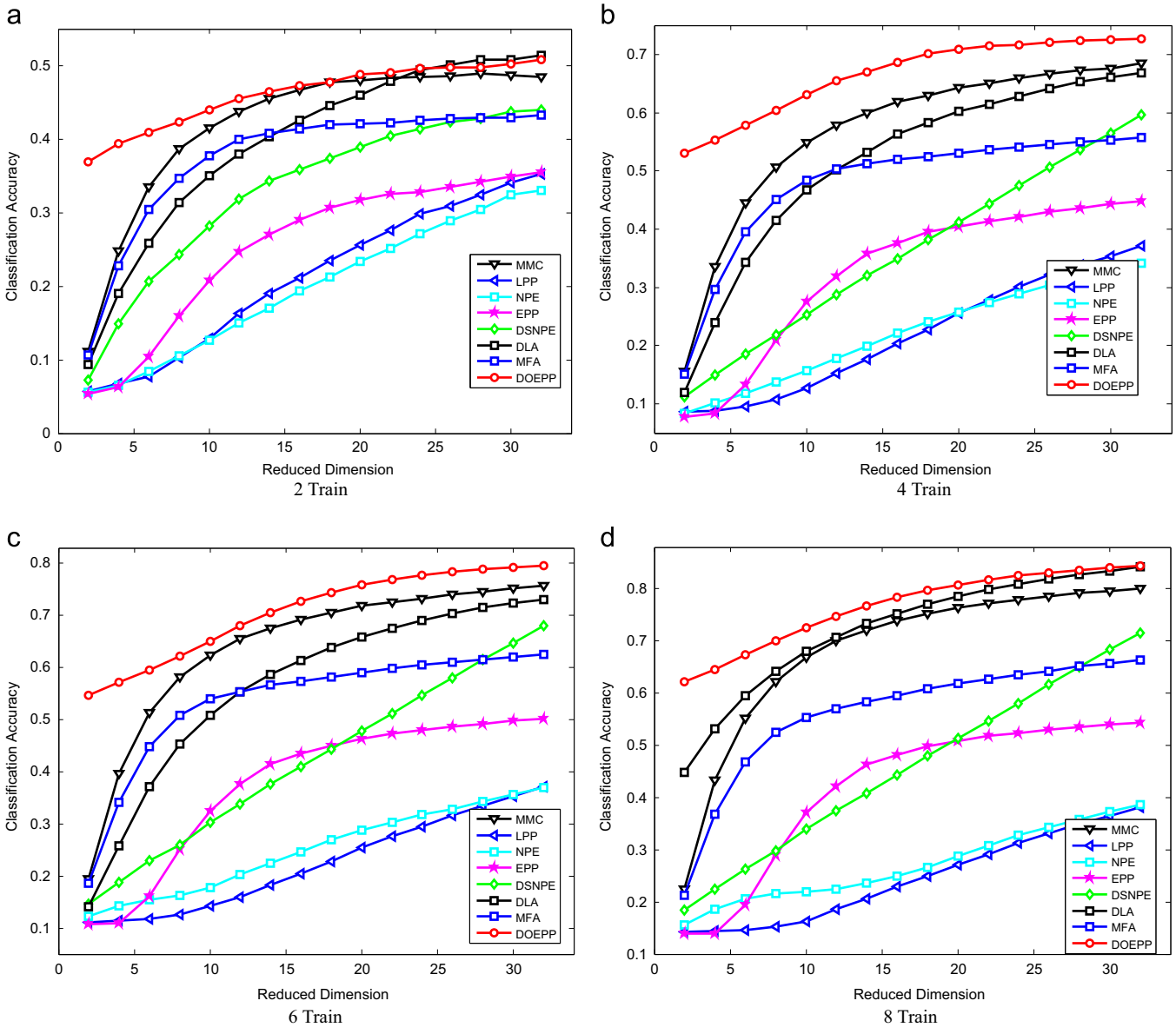

Fig. 14. Recognition rate of eight methods versus subspace dimension with a different number of training samples on Extended YaleB.

Table 8
Average recognition rates (%) ± standard deviation of eight algorithms on COIL-20.

Mode	MMC	LPP	NPE	DSNPE	EPP	MFA	DLA	DOEPP
2 Train	69.70 ± 6.30	68.00 ± 5.76	64.75 ± 5.72	66.75 ± 6.63	71.15 ± 6.31	71.60 ± 6.32	73.20 ± 4.42	71.55 ± 6.12
3 Train	77.00 ± 6.08	73.40 ± 5.44	69.93 ± 5.94	70.56 ± 6.44	77.23 ± 5.82	77.83 ± 5.26	79.66 ± 4.46	78.90 ± 5.74
4 Train	79.07 ± 4.08	74.40 ± 4.85	71.30 ± 4.49	72.17 ± 5.28	80.00 ± 3.91	79.30 ± 3.98	82 ± 3.66	80.65 ± 5.28
5 Train	83.54 ± 3.73	78.88 ± 4.26	74.48 ± 5.10	76.22 ± 4.27	82.70 ± 3.95	82.82 ± 4.82	85.62 ± 3.80	84.96 ± 4.81
6 Train	85.76 ± 3.01	81.18 ± 3.36	77.31 ± 3.85	78.11 ± 4.09	85.55 ± 2.95	84.98 ± 3.52	87.86 ± 2.95	86.67 ± 4.49
7 Train	87.00 ± 2.42	82.47 ± 2.68	78.68 ± 3.29	78.82 ± 3.68	86.52 ± 2.86	86.28 ± 2.81	88.68 ± 2.67	89.90 ± 2.52
8 Train	88.57 ± 2.56	83.75 ± 3.13	80.40 ± 2.68	79.52 ± 3.22	88.55 ± 2.35	86.97 ± 2.92	90.03 ± 2.46	91.34 ± 2.49

Fig. 16 gives comparisons of the recognition rates of eight algorithms under different reduced dimensions on COIL-20. Table 8 lists the average recognition rates and the corresponding standard deviation when the number of reduced features fixed to 200. From Figs. 15, 16 and Table 8, we observe that the performance of DOEPP parallels that of DLA, and outperform again the remaining methods. Note that DOEPP outperforms the other algorithms in the very low dimensional feature space, as shown in Fig. 16. Overall, DOEPP seems to be slightly better than DLA.

4.3. Parameter determination

For the classification problem, the two control parameters α and β will have an effect on the recognition rates (RR). Since parameter determination is still an open problem, we determine

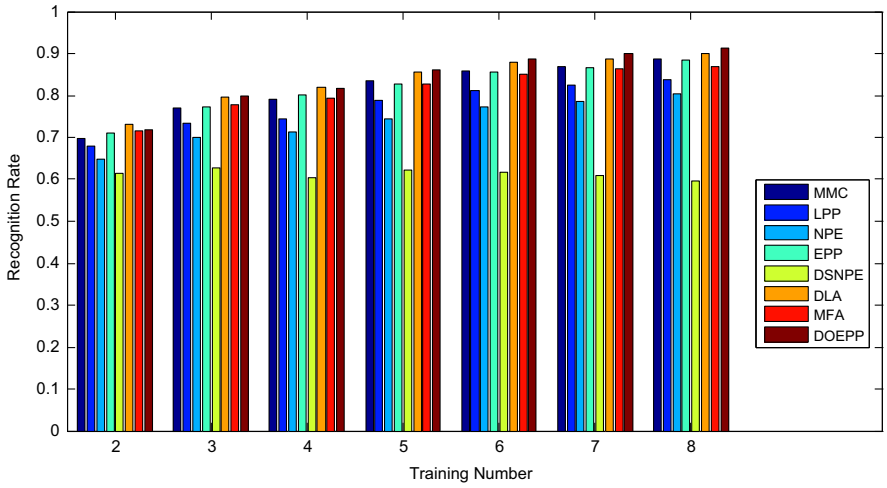

Fig. 15. Best recognition rate of eight methods versus training numbers on COIL-20.

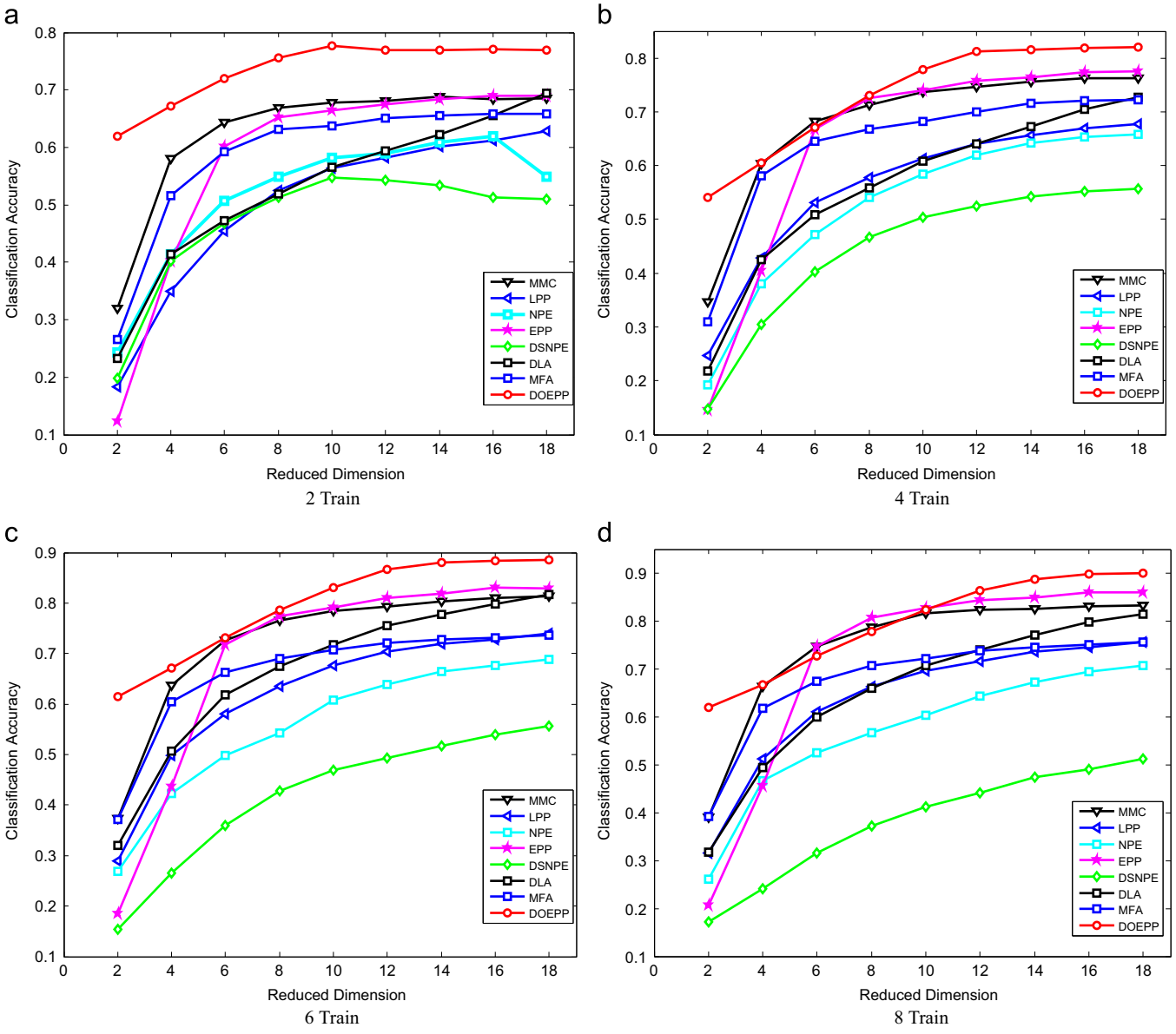

Fig. 16. Recognition rate of eight methods versus subspace dimension on COIL-20 with a different number of training samples.

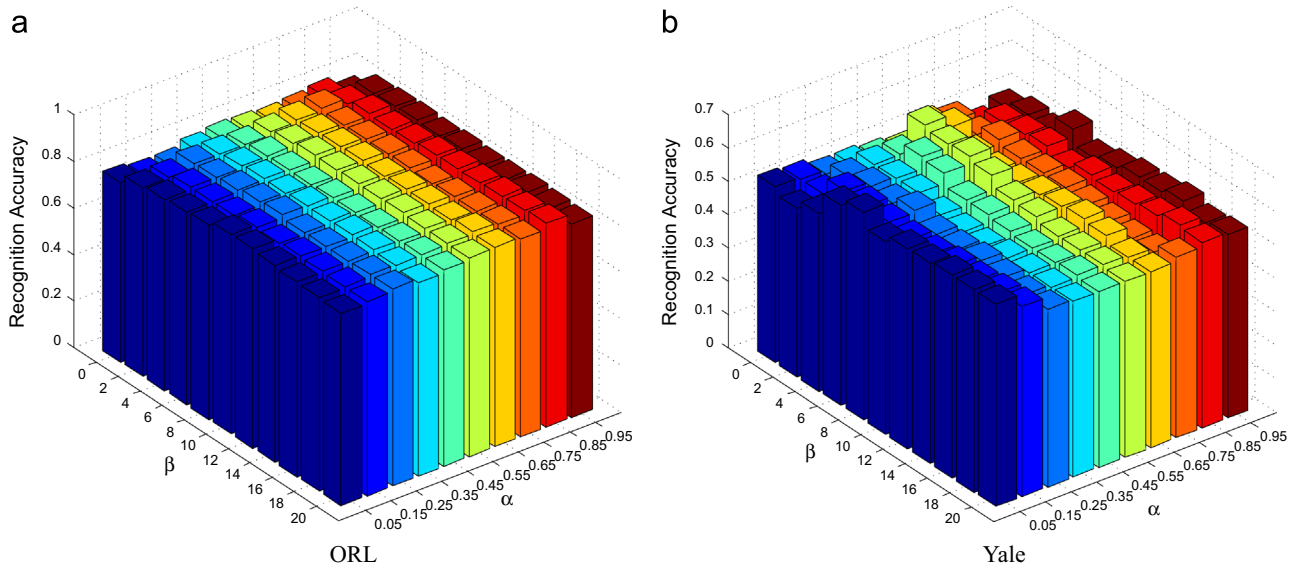

Fig. 17. The recognition rate of DOEPP versus the variation of α and β on ORL and Yale database.

the most important parameters, i.e., α and β is a heuristic way. More concretely, we determine two parameters by grid search at first and then change them within certain ranges. Therefore, we set $\alpha = [0, 0.05, 0.1, 0.15 \ldots 0.95, 1]$ and $\beta = [0, 2, 4 \ldots 18, 20]$ on ORL and Yale. The experimental results are shown in Fig. 17.

As seen from Fig. 17, parameter determination takes influence on the performance of DOEPP. Different combinations of parameters may result in different embeddings. Then, the recognition results of classification change.

5. Conclusion

To enhance the performance of manifold learning, we propose a new discriminative orthogonal elastic preserving projections (DOEPP) for classification. EPP can explore the local sub-manifold and global structure information simultaneously. We introduce the discriminant information of class labels and the orthogonal constraint into the objective function of DOEPP. Therefore, DOEPP does not only incorporate the advantages of both EPP and MMC, but also can improve the discriminative performance for classification tasks. Extensive experimental results on real datasets show that DOEPP has more discriminative power than others. In our future work, we will extend DOEPP to nonlinear form by kernel trick and tensor analysis, which can process the tensor data directly.

Acknowledgements

This research was supported partially by the National Natural Science Foundation of China under Grant 61473302, 61170159 and 61503396.

References

[1] Daniel Engel, Lars Hüttenberger, Bernd Hamann, A survey of dimension reduction methods for high-dimensional data analysis and visualization, in: Christoph Garth and Ariane Middel and Hans Hagen (Eds.), Visualization of Large and Unstructured Data Sets: Applications in Geospatial Planning, Modeling and Engineering - Proceedings of IRTG 1131 Workshop 2011, vol. 27, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2012.

[2] D.L. Donoho, et al., High-dimensional data analysis: the curses and blessings of dimensionality, in: AMS Math Challenges Lecture, 2000, pp. 1–32.

[3] L. Tingjin, Z. Jun, L. Lin, Speed up junction detector based on azimuth consensus by Harris corner, Opt. Rev. 21 (2) (2014) 135–142.

[4] C. Hou, C. Zhang, Y. Wu, Y. Jiao, Stable local dimensionality reduction approaches, Pattern Recognit. 42 (9) (2009) 2054–2066.

[5] C. Hou, C. Zhang, Y. Wu, F. Nie, Multiple view semi-supervised dimensionality reduction, Pattern Recognit. 43 (3) (2010) 720–730.

[6] J. Gui, W. Jia, L. Zhu, S.-L. Wang, D.-S. Huang, Locality preserving discriminant projections for face and palmprint recognition, Neurocomputing 73 (13–15) (2010) 2696–2707 , Pattern Recognition in Bioinformatics Advances in Neural Control.

[7] J. Gui, S.-L. Wang, Y.-K. Lei, Multi-step dimensionality reduction and semi-supervised graph-based tumor classification using gene expression data, Artif. Intell. Med. 50 (3) (2010) 181–191.

[8] Y.-K. Lei, Y.-M. Xu, J.-A. Yang, Z.-G. Ding, J. Gui, Feature extraction using orthogonal discriminant local tangent space alignment, Pattern Anal. Appl. 15 (3) (2012) 249–259.

[9] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. Syst. 2 (1) (1987) 37–52.

[10] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugenics 7 (2) (1936) 179–188.

[11] H. Li, T. Jiang, K. Zhang, Efficient and robust feature extraction by maximum margin criterion, IEEE Trans. Neural Netw. 17 (1) (2006) 157–165.

[12] H.S. Seung, D.D. Lee, The manifold ways of perception, Science 290 (5500) (2000) 2268–2269.

[13] J.B. Tenenbaum, V. De Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.

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

[15] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, in: Neural Information Processing Systems, vol. 14, 2001, pp. 585–591.

[16] D.L. Donoho, C. Grimes, Hessian eigenmaps: locally linear embedding techniques for high-dimensional data, Proc. Natl. Acad. Sci. 100 (10) (2003) 5591–5596.

[17] S. Roweis, L.K. Saul, G.E. Hinton, Global coordination of local linear models, Adv. Neural Inf. Process. Syst. 2 (2002) 889–896.

[18] K.Q. Weinberger, L.K. Saul, Unsupervised learning of image manifolds by semidefinite programming, Int. J. Comput. Vis. 70 (1) (2006) 77–90.

[19] D. Hu, G. Feng, Z. Zhou, Two-dimensional locality preserving projections (2dlpp) with its application to palmprint recognition, Pattern Recognit. 40 (1) (2007) 339–342.

[20] H. Zhang, Q. Jonathan Wu, T.W. Chow, M. Zhao, A two-dimensional neighborhood preserving projection for appearance-based face recognition, Pattern Recognit. 45 (5) (2012) 1866–1876.

[21] Wei Li, Qiuqi Ruan, Gaoyun An, Jun Wan, Signal Processing (ICSP), 2012 IEEE 11th International Conference on Discriminative uncorrelated neighborhood preserving projection for facial expression recognition, vol. 2, Beijing, China, 2012, pp. 801–805.

[22] Y. Huang, D. Xu, F. Nie, Patch distribution compatible semisupervised dimension reduction for face and human gait recognition, IEEE Trans. Circuits Syst. Video Technol. 22 (3) (2012) 479–488.

[23] T. Zhang, J. Yang, D. Zhao, X. Ge, Linear local tangent space alignment and application to face recognition, Neurocomputing 70 (7) (2007) 1547–1553.

[24] T.L. Griffiths, M.L. Kalish, A multidimensional scaling approach to mental multiplication, Mem. Cogn. 30 (1) (2002) 97–106.

[25] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. Le Roux, M. Ouimet, Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering, in: Advances in Neural Information Processing Systems, vol. 16, 2004, pp. 177–184.
[26] X. Niyogi, Locality preserving projections, in: Neural Information Processing Systems, vol. 16, 2004. p. 153.
[27] X. He, D. Cai, S. Yan, H.-J. Zhang, Neighborhood preserving embedding, in: The 10th IEEE International Conference on Computer Vision, 2005. ICCV 2005, vol. 2, IEEE, Beijing, China, 2005, pp. 1208–1213.
[28] E. Kokiopoulou, Y. Saad, Orthogonal neighborhood preserving projections, in: The Fifth IEEE International Conference on Data Mining, IEEE, Houston, Texas, USA, 2005, pp. 8–pp.
[29] E. Kokiopoulou, Y. Saad, Orthogonal neighborhood preserving projections: a projection-based dimensionality reduction technique, IEEE Trans. Pattern Anal. Mach. Intell. 29 (12) (2007) 2143–2156.
[30] D. Cai, X. He, Orthogonal locality preserving indexing, in: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Salvador, Brazil, 2005, pp. 3–10.
[31] F. Zang, J. Zhang, J. Pan, Face recognition using elasticfaces, Pattern Recognit. 45 (11) (2012) 3866–3876.
[32] D. Xu, S. Yan, D. Tao, S. Lin, H.-J. Zhang, Marginal fisher analysis and its variants for human gait recognition and content-based image retrieval, IEEE Trans. Image Process. 16 (11) (2007) 2811–2821.
[33] T. Zhang, D. Tao, X. Li, J. Yang, Patch alignment for dimensionality reduction, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1299–1313.
[34] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, S. Lin, Graph embedding and extensions: a general framework for dimensionality reduction, IEEE Trans. Pattern Anal. Mach. Intell. 29 (1) (2007) 40–51.
[35] L.K. Saul, S.T. Roweis, Think globally, fit locally: unsupervised learning of low dimensional manifolds, J. Mach. Learn. Res. 4 (2003) 119–155.
[36] J. Gui, Z. Sun, W. Jia, R. Hu, Y. Lei, S. Ji, Discriminant sparse neighborhood preserving embedding for face recognition, Pattern Recognit. 45 (8) (2012) 2884–2893.

Tingjin Luo is a Ph.D. candidate with the College of Science at the National University of Defense Technology, Changsha, China. He received his Master degree and B.S. degree with the College of Information System and Management at the same university in 2013 and 2011, respectively. His research interests include machine learning, multimedia analysis and computer vision.

Chenping Hou received his B.S. and Ph.D. degree in Applied Mathematics from National University of Defense Technology, Changsha, China in 2004 and 2009, respectively. He is now a lecture in Department of Mathematics and System Science. He is a member of both IEEE and ACM. His current research fields include pattern recognition, machine learning and computer vision.

Dongyun Yi is a Professor in the Department of Mathematics and System Science at the National University of Defense Technology in Changsha, China. He has worked as a Visiting Researcher of the University of Warwick, in 2007. His research interests include systems science, statistics and data processing.

Jun Zhang received her B.S. degree in Application Mathematics from Information Engineering University, Zhengzhou, China, in July 1996. She then received M.S. degree in Management Science and Engineering and Ph.D. in System Engineering from National University of Defense Technology , Changsha, China, in 2002 and 2008, respectively. Her research interests include image processing, image coding, compressive sensing, target detection and data mining.