Feature Extraction via Sparse Difference Embedding (SDE)

Minghua Wan[1*,2,3], Zhihui Lai[1]

[1] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China

[2] Key Laboratory of Trusted Cloud Computing and Big Data Analysis, Nanjing Xiaozhuang University, Nanjing, 211171, P.R. China

[3] Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University), Fuzhou, 350121, China

[e-mail: wmh36@sina.com, Lai_zhi_hui@163.com]

*Corresponding author: Minghua Wan

Received December 11, 2016; revised April 2, 2017; accepted April 28, 2017; published July 31, 2017

Abstract

The traditional feature extraction methods such as principal component analysis (PCA) cannot obtain the local structure of the samples, and locally linear embedding (LLE) cannot obtain the global structure of the samples. However, a common drawback of existing PCA and LLE algorithm is that they cannot deal well with the sparse problem of the samples. Therefore, by integrating the globality of PCA and the locality of LLE with a sparse constraint, we developed an improved and unsupervised difference algorithm called Sparse Difference Embedding (SDE), for dimensionality reduction of high-dimensional data in small sample size problems. Significantly differing from the existing PCA and LLE algorithms, SDE seeks to find a set of perfect projections that can not only impact the locality of intraclass and maximize the globality of interclass, but can also simultaneously use the Lasso regression to obtain a sparse transformation matrix. This characteristic makes SDE more intuitive and more powerful than PCA and LLE. At last, the proposed algorithm was estimated through experiments using the Yale and AR face image databases and the USPS handwriting digital databases. The experimental results show that SDE outperforms PCA LLE and UDP attributed to its sparse discriminating characteristics, which also indicates that the SDE is an effective method for face recognition.

Keywords: Feature extraction; Principal component analysis (PCA); Locally linear embedding (LLE); Sparse constraint; Lasso regression

This work is partially supported by the National Science Foundation of China under grant no. 61462064, 61362031, 61463008, 61403188, 61503195 the China Postdoctoral Science Foundation under grant No. 2016M600674, the Natural Science Fund of Jiangsu Province(Grant No. BK20161580).

https://doi.org/10.3837/tiis.2017.07.016 **ISSN : 1976-7277**

1. Introduction

Subspace learning methods have attracted much attention in the field of feature extraction [1–4]. Two of the most fundamental linear methods are Principal Component Analysis (PCA) [5] and Linear Discriminant Analysis (LDA) [6], which have been successfully applied to many classification problems such as speech recognition, face recognition and multimedia information retrieval. PCA aims to find a set of optimal orthogonal vectors in the sense of minimum mean square error and preserves the global Euclidean structure of the data space. LDA tries to find a set of optimal projection vectors by maximizing the ratio of the between-class scatter to the within-class scatter of the training samples. In order to solve the small sample size (SSS) problems in image vectors based on face recognition, many schemes [1–3] have been proposed.

Some researchers have shown that the face space is more likely to exist in the low -dimensional nonlinear manifold subspace [7]. Therefore, researchers have proposed a large number of nonlinear manifold learning methods. The representative methods are Locally Linear Embedding (LLE) [8], Isomaptic Map (Isomap) [9], Laplacian Eigenmap (LE) [10] and so on. He et al. subsequently proposed Locality Preserving Projections (LPP) [11,12], which is a linear subspace learning method derived from Laplacian Eigenmap. Yang et al. proposed an Unsupervised Discriminant Projection (UDP) [13] algorithm to address the limitation of LPP for the clustering and classification tasks, considering the nonlocal and local quantities at the same time. In the past few years, manifold learning and sparse representation have been widely used for feature extraction and dimensionality reduction[14].

However, all the methods mentioned above find it difficult to give a reasonable interpretation of which features or variables play an important role in real-world applications such as feature extraction and classification [15]. In recent years, sparse subspace learning has aroused great interest from researchers [16,17]. The more effective sparse feature extraction methods are Lasso [18], Lars [19], Elastic Net [20] and so on[21,22]. Robust Structured Subspace Learning (RSSL)[23] is adopted as an intermediate space to reduce the semantic gap between the low-level visual features and the high-level semantics by integrating image understanding and feature learning into a joint learning framework. With the norm, these methods can make the corresponding coefficient of the partial characteristics shrink to zero. Zhang [24] proposed three kinds of local information, namely, local similarity information, local intra-class pattern variation and local inter-class pattern variation.

In order to obtain the local structure and global structure of the data at the same time, and eventually get a sparse transformation matrix, on the basis of sparse subspace learning and manifold learning, in this paper we propose a feature extraction method called Sparse Difference Embedding (SDE). PCA is the global linear method which preserves total variance by maximizing the feature covariance matrix trace. The low-dimensional data obtained by LLE preserves the topology of the original space. Integrating PCA and LLE, SDE seeks to minimize the difference, rather than the ratio, between the local minimizing embedding obtained by LLE and the global maximizing variance obtained by PCA, and then join a sparse constraint in the objective function [25]. The results of experiments on the Yale and AR face image database verify the effectiveness of the proposed method.

The rest of this paper is organized as follows. In section 2, we review the ideas of the relevant methods. In section 3, we propose the idea of SDE in detail. In section 4, experiments are presented to verify the effectiveness of SDE on the Yale and AR face image databases and the USPS handwriting digital databases, and compare it with other methods (PCA, LDA and LLE). Finally we give concluding remarks and a discussion of future work in section 5.

2. Outline of relevant methods

Let us consider a set of N data vector $X=\{x_1,x_2,...,x_N\},x_i\in R^n$, taking values in a n dimensional image space. Let us also consider a linear transformation mapping the original n dimensional space into a d dimensional feature space $Y=\{y_1,y_2,...,y_N\}$, where $y_i\in R^d$ and $n>d$. The new feature vectors $y_i\in R^d$ are defined by the following linear transformation:

$$y_i=U^T x_i , \qquad i=1,...,N \tag{1}$$

where $U\in R^{n\times d}$ is a transformation matrix.

2.1 Principal Component Analysis (PCA)

PCA is one of the most commonly used linear methods. Assume that a point in space is projected into a vector. First, the central point of the original space is defined as:

$$\bar{x}=\frac{1}{N}\sum_{i=1}^{N} x_i \tag{2}$$

Assuming that U is the transformation matrix, the variance after the projection is defined as:

$$\frac{1}{N}\sum_{i=1}^{N}(U^T x_i-U^T \bar{x})^2=U^T SU \tag{3}$$

Using the Lagrange multiplier method, the optimization of the content on the right-hand side of the equal sign is defined as:

$$U^T SU+\lambda(1-U^T U)=0 \tag{4}$$

Take the derivative and find the integration of the formula, and make it equal to 0:

$$SU=\lambda U \tag{5}$$

Then we can obtain the matrix of the biggest variance by taking the projection matrix consisting of the front d eigenvectors.

2.2 Locally Linear Embedding (LLE)

LLE is one of the most classic methods of manifold learning, which can better maintain the original manifold structure after dimensionality reduction. The main process of the algorithm is divided into three steps.

The first step of LLE is to select k -nearest neighbors of each data point x_i using Euclidean distances.

The second step of LLE is to calculate the reconstructing weight matrix $W=\left[w_{ij}\right]_{N\times N}$, which reconstructs each point x_i from its k-nearest neighbors. We can obtain the coefficient matrix W by minimizing the reconstruction error:

$$\min J_L(W)=\sum_{i=1}^{N}\left\|x_i-\sum_{j=1}^{N}w_{ij}x_j\right\|^2 \tag{6}$$

where $w_{ij}=0$ if x_i and x_j are not neighbors, and the rows of W sum to 1: $\sum_{j=1}^{N}w_{ij}=1$.

The reconstruction error for x_i can be converted to this form:

$$\begin{aligned}\xi_i &= \left\|x_i-\sum_{j=1}^{N}w_{ij}x_j\right\|^2 = \left\|\sum_{j=1}^{N}w_{ij}\left(x_i-x_j\right)\right\|^2 \\ &= \sum_{j=1}^{N}w_{ij}\left(x_i-x_j\right)\sum_{t=1}^{N}w_{it}(x_i-x_t) = \sum_{j=1}^{N}\sum_{t=1}^{N}w_{ij}w_{it}G_{jt}^{i}\end{aligned} \tag{7}$$

where $G_{jt}^{i}=\left(x_i-x_j\right)^T(x_i-x_t)$, called the local Gram matrix. By solving the least-squares problem with the constraint $\sum_{j=1}^{N}w_{ij}=1$, the optimal coefficients are given:

$$w_{ij}=\frac{\sum_{t=1}^{N}\left(G^i\right)_{jt}^{-1}}{\sum_{p=1}^{N}\sum_{q=1}^{N}\left(G^i\right)_{pq}^{-1}} \tag{8}$$

After repeating the first step and the second step performed on all the N data points, we can calculate the reconstruction weights to construct a weight matrix $W=\left[w_{ij}\right]_{N\times N}$.

The third step of LLE is to reconstruct represented y_i by the weight matrix W. To maintain the intrinsic geometrical feature of the data after the embedding process, the reconstruction error function must be minimized:

$$\min J_L(Y)=\sum_{i=1}^{N}\left\|y_i-\sum_{j=1}^{N}w_{ij}y_j\right\|^2 \tag{9}$$

where y_i is the mapping output of x_i, y_j is a neighbor of y_i.

Considering the transformation $y_i=U^T x_i$, the objective function reduces to

$$\begin{aligned} J_L(U) &= \sum_{i=1}^{N}\left\| y_i - \sum_{j=1}^{N} w_{ij} y_j \right\|^2 = \sum_{i=1}^{N} tr\left\{ \left(y_i - \sum_{j=1}^{N} w_{ij} y_j \right)\left(y_i - \sum_{j=1}^{N} w_{ij} y_j \right)^T \right\} \\ &= tr\left\{ \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{N} w_{ij} y_j \right)\left(y_i - \sum_{j=1}^{N} w_{ij} y_j \right)^T \right\} \\ &= tr\left\{ Y\left(I - W^T\right)\left(I - W^T\right)^T Y^T \right\} \\ &= tr\left\{ Y(I-W)^T (I-W) Y^T \right\} \\ &= tr\left\{ U^T X\, M X^T U \right\} \end{aligned} \tag{10}$$

where $M = (I-W)^T (I-W)$.

2.3 Lasso regression

In this paper, we use the least absolute shrinkage and selection operator (Lasso) algorithm to obtain a sparse sample solution. This algorithm is obtained by constructing a penalty function to obtain a refined model. It is a biased estimate of the processing of complex linear data. The basic idea of Lasso is to minimize the sum of squared residuals under the condition that the absolute value of the regression coefficient is less than a constant, which can produce a regression coefficient that is strictly equal to 0; then we can obtain an explanatory model.

Assume that we have data $(X^i, y_i), i=1,2,...,N$, and here $X^i = (x_{i1},...x_{ip})^T$ are the variables of the ith observation value, y_i are the corresponding response variables, and p is the number of elements in x_i . Consider the linear regression model:

$$y_i = \alpha_i + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i \tag{11}$$

In the general regression structure, assume that the observation values are independent of each other, or that the response variable y_i is independent of the condition of the observed values that are given, that is:

$$\frac{1}{N}\sum_j x_{ij} = 0, \frac{1}{N}\sum_j x_{ij}^2 = 0, \tag{12}$$

The Lasso is estimated to be:

$$\begin{aligned} &(\hat{\alpha}, \hat{\beta}) = \arg\min\{\sum_i (y_i - \alpha_i - \sum_j \beta_j x_{ij})^2\} \\ &s.t. \sum_j |\beta_j| \le t \end{aligned} \tag{13}$$

Here $t \ge 0$, it is a harmonic parameter, and for all t, the estimate of $\alpha : \hat{\alpha} = \bar{y}$. Assuming that $\bar{y} = 0$, then α will be omitted. The control of harmonic parameters t makes the regression coefficient smaller. Let $t_0=\sum_j|\beta_j|, t\leq t_0$, and it will reduce the number of regression coefficients and tend to 0; some of the coefficients will even equal 0.

3. Sparse Difference Embedding (SDE)

PCA is a global method but it cannot find the local structure of the data. LLE sought a manifold approximated as linear on the local neighbourhood. Combining the global structure of PCA and the local structure of LLE, we propose a feature extraction method called Sparse Difference Embedding (SDE), which is an unsupervised linear method. SDE attempts to create neighbour samples of the original space that are still in the same neighbourhood after being projected into subspace, or far away from each other if not in the same neighbourhood. In addition, we join the sparse constraint in the objective function with solving a constrained optimization problem through the Lasso algorithm to achieve this goal.

3.1 Locally minimizing embedding

To begin with, we propose to minimize the local scatter compactness of each data point by linear coefficients that reconstruct the data point from other points. The technique of local representation is the same as LLE. Considering Equation (9), the objective function is:

$$\min J_L(Y)=\sum_{i=1}^{N}\left\|y_i-\sum_{j=1}^{N} w_{ij}y_j\right\|^2 \tag{14}$$

The objective function reduced from Equation (10) is as follows:

$$J_L(U)=\sum_{i=1}^{N}\left\|y_i-\sum_{j=1}^{N} w_{ij}y_j\right\|^2 = tr\left\{V^T X\ MX^T V\right\} \tag{15}$$

3.2 Globally maximizing variance

Secondly, we propose to maximize the sum of pairwise squared distances between outputs, where PCA preserves the global geometric structure of data in a transformed low-dimensional space. Therefore, maximizing the global scatter of samples is considered:

$$\max J_G(Y)=\sum_{i=1}^{N}\|y_i-\bar{y}\|^2 \tag{16}$$

Considering the transformation in Equation (1), the objective function can be simplified as

$$\begin{aligned} J_G(U) &= \sum_{i=1}^{N}\|y_i-\bar{y}\|^2 = \sum_{i=1}^{N}\left\|U^T(x_i-\bar{x})\right\|^2 \\ &= \sum_{i=1}^{N} tr\left\{U^T(x_i-\bar{x})(x_i-\bar{x})^T U\right\} \\ &= tr\left\{U^T S_T U\right\} \end{aligned} \tag{17}$$

where $\bar{y}=\frac{1}{N}\sum_{i=1}^{N} y_i$, $\bar{x}=\frac{1}{N}\sum_{i=1}^{N} x_i$, and S_T is the total scatter matrix.

3.3 Optimization criterion of SDE

Lastly, when local minimizing embedding and global maximizing variance have been constructed, an intuitive motivation is to find a common projection that minimizes local scatter and maximizes global scatter at the same time, and then joins the sparse constraint in the objective function. Actually, we can obtain such a projection by the following multi-object optimized problem with a sparse constraint, that is:

$$\begin{cases} \min\ tr\left\{U^T XMX^T U\right\} \\ \max\ tr\left\{U^T S_T Y\right\} \end{cases}$$
$$\text{s.t. } U^T XDX^T U = I \tag{18}$$
$$Card(U) = K$$

$Card(U)$ is the number of the non-zero elements in U. SDE seeks to minimize the difference, rather than the ratio, between the local minimizing embedding and the global maximizing variance. So it can be changed into the following constrained problem:

$$\begin{cases} \min\ tr\left\{U^T\left((1-\alpha)XMX^T - \alpha S_T\right)U\right\} \\ \text{s.t. } U^T XDX^T U = I \end{cases} \tag{19}$$
$$Card(U) = K$$

$\alpha(0 \le \alpha < 1)$ is the balance parameter. The minimization can be solved by the Lagrange multiplier method:

$$\frac{\partial}{\partial U} tr\left\{u^T\left((1-\alpha)XMX^T - \alpha S_T\right)u \ -\lambda_i\left(u^T XDX^T u - I\right)\right\} = 0 \tag{20}$$

Let $\frac{\partial L(U,\lambda)}{\partial U} = 0$, then

$$\begin{cases} \left[(1-\alpha)XMX^T - \alpha S_T\right]u_i = \lambda_i XDX^T u_i \\ s.t. Card(U) = K \end{cases} \tag{21}$$

where u_i is a generalized eigenvector corresponding to a generalized eigenvalue λ_i.

However, u_i is not a sparse matrix. Using the Lasso algorithm, with a L_1-norm on u_i, we have:

$$U = \arg\min\left(\sum_{i=1}^{m}(u_i^T x_i - y_i)^2 + \beta\sum_{j=1}^{n}|u_j|\right) \tag{22}$$

However, the number of features selected by Lasso is limited by the number of samples, so we integrate the Lasso regression and Ridge regression [20]:

$$U = \arg\min\left(\sum_{i=1}^{m}(u_i^T x_i - y_i)^2 + \beta\sum_{j=1}^{n}|u_j| + \gamma\sum_{j=1}^{n}(\overline{u_j})^2\right) \tag{23}$$

Then we get an optimal sparse transformation matrix.

3.4 SDE algorithm

The following steps summarize the SDE algorithm described previously:

Step 1. Compute the $J_L(U)$ and $J_G(U)$ matrices using Equations (15) and (17), respectively.

Step 2. Compute the multi-object optimized problem with a sparse constraint using Equation (18).

Step 3. Integrate the Lasso regression and Ridge regression using Equations (23).

Once the project matrix U is obtained through using the SDE algorithm, the nearest neighbour classification becomes straightforward.

4. Experiments and results

In order to verify the effectiveness of the proposed method, we compared it with several other methods (PCA, LDA, UDP and LLE) on the Yale and AR face image databases and the USPS handwriting digital databases. The ORL database was used to examine the performance of the algorithm under conditions where the pose and sample size were varied. Evaluation of the SDE algorithm on variations in both facial expressions and illumination was performed by using the Yale face database. Euclidean distance and nearest neighbour classifier were used in all the experiments. All the experiments used PCA for processing, with about 95% energy of the pictures held. The experiments were carried out on the same computer (Intel (R) Core i3-2130 3.40GHz, Matlab 2014a).

4.1 Experiments using the USPS handwriting digital database

The USPS handwriting digital data includes 10 classes designated from “0” to “9”. Each class has 1100 examples. In this experiment, a subset was selected from the original database. Each image is then cropped to have the size of 16×16. There are 100 images for each class in the subset and the total number is 1000. **Fig. 1** displays a subset of digital “2” from original USPS handwriting digital database.

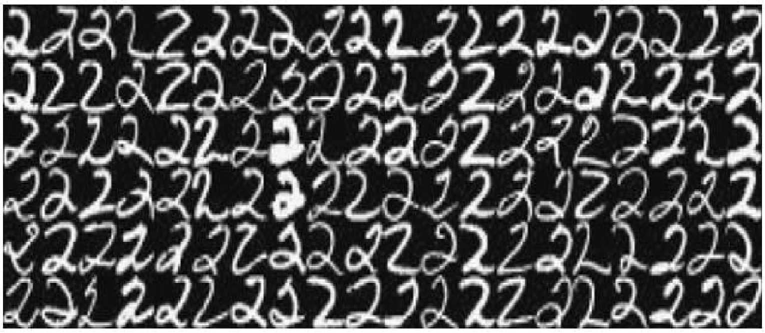

Fig. 1. The sample digital images “2”from the USPS handwriting database.

For each individual, l (=20, 30, 40, 50, 60) images were randomly selected as training samples, and the rests were used for test. For feature extraction, PCA (eigenface), LDA (Fisherface), LLE, UDP and the proposed SDE were used. It is noted that LDA, LLE, UDP and the proposed SDE all involve using a PCA phase. For different algorithms, the optimal PCA dimension may be different, and it is possible it is still an open problem to choose the optimal dimension of PCA. For fair comparisons, in this phase, we kept nearly 95% image energy and selected the number of principal components as 30 for each method. In general, the recognition rates varied with the dimension of the face subspace. The best recognition accuracy of different algorithms is shown in **Table 1**.

Table 1. The best recognition accuracy (%) of different algorithms, with the corresponding dimension shown in parentheses.

	PCA	**LDA**	**LLE**	**UDP**	**SDE**
20	80.88 (20)	82.72 (7)	78.93 (28)	80.53(30)	82.03 (12)
30	84.56 (20)	85.83 (9)	82.75 (30)	85.71(30)	86.17 (15)
40	86.72 (29)	86.80 (8)	85.70 (29)	87.51(27)	88.25 (20)
50	87.96 (26)	88.00 (9)	86.78 (30)	88.45(20)	89.46 (20)
60	88.90 (27)	88.57 (9)	88.82 (30)	89.26(29)	90.40 (30)

4.2 Experiments on the Yale face database

The Yale face database (http://www.cvc.yale.edu/projects/yalefaces/yalefaces.html) contains 165 gray scale images of 15 individuals, each individual has 11 images. The images demonstrate variations in lighting condition, facial expression (normal, happy, sad, sleepy, surprised, and wink). **Fig. 2** show the sample images from the Yale database. In the experiments, l images (l varied from 4 to 6) were randomly selected from the image gallery of each individual to form the training sample set. The remaining $11-l$ images were used for testing. For each experiment with a different training sample size, we independently ran each experiment 10 times. The maximal recognition accuracy of different algorithms is shown in **Table 2**. **Fig. 3** shows the variation of average recognition rate by using different algorithms with the dimension.

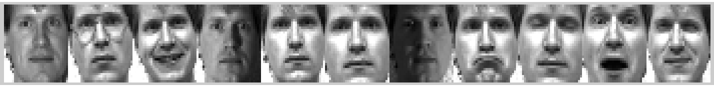

Fig. 2. Images of one person on the Yale face database.

Table 2. The maximal recognition rates (%), with the corresponding dimension shown in parentheses, of the five methods on the Yale face database versus the variation of the training sample sizes.

	PCA	LDA	LLE	UDP	SDE
4	61.30(50)	74.29(6)	70.51(47)	80.50(50)	86.76(27)
5	60.75(41)	77.33(8)	70.49(47)	82.76(47)	86.11(31)
6	61.52(50)	81.11(8)	72.18(43)	84.75(49)	88.40(48)

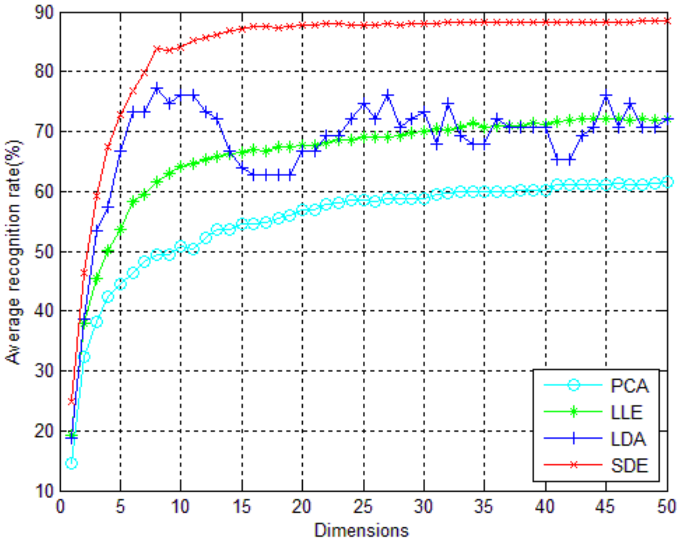

Fig. 3. The average recognition rate (%) of different algorithms varying with feature dimensions when the training sample size is 6 on the Yale face database.

4.3 Experiments on AR the face database

The AR face database (http://cobweb.ecn.purdue.edu/~aleix/aleix_face_DB.html) contains over 4,000 color face images of 126 people (70 men and 56 women), including frontal views of faces with different facial expressions, lighting conditions, and occlusions. **Fig. 4** show the sample images from the AR database. In the experiments, we chose a subset of the AR face database which contained the front 40 people, where each person had 10 images. Then l images (l varied from 4 to 6) were randomly selected from the image gallery of each individual to form the training sample set. The remaining $10-l$ images were used for testing. For each experiment with a different training sample size, we independently ran each experiment 10 times. The average recognition rate of different algorithms when 5 images per person were randomly selected for training is shown in **Fig. 5,** varying with dimensions.

Fig. 4. Images of one person on the AR face database.

Table 3. The maximal recognition rates (%), with the corresponding dimension shown in parentheses, of the five methods on the AR face database versus the variation of the training sample sizes.

	PCA	LDA	LLE	UDP	SDE
4	56.81(50)	59.58(32)	74.45(47)	80.15(50)	81.33(47)
5	56.54(49)	67.50(22)	76.65(50)	84.80(50)	87.15(49)
6	55.42(47)	79.30(25)	74.95(50)	86.08(50)	92.18(50)

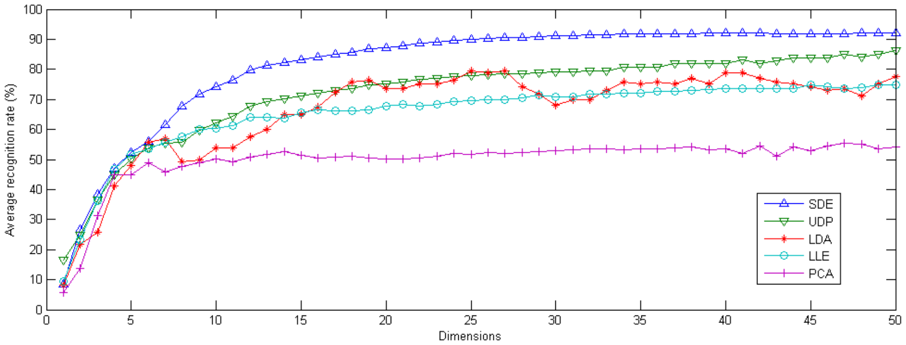

Fig. 5. The average recognition rate (%) of different algorithms varying with feature dimensions when the training sample size is 6 on the AR face database.

4.4 Analysis

From the results in **Table 1**, **Table 2**, **Fig. 3** and **Fig. 5**, we can see first that with a different training sample size, the maximal recognition rates (%) of SDE are significantly higher than that of the other three methods. Secondly, when the training sample size is 6, the result of SDE is more robust than that of PCA, LDA, UDP and LLE; with the increase in the feature dimension, the recognition rate is continuously rising based on the test results by using the USPS handwriting digital database, and Yale and AR face image databases. For tests conducted on images from the USPS, and Yale and AR face image databases, the best performances of average recognition rates of using SDE exceed the best performances of using the other methods. Experimental results of using the USPS, Yale face and AR face databases (**Tables 1** to **3**) show that the increases in the best performances of average recognition rates by using SDE are higher than the results from the other algorithms. The reason may be that, on the one hand, SDE can obtain the local structure and non-local structure of the samples at the same time, but PCA, LDA, UDP and LLE can obtain only the local or non-local structures. On the other hand, the sparsity can extract more effective discriminant features from the training samples. The advantage of the SDE algorithm is that it can efficiently handle the vagueness and ambiguity of samples being degraded by poor illumination, shape and facial expression variation, yet it does not change the class of samples. The affinity weights of the novel neighbourhood graphs (intra-class and inter-class) instead of the weights of the binary pattern are defined, which reduces the sensitivity of the method to the substantial sample variations caused by illumination, shape and viewing conditions. Furthermore, the outlier images may cause errors in the estimation of intra-class scatter and global covariance and lead to low recognition rates for PCA and LDA methods, while the SDE method minimizes the effects of outlier images. The factors discussed above contribute to the superior effectiveness of the SDE method.

5. Conclusion

In this paper, we developed an unsupervised linear method called Sparse Difference Embedding (SDE), which combined the two methods of PCA and LLE, preserving the global structure and local structure data at the same time. In addition, we joined the sparse constraint in the method with solving a constrained optimization problem by the Lasso algorithm, and finally we obtained a sparse solution. The most important and interesting observation is that the sparse projections learned by SDE have a direct physical interpretation on which features or variables are contributive to feature extraction and discrimination. The results of the experiments on the Yale face database, AR face database and USPS handwriting digital databases showed that, in comparison with PCA, LDA, UDP and LLE, SDE had a higher performance in feature extraction. Specifically, for face recognition, the sparse face subspaces show us an intuitive, semantic and insightful understanding of the feature extraction. In future work, we will extend the SDE algorithm to kernel and tensor form via the kernel and tensor methods, respectively.

References

[1] W. Fang, P. Ma, Z. Cheng, D. Yang, X. Zhang, “2-dimensional projective non-negative matrix factorization and its application to face recognition [J],” *Acta Automatica Sinica*, 38: 1503-1512, 2012. Article (CrossRef Link).

[2] M. Shao, D. Kit, Y. Fu, “Generalized transfer subspace learning through low-rank constraint [J],” *International Journal of Computer Vision*, 109: 74-93, 2014. Article (CrossRef Link).

[3] Y. Xie, W. Zhang, Y. Qu, “Discriminative subspace learning with sparse representation view-based model for robust visual tracking [J],” *Pattern Recognition*, 47: 1383-1394, 2014. Article (CrossRef Link).

[4] S. Wang, W. Pedrycz, Q. Zhu, “Subspace learning for unsupervised feature selection via matrix factorization [J],” *Pattern Recognition*, 48: 10-19, 2015. Article (CrossRef Link).

[5] M. Turk, A. Pentland, “Eigenfaces for recognition [J],” *Cognitive Neuroscience*, 3: 71-86, 1991. Article (CrossRef Link).

[6] D. Swets, J. Weng, “Using discriminant eigenfeatures for image retrieval [J], *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18: 831-836, 1996. Article (CrossRef Link).

[7] Wan, Minghua, et al., "Local graph embedding based on maximum margin criterion via fuzzy set[J]," *Fuzzy Sets and Systems* 318, 120-131, 2017. Article (CrossRef Link).

[8] S. Roweis, L. Saul, “Nonlinear Dimensional Reduction by Locally Linear Embedding [J],” *Science*. 290: 2323-2326, 2000. Article (CrossRef Link).

[9] J. Tenenbaum, V. DeSilva, J. Langford, “A Global Geometric Framework for Nonlinear Dimensionality Reduction [J],” *Science*, 290: 2319-2323, 2000. Article (CrossRef Link).

[10] M. Belkin, P. Niyogi, “Laplacian Eigenmaps for Dimensionality Reduction and Data Representation [J],” *Neural Computation*.15: 1373-1396, 2003. Article (CrossRef Link).

[11] Wan, Minghua, et al., "Feature extraction using two-dimensional maximum embedding difference[J]," *Information Sciences*, 274, 55-69, 2014. Article (CrossRef Link).

[12] X. He, S. Yan, Y. Hu, Niyogi, H. Zhang. Face Recognition Using Laplacianfaces [J],” *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27: 328-340, 2005. Article (CrossRef Link).

[13] J. Yang, D. Zhang, J.Y. Yang, B. Niu, “Globally Maximizing, Locally Minimizing: Unsupervised Discriminant Projection with Applications to Face and Palm Biometrics [J],” *IEEE Trans Pattern Anal and Mach Intelligence*, 29: 650-664, 2007. Article (CrossRef Link).

[14] Gao S Q, Jing X Y, Lan C, et al, “Feature extraction based on sparsity embedding with manifold information for face recognition[C],” in *Proc. of Computational Intelligence and Software Engineering (CiSE), 2010 International Conference on. IEEE*, 2010. Article (CrossRef Link).

[15] Z. Lai, “Sparse facial feature extraction via Manifold learning [D],” *Nanjing University of Technology and Engineering*, 2011.

[16] A. Wagner, J. Wright, “A. Ganesh, et al. Toward a practical face recognition system: Robust alignment and illumination by sparse representation [J],” *IEEE Trans on Pattern Analysis and Machine Intelligence*, 34: 372-386, 2012. Article (CrossRef Link).

[17] J. Yang, L. Zhang, Y. Xu, “Beyond sparsity: The role of L1-optimizer in pattern classification [J],” *Pattern Recognition*, 45: 1104-1118, 2012. Article (CrossRef Link).

[18] Tibshirani, Robert, "Regression shrinkage and selection via the lasso: a retrospective," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.3, 273-282, 2011. Article (CrossRef Link).

[19] B. Efron, T. Hastie, I. Johnstone, “Least angle regression [J],” *Annals of Statistics*, 32(2):407-99, 2004. Article (CrossRef Link).

[20] H. Zou, T. Hastie, “Regression shrinkage and selection via the Elastic Net, with applications to microarrays [J],” *Journal of the Royal Statistical Society Series B: Methodological*, 67: 301-320, 2005. Article (CrossRef Link).

[21] Li Z, Liu J, Tang J, et al, “Robust structured subspace learning for data representation[J],” *IEEE transactions on pattern analysis and machine intelligence*, 37(10): 2085-2098, 2015. Article (CrossRef Link).

[22] Li Z, Tang J, “Unsupervised feature selection via nonnegative spectral analysis and redundancy control[J],” *IEEE Transactions on Image Processing*, 24(12): 5343-5355, 2015. Article (CrossRef Link).

[23] Li Z, Tang J, “Weakly Supervised Deep Matrix Factorization for Social Image Understanding[J],” *IEEE Transactions on Image Processing*, 26(1): 276-288, 2017. Article (CrossRef Link).

[24] Zhang D, He J, Zhao Y, et al, "Global plus local: A complete framework for feature extraction and recognition[J]," *Pattern Recognition*, 47(3): 1433-1442, 2014. Article (CrossRef Link).

[25] M. Imani, H. Ghassemian, "Ridge regression-based feature extraction for hyperspectral data [J]," *International Journal of Remote Sensing*, 36: 1728-1742, 2015. Article (CrossRef Link).

Minghua Wan received his BS degree in automated institute from the Nanchang University of Aviation in 2003 and his MS and PhD degrees in pattern recognition and intelligent systems from the Nanjing University of Science and Technology (NUST) in 2007 and 2011, respectively. He is the author of more than 20 scientific papers in pattern recognition and computer vision. He is a associate professor at the Nanjing Audit University. His current research interests include face recognition and detection, and image processing.

Zhihui Lai received the B.S degree inmathematics from South China NormalUniversity, M.S degree from Jinan University,and the PhD degree in pattern recognition and intelligence system from Nanjing Universityof Science and Technology (NUST), China, in 2002, 2007 and 2011,respectively. He was a postdoctoral fellow at Bio-Computing ResearchCenter, Shenzhen Graduate School, Harbin Institute of Technology, China. He has also been a research associate, postdoctoral fellow and research fellow since 2010 at The Hong Kong Polytechnic University. He has published 50+ academic papers, including 20+ papers published on IEEE Transactions.His research interests include face recognition, image processing and content-basedimage retrieval, pattern recognition, compressive sense, human visionmodelization and applications in the fields of intelligent robot research.