2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

A Supervised Neighborhood Preserving Embedding for Face Recognition

Xing Bao
School of Computer Science and Technology& Provincial key Laboratory for Computer information processing technology Soochow University Suzhou 215006, Jiangsu, China
20124227034@suda.edu.cn

Li Zhang
School of Computer Science and Technology& Provincial key Laboratory for Computer information processing technology Soochow University Suzhou 215006, Jiangsu, China
zhangliml@suda.edu.cn

Bangjun Wang
School of Computer Science and Technology& Provincial key Laboratory for Computer information processing technology Soochow University Suzhou 215006, Jiangsu, China
wangbangjun @suda.edu.cn

Jiwen Yang
School of Computer Science and Technology& Provincial key Laboratory for Computer information processing technology Soochow University Suzhou 215006, Jiangsu, China
jwyang @suda.edu.cn

***Abstract*—Neighborhood preserving embedding (NPE) is an approximation to locally linear embedding (LLE), which has an ability to preserve local neighborhood structure on data manifold. As an unsupervised dimensionality reduction method, NPE can be applied to face recognition for preprocessing. However, NPE could not utilize the label information in the classification tasks. To make the data in a reduced subspace separable, this paper proposes a supervised neighborhood preserving embedding which could learn a projection matrix by using both the geometrical manifold structure and the label information of the given data. In addition, the projection matrix could be found by solving a linear set of equations. Experimental results on ORL and Yale face image datasets show that the proposed method has a high recognition rate.**

Keywords—face recognition; dimension reduction; label information; local preserving embedding

I. INTRODUCTION

In the real world, data exists in the form of high-dimensionality, including image data. Dealing with high-dimensional data directly would cause a large computational complexity, the curse of dimensionality and other problems [1]. It is an effective way to overcome the problems caused by high-dimensional data by projecting the high-dimensional data into low-dimensional subspaces. Therefore, dimensionality reduction plays an important role in their specific applications, including data visualization, data compression [2], pattern recognition and classification [3], multimedia information retrieval and others.

The two classical linear embedding methods are linear discriminant analysis (LDA) [4] and principal component analysis (PCA) [5]–[6], which are demonstrated to be computationally efficient and suitable for practical applications. LDA is a supervised dimensionality reduction algorithm. This algorithm aims to find the optimal projection vector on which the data points of different classes are far from each other and the data points of the same class are to be as close to each other as possible. PCA, an unsupervised method, is famous for the low-dimensional representation of high-dimensional data. In other words, LDA tries to find the expected projection direction of the data for classification tasks, while PCA seeks for an effective way to represent data for compressing data.

Manifold learning is a typical nonlinear dimensionality reduction method. Usually, manifold learning is first to construct a data adjacency graph to characterize the data distribution or geometry, and then seek for an optimal mapping or a projection direction to effectively maintain the structure. Most manifold learning algorithms, such as laplacian eigenmaps (LE) [7], locally linear embedding (LLE) [8], locality preserving projection (LPP) [9] and neighborhood preserving embedding (NPE) [10], even including the classic PCA and LDA which can be unified under the framework of adjacency graph construction and dimensionality reduction.LPP is a linear approximation to LE. The goal of LPP is to project the high-dimensional data into a low-dimensional manifold subspace that can better preserve the original data's locality. In other words, the adjacent data points in the original data space can also maintain adjacent relationship respectively in the projection subspace.

NPE is a linear approximation to LLE and has an ability to maintain the local neighborhood information of data manifold. NPE has received extensive attention in face recognition [11-19]. However, in face recognition tasks, NPE is used as an unsupervised dimensionality reduction method, and cannot take advantage of the label information on given data. To utilize the label information, Du et al proposed a new subspace learning method called neighborhood preserving discriminant embedding (NPDE) [20]. NPDE keeps the data information in the local neighborhood manifold structure while emphasizing the discrimination information of high-dimensional data. It can ensure the minimum reconstruction error of local neighborhood and make the projection sample points with minimum within-class scatter and maximum between-scatter. Unlike many existing techniques such as LPP and NPE, in which the local neighborhood information is preserved during the dimension reduction procedure, sparsity preserving projection (SPP) [21] aims to preserve the sparse reconstructive relationship of the data, which is achieved by

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

minimizing a l_1 regularization-related objective function. It is well known that maximum margin criterion (MMC) [23] is a method proposed to maximize the trace of the difference of the between-class scatter matrix and within-class scatter matrix. Thus, discriminant sparse neighborhood preserving embedding (DSNPE) [22] was proposed by introducing MMC into the objective function of SPP, which has two advantages: (1) it retains the sparsity characteristic of SPP; (2) it emphasizes the discriminative information by incorporating MMC, which can make the class mean vectors have a wide spread and make every class scatter in a small space. Furthermore, to further increase the discriminative power of DSNPE, it integrates additional discriminant information.

This paper proposes a novel supervised neighborhood preserving embedding (SNPE). Different from NPDE, SNPE utilizes the label information to construct attraction vectors each of which would attract points in the same class. Meanwhile, SNPE requires preserving the local neighborhood structure on data manifold. By doing so, the embedded points in the same class would be close to each other, while the points in the different classes would be far away from each other. In addition, SNPE is cast into a linear set of equations, which is easier to solve. .

The rest of this paper is organized as follows. In Section 2, we briefly review NPE and NPDE. Section 3 proposes SNPE. In Section 4, we compare SNPE with some related works and give experimental results. Conclusions are made in Section 6.

II. RELATED WORK

A. Neighborhood Preserving Embedding

Let the training samples matrix be $\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n]$, where $\mathbf{x}_i \in R^m$, m is the dimensionality of the training samples and n is the number of the training samples. NPE is intended to reduce the dimensionality of data and maintain the inherent local neighbor manifold structure at the same time. It seeks for an optimal transformation matrix $\mathbf{A}=[\mathbf{a}_1,\mathbf{a}_2,...,\mathbf{a}_m]$, where $\mathbf{a}_i \in R^d$, which can map the high-dimensional data into a relatively low-dimensional feature subspace.

Similar to LLE, NPE evaluates the affinity weight matrix by using local least squares approximation. The local approximation error in NPE is measured by minimizing the following cost function:

$$\varphi(\mathbf{W})=\sum_{i=1}^{n}\left\|\mathbf{x}_i-\sum_{j=1}^{n}\mathbf{W}_{ij}\mathbf{x}_j\right\|^2 \tag{1}$$

under two constraints: (1) if $\mathbf{x}_j$ is not one of k neighbors of $\mathbf{x}_i$, then $\mathbf{W}_{ij}=0$; otherwise, $\mathbf{W}_{ij}\neq 0$ and $\sum_{j=1}^{n}\mathbf{W}_{ij}=1$, $j=1,2,...,n$. A reasonable criterion for choosing a "good" projection is minimizing the cost function:

$$\begin{aligned}\mathbf{A}&=\arg\min_{\mathbf{A}}\left[\sum_{i=1}^{n}\left\|\mathbf{y}_i-\sum_{j=1}^{n}\mathbf{W}_{ij}\mathbf{y}_j\right\|^2\right] \\ &=\arg\min_{\mathbf{A}} tr(\mathbf{A}^T\mathbf{X}\mathbf{M}\mathbf{X}^T\mathbf{A})\end{aligned} \tag{2}$$

which subject to $\mathbf{A}^T\mathbf{X}\mathbf{X}^T\mathbf{A}=\mathbf{I}$, where $tr(\bullet)$ is the trace of matrix $\bullet$, $\mathbf{I}$ is the identity matrix, $\mathbf{Y}=\mathbf{A}\mathbf{X}$ and $\mathbf{M}=(\mathbf{I}-\mathbf{W})^T(\mathbf{I}-\mathbf{W})$ is a sparse, symmetric, and semi-positive definite matrix.

By using Lagrange multiple technique, NPE leads to the following generalized eigenvector problem:

$$\mathbf{X}\mathbf{M}\mathbf{X}^T\mathbf{A}=\lambda\mathbf{X}\mathbf{X}^T\mathbf{A} \tag{3}$$

B. Neighborhood Preserving Discriminant Embedding

NPE is an unsupervised learning method, and it could not utilize the label information on given data in classification tasks. Therefore, NPDE was proposed in [20]. NPDE keeps the local neighborhood structure on data manifold and simultaneously emphasizes the discrimination information of data. It can make the local neighborhood reconstruction error minimal, and maintain points with minimum within-class scatter and maximum between-scatter in the subspace. Similarly, both of two methods involve solving the characteristic of the matrix decomposition problem.

Let $\mathbf{X}=[\mathbf{X}_1,\mathbf{X}_2,...,\mathbf{X}_C]$ represent the training samples matrix in original subspace, where $\mathbf{X}_c$ denotes the sample matrix belonging to the cth class, and C is the total number of classes. Let n_c be the number of the training samples in the cth class. So, the total number of the training samples is $n=\sum_{c=1}^{C}n_c$. Let $\mathbf{Y}=[\mathbf{Y}_1,\mathbf{Y}_2,...,\mathbf{Y}_C]^T$ be the projected training sample matrix in the low dimensional subspace, we have the objective function of NPDE:

$$\min\quad J=\frac{\sum_{c=1}^{C}\sum_{i=1}^{n_c}[\mathbf{y}_i^c-\sum_{j=1}^{n_c}\mathbf{W}_{ij}^c\mathbf{y}_j^c]}{\sum_{c=1}^{C}n_c(\mathbf{u}_c-\mathbf{u})(\mathbf{u}_c-\mathbf{u})^T} \tag{4}$$

where $\mathbf{y}_i^c$ and $\mathbf{y}_j^c$ respectively denote the ith and jth embedded vectors in the cth class, $\mathbf{W}_{ij}^c$ denotes the reconstruction weighting coefficient of training samples in the cth class, $\mathbf{u}_c$ denotes the mean of embedded vectors in the cth class and $\mathbf{u}$ denotes the mean of all embedded vectors.

The objective function of NPDE can be reduced to (5)

$$\min\quad J=\frac{\boldsymbol{\alpha}^T\mathbf{X}\mathbf{M}\mathbf{X}^T\boldsymbol{\alpha}}{\boldsymbol{\alpha}^T\mathbf{S}_B\boldsymbol{\alpha}} \tag{5}$$

where $\mathbf{S}_B = \sum_{c=1}^{C} n_c(\mathbf{m}_c - \mathbf{m})(\mathbf{m}_c - \mathbf{m})^T$ denotes the between-class scatter matrix in the original space, $\mathbf{m}_c = \frac{1}{n_c}\sum_{i=1}^{n_c}\mathbf{x}_i^c$, $\mathbf{m} = \frac{1}{n}\sum_{c=1}^{C}\sum_{i=1}^{n_c}\mathbf{x}_i^c$, the matrix $\mathbf{M}$ is

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_1 & 0 & \dots & 0 \\ 0 & \mathbf{M}_2 & \dots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \dots & \mathbf{M}_C \end{bmatrix}$$

with $\mathbf{M}_c = (\mathbf{I} - \mathbf{W}_c)^T(\mathbf{I} - \mathbf{W}_c)$, $\mathbf{W}_c \in R^{n_c \times n}$ whose ith row and jth column is $\mathbf{W}_{ij}^c$.In addition, the rank of $\mathbf{S}_B$ is $C-1$.

The optimal transformation matrix can be obtained by minimizing the objective function of (5). Minimizing the objective function of (5) is equivalent to minimize the numerator term and maximize the denominator term of the objective function simultaneously.

III. Supervised Neighborhood Preserving Embedding

In order to incorporate the label information on the given data, this paper proposes an alternative supervised NPE method, called SNPE. In our method, we construct attraction vectors by using the label information of training samples and make the samples in the subspace drawn to these attraction points.

Assume that there has a set of training samples $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$, where $\mathbf{x}_i \in R^m$, $\mathbf{y}_i \in \{1, 2, \cdots, c\}$, m is the dimensionality of the training samples, n is the total number of the training samples, and C is the number of classes. Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n] \in R^{m \times n}$ be the training samples matrix. For each training sample, we construct an attraction point $\mathbf{h}_i \in R^C$ by using its label information. If $\mathbf{y}_i = c$, then the cth entry of $\mathbf{h}_i$ is 1 and other entries are zero. Thus training samples belonging to the same class share the same attraction point. We hope that each sample in the subspace could be attracted to its attraction point.

To make the tradeoff between the geometric characteristics of low-dimensional coordinate point and label information, we seek for an optimal projection matrix $\mathbf{A} \in R^{m \times C}$, which can project training points into a relatively low-dimensional feature subspace. The sample matrix in the subspace could be represented as $\mathbf{Y} = \mathbf{A}^T\mathbf{X}$.

Based on NPE, we introduce a discriminate information term and obtain the following optimal problem:

$$\min_{\mathbf{A}} \quad \frac{1}{2}\sum_{i=1}^{n}\left\|\mathbf{y}_i - \sum_{j=1}^{k}\mathbf{W}_{ij}\mathbf{y}_j\right\|^2 + \beta\sum_{i=1}^{n}\|\mathbf{y}_i - \mathbf{h}_i\|^2 \qquad (6)$$

where $\beta \in [0, +\infty)$ is a balance parameter which is used to balance the importance of the label information. The reconstruction weighting coefficient $\mathbf{W}$ can be obtained by solving (1).

From (2) we can know that the first term in (6) can be written as

$$\sum_{i=1}^{n}\left\|\mathbf{y}_i - \sum_{j=1}^{k}\mathbf{W}_{ij}\mathbf{y}_j\right\|^2 = \mathbf{YMY} = \mathbf{A}^T\mathbf{XMX}^T\mathbf{A} \qquad (7)$$

and the second term in (6) can be modified as

$$\begin{aligned}\beta\sum_{i=1}^{n}\|\mathbf{y}_i - \mathbf{h}_i\|^2 &= \beta(\mathbf{Y} - \mathbf{H})(\mathbf{Y} - \mathbf{H})^T \\ &= \beta(\mathbf{A}^T\mathbf{X} - \mathbf{H})(\mathbf{A}^T\mathbf{X} - \mathbf{H})^T\end{aligned} \qquad (8)$$

where the attraction matrix $\mathbf{H} = [\mathbf{h}_1, \cdots, \mathbf{h}_n] \in R^{C \times n}$. Obviously, (6) is equivalent to the following optimal problem

$$\min_{\mathbf{A}} \quad tr\left(\mathbf{A}^T\mathbf{XMX}^T\mathbf{A} + \beta(\mathbf{A}^T\mathbf{X} - \mathbf{H})(\mathbf{A}^T\mathbf{X} - \mathbf{H})^T\right) \qquad (9)$$

We show the solution to (9) in the following theorem.

Theorem 1. Given the symmetric, and semi-positive definite matrix $\mathbf{M} \in R^{n \times n}$, the real matrix $\mathbf{H} \in R^{C \times n}$, $\beta \in R^+$ and a full rank matrix $\mathbf{X} \in R^{m \times n}$, the object function of (9) has its optimal value when $\mathbf{A} = \beta(\mathbf{XMX}^T + \beta\mathbf{XX}^T)^{-T}\mathbf{XH}^T$.

Proof. Let $L(\mathbf{A}) = \mathbf{A}^T\mathbf{XMX}^T\mathbf{A} + \beta(\mathbf{A}^T\mathbf{X} - \mathbf{H})(\mathbf{A}^T\mathbf{X} - \mathbf{H})^T$. Obviously $tr(L(\mathbf{A}))$ will have the minimal value when the derivative of $L(\mathbf{A})$ equals to 0, that is

$$\frac{\partial L(\mathbf{A})}{\partial \mathbf{A}} = 0 \Rightarrow \mathbf{A}^T\mathbf{XMX}^T + \beta(\mathbf{A}^T\mathbf{X} - \mathbf{H})\mathbf{X}^T = 0 \qquad (10)$$

(10) can be modified as

$$(\mathbf{XMX}^T + \beta\mathbf{XX}^T)^T\mathbf{A} = \beta\mathbf{XH}^T \qquad (11)$$

Now we need to certify that $\mathbf{XMX}^T + \beta\mathbf{XX}^T$ is invertible, which could be rewritten as

$$\mathbf{XMX}^T + \beta\mathbf{XX}^T = \mathbf{X}(\mathbf{M} + \beta\mathbf{I})\mathbf{X}^T \qquad (12)$$

where $\mathbf{I} \in R^{n \times n}$ is the identity matrix.

The matrix $\mathbf{X} \in R^{m \times n}$ is a full rank real matrix. If $m < n$, $\mathbf{X}$ is a row full rank matrix, otherwise $\mathbf{X}$ is a column full rank matrix. Assume $\mathbf{X}$ is a row full rank matrix, and the rank of $\mathbf{X}$ is m. The case of column full rank is similar to that of row full rank. Thus, we only observe that $\mathbf{X}(\mathbf{M} + \beta\mathbf{I})\mathbf{X}^T$ is invertible under the circumstance of row full rank.

Since $\mathbf{M} + \beta\mathbf{I}$ is real symmetric positive definite matrix, we can apply the square root decomposition on $\mathbf{M} + \beta\mathbf{I}$ as follows:

$$\mathbf{M} + \beta\mathbf{I} = \mathbf{LL}^T$$

where $\mathbf{L} \in \mathbf{R}^{n \times n}$ is a positive definite full rank matrix. So (12) can be represented as

$$\mathbf{X}(\mathbf{M} + \beta\mathbf{I})\mathbf{X}^T = \mathbf{DD}^T$$

where $\mathbf{D} = \mathbf{XL}$ is still a full rank matrix. We only verify $\mathbf{DD}^T$ is invertible as follows.

For an arbitrary vector $\mathbf{s} \in R^m$, where $\mathbf{s}$ is a nonzero vector, we can get $\mathbf{s}^T(\mathbf{D}^T\mathbf{D})\mathbf{s} = (\mathbf{Ds})^T(\mathbf{Ds})$. Let $\mathbf{Ds} = [t_1, t_2, ..., t_m]^T$, thus $\mathbf{s}^T(\mathbf{D}^T\mathbf{D})\mathbf{s} = t_1^2 + t_2^2 + ... + t_m^2 \geq 0$. If $\mathbf{s}^T(\mathbf{D}^T\mathbf{s})\mathbf{D}=0$, then $t_1=t_2=\cdots=t_m=0$. The linear set of equations $\mathbf{Ds}=\mathbf{0}$ has a zero solution, or $\mathbf{s}=\mathbf{0}$ where $\mathbf{0}$ denotes a vector of all zeros. Thus, we have $rank(\mathbf{D}) < m$ which contradicts the known or $\mathbf{s}$ is a nonzero vector. Therefore, we can get $t_1^2 + t_2^2 + ... + t_m^2 > 0$. Finally we can observe that $\mathbf{DD}^T$ is non-singular, that is, $\mathbf{X}(\mathbf{M} + \beta\mathbf{I})\mathbf{X}^T$ is invertible.

This means $L(\mathbf{A})$ would get the minimal value when

$$\mathbf{A} = \beta\left(\mathbf{XMX}^T + \beta\mathbf{XX}^T\right)^{-T}\mathbf{XH}^T \qquad (13)$$

This completes the proof.

From Theorem 1, equation (9) is cast into a linear set of equations, which is very easier to solve. Our approach is a promotion to the unsupervised NPE algorithm. We add discriminative information by constructing an attraction matrix and introduce a parameter β to control the importance of the label information. If the label information is not untrustworthy, we can let $\beta = 0$. Equation (9) is equivalent to (2).

The improvement of SNPE algorithm over NPE method benefits mostly from two aspects: one aspect is that SNPE tries to find the subspace that best discriminates different face classes; the other aspect is that SNPE reduces the energy of noise and transformation difference as much as possible.

IV. EXPERIMENTS

To verify the effectiveness of SNPE, two experiments are carried out here. The first one is performed on a two-dimensional artificial dataset. In the second experiment, two well-known and benchmark face image databases (ORL and Yale [24]) are used to evaluate the performance of SNPE by comparing with PCA, LDA, LPP, NPE, and NPDE. To make the comparison fair, for all the evaluated algorithms, we first apply PCA on the face data to reduce the dimensionality and remove the noise. A nearest neighbor (NN) classifier is employed to classify the projected samples. The experiments are implemented on MATLAB platform.

A. Artificial dataset

In order to compare SNPE with NPE and NPDE and analyze the involved parameter, we generate a two-class synthetic dataset which can better be visualized in the 3-dimensional space. The first class is generated from the Gaussian distribution with a mean $[0,0,0]^T$ and a covariance matrix $\mathbf{I} \in R^{3\times3}$, while the second class is generated from the Gaussian distribution with a mean $[2,2,2]^T$ and a covariance matrix $\mathbf{I} \in R^{3\times3}$. We randomly generate 100 datasets. Each class has 100 training and 100 test points in each trial. We try to project these points into a two-dimensional space. NPE, NPDE and SNPE all have a neighborhood parameter k when constructing the adjacent graph. In addition, SNPE is also involved in a control parameter β which has an effect on results of embedding projection theoretically.

We first observe the effect of β on the recognition accuracy of SNPE when setting $k = 5$. Let β change in the set $\{2^{-4}, 2^{-3}, \cdots, 2^4\}$. We repeat 100 times independently and report the average recognition accuracy as depicted in Fig 1. From this figure, we can see that the curve remains flat with varying the parameter β since the standard deviation on the whole parameter set is 1.089×10^{-6}, which is very small so that it can be ignored approximately. Therefore, in order to make the tradeoff between the original manifold geometry and label information of training samples, β=1 is an ideal choice in the experiment.

Since NPE, NPDE and SNPE are related to the adjacency graph neighborhood parameter k, which has different effects on projection results. For these three methods, k varies from 1 to 15. The experiments are repeated 100 times and the average accuracy is recorded. Fig. 2 illustrates that recognition accuracy curves of four methods vary with different neighborhood parameters. The curve of "NN" in Fig. 2 denotes the result of the nearest neighbor classifier without dimensionality reduction. Since "NN" is independent of the graph neighborhood parameter, the result of "NN" is a fixed value. Observation on Fig. 2 indicates that SPNE is much better than the other three methods and is relative stable when varying k. NPE and NPDE are sensitive to selection for neighbor parameter with different degrees. NPDE has a relatively large fluctuation. These three methods almost have a higher accuracy when $k = 3$. The following experiments, k is taken to 3.

Fig. 3(a) shows the randomly generated data points in one trial. Two-class points are represented by '+' and 'o', respectively. Fig 3 (b), (c) and (d) show the projected points in the two-dimension space obtained by NPE, NPDE and SNPE, respectively. Note that the data points overlap significantly in the original 3-dimensional space, the dimensionality reduction data points would still overlap. We can see that the projected point obtained by NPDE is greatly overlapping, which shows that the method cannot utilize the label information well and perform even worse than NPE. Relatively speaking, SNPE is better than the other two methods, which means it is easy to perform classification tasks. Visualization on points in Fig. 3 also reflects the results of Fig. 2.

B. ORL Database

The ORL face database consists of a total of 400 face images, and of a total of 40 people (10 samples per person). For some subjects, the images were taken at different times, varying the lighting, facial expressions (open/closed eyes, smiling/not smiling) and facial details (glassed/no glassed). All the images were taken against a dark homogeneous background with the subjects in an upright, front position

(with tolerance for some side movement). All images are grayscale and normalized to a resolution of 92×112 pixels.

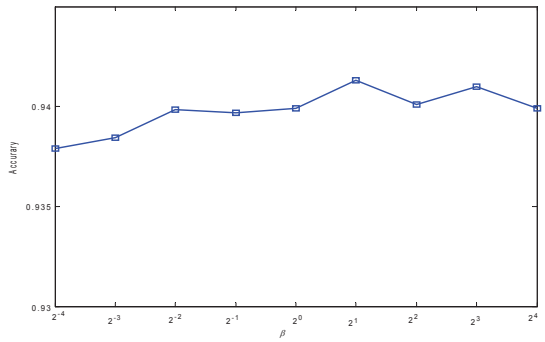


Fig.1 Recognition accuracy under different control parameters

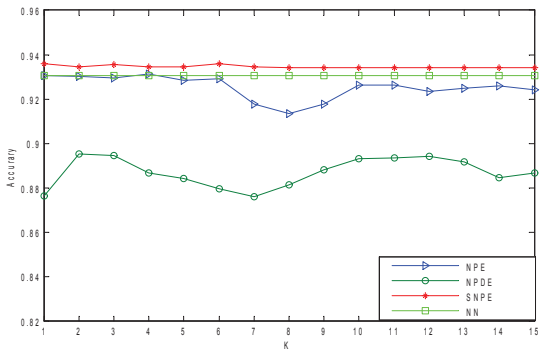
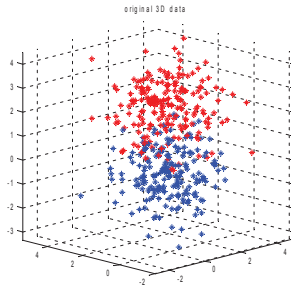
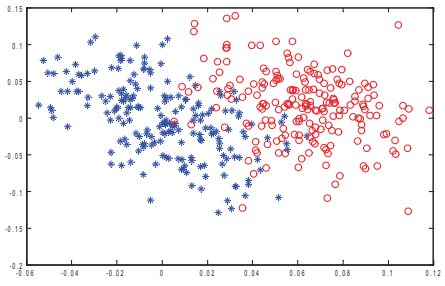


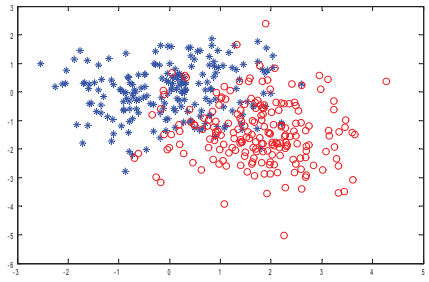
Fig.2 Recognition accuracy under different neighborhood parameter



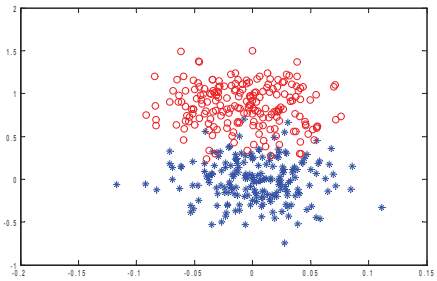
(a)Random data



(b) NPE



(c) NPDE



(d) SNPE

Fig.3 Random data and the results of visualization

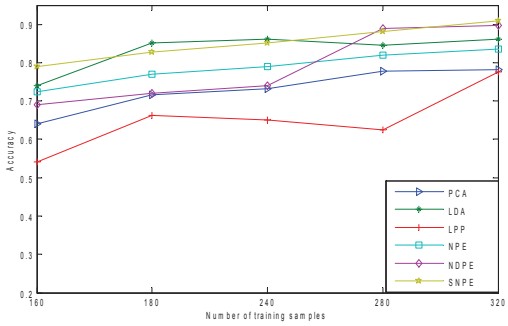


Fig.4. Results on the ORL database

In this experiment, we certify the recognition accuracy of several algorithms with different training samples. We randomly select n samples of each individual for training, and the rest $(10-n)$ for testing. The number of the training sample set varies from 160 to 320. To overcome the complication of singular matrices, we first apply PCA on the face data to reduce dimensionality and remove the noise and remove its null space so that the resulting matrix is non-singular. Then the proposed method is used for feature extraction, that is, other methods are used in this 100-dimensional space, including LDA, LPP, NPE, NPDE and SNPE. For PCA, the dimensionality of subspace is 100. Since the rank of $\mathbf{S}_B$ is $C-1$ in (5), the final dimensionality of LDA is 39, and the other approach is 40. Finally, the nearest neighbor classifier is used for classification.

We perform 100 trials to randomly choose the training set and calculate the average recognition rates. To compare PCA, LDA, LPP, NPE, NPDE and SNPE under the condition of different training samples, we give the average classification rate curves in Fig. 4. We observe that the

recognition accuracy of each method increases when increasing the number of training samples. SNPE is always better than PCA, NPE and LPP. In addition, SNPE performs better than other methods in the case of a few training samples. Compared with NDPE, SNPE is dominant with less training samples. LDA method is only better than SNPE when the number of training samples is from 200 to 240.

C. Yale Database

The Yale face database contains 165 gray scale images of 15 individuals, and each individual has 11 images. The images demonstrate variations in lighting condition, facial expression (normal, happy, sad, sleepy, surprised, and wink). Fig. 5 shows sample images of one person.

We randomly select some samples of each individual for training, and the rest of the Yale database for testing. The number of the training samples set changes from 45 to 120. The experimental setup is the same as Section 4.2. First, we reduce dimension of the training samples to 100 by using PCA, and then other methods are used for the second dimensionality reduction. The dimension for PCA is 100, for LDA is 14, and for the other approaches are 15. The average results on 100 independent experiments are shown in Fig. 6.

Fig.5 Images of one person in Yale.

Fig. 6 depicts that the average classification rate curves of six methods with different number of training samples. From the experimental results, we can also see that the performance of supervised techniques (or LDA, NPDE, and SNPE) is always better than unsupervised techniques (or PCA, NPE and LPP). In three supervised methods, SNPE is the best one, which indicates that SNPE can make full use of the label information.

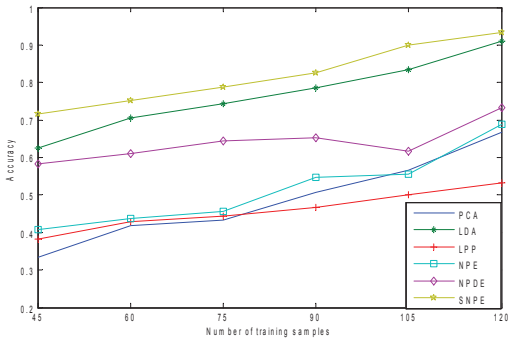


Fig.6 Results on the Yale database

V. Conclusion

This paper proposes a supervised neighborhood preserving embedding approach which can not only maintain the geometrical manifold structure but also use the label information on the given data. SPNE first builds the neighbor graph in the high-dimensional space, and then find the weight matrix of the adjacency graph. Finally, SPNE learns the projection matrix by using the label information and projects samples from the high-dimensional space into a low-dimensional space. As can be seen in the experiment, SNPE has an advantage in face recognition tasks. SNPE gives better results than other algorithm under different number of training samples. Moreover, SNPE also has a higher recognition rate when the number of training samples is not too much.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 61373093 and 61033013, by the Natural Science Foundation of Jiangsu Province of China under Grant Nos. BK2011284 and BK201222725, by the Natural Science Foundation of the Jiangsu Higher Education Institutions of China under Grant No.13KJA520001, and by the Qing Lan Project.

References

[1] J.H. Friedman, “On bias, variance, 0/1—loss, and the curse of dimensionality,” Data mining and knowledge discovery, 1997, pp: 55-77.

[2] J. Amador, “Random Projection and Orthonormality for Lossy Image Compression, ”Image and Vision Computing, 2007.

[3] R. Duda, P. Hart, and D. Stork, Pattern Classification, 2nd ed., Wiley, 2000, pp.1-15.

[4] R. A. Fisher, “The Use of Multiple Measurements in Taxonomic Problems, ”Annals of Eugenics, 1936.

[5] I .T .Jolliffe, Principal component analysis,New York: Springer- Verlag, ch. 5.

[6] J. Yang, D. Zhang, A.F. Frangi, J.Y. Yang, “Two-dimensional PCA: a new approach to appearance-based face representation and recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, pp: 131–137.

[7] B. Mikhail, N. Partha, “Laplacian eignmaps for dimensionality reduction and data representation,” Neural Computation, 2003, pp: 1373-1396.

[8] T. R. Sam, K.S. Lawrence, “Nonlinear Dimensionality Reduction by Locally Linear Embedding,” Science，2000,pp: 2323-2326.

[9] X.F. He, N. Partha, “Locality Preserving Projections,” Proceedings of the 17th Annual Conference on Neural Information Processing Systems, Vancouver, 2003, pp: 153-160.

[10] X.F. He, D. Cai, S. C, Yan et al, “Neighborhood preserving embedding,” In: Proceedings of 10th IEEE International Conference on Computer Visions[C], Washington D C, America, 2005, pp: 2208-1213.

[11] H. Murase, S.K. Nayar, “Visual learning and recognition of 3-D objects from appearance,” International Journal of Computer Vision, 1995, pp: 5–24.

[12] P.N. Belhumeur, J.P. Hespanha, D.J. Krieg man, “Eigen faces vs. Fisher faces: recognition using class specific linear projection,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, pp: 711–720.

[13] R. Chellappa, C.L. Wilson, S. Sirohey, “Human and machine recognition of faces: a survey,” Proc. IEEE, 1995, pp: 705–740.

[14] W. Zhao, R. Chellappa, A. Rosenfeld, P.J. Phillips, “Face recognition: a literature survey,” ACM Compute. Surv. 2003, pp: 399–458.

[15] A.M. Martinez, A.C. Kak, “PCA versus LDA,” IEEE Trans. Pattern

Anal. Mach. Intell, 2001, pp: 228–233.

[16] M. Turk, A. Pentland, “Eigenfaces for recognition,” J. Cognitive Neurosci, 1991, pp: 71–86.

[17] J. Yang, D. Zhang, A.F. Frangi, J-Y. Yang, “Two-dimensional PCA: a new approach to appearance-based face representation and recognition,”IEEE Trans. Pattern Anal. Mach. Intell, 2004, pp: 131–137.

[18] Y. Xu, D. Zhang, J. Yang, J.-Y. Yang, “An approach for directly extracting features from matrix data and its application in face recognition, Neurocom- putting,” 2008, pp: 1857–1865.

[19] Y. Xu, D. Zhang, J.-Y. Yang, “A feature extraction method for use with bimodal biometrics,” Pattern Recognition 2010, pp: 1106–1115.

[20] HS.Du,XL.Chai, “Face recognition method using neighborhood preserving discriminant embedding,”2010,pp: 625–629.

[21] L.S. Qiao, S.C. Chen, X.Y. Tan, “Sparsity preserving projections with applications to face recognition,” Pattern Recognition 2010, pp:331–341. January

[22] J. Gui, Z. N. Sun, “Discriminant sparse neighborhood preserving embedding for face recognition,” Pattern Recognition, 2012, pp: 2884–2893.

[23] G.F. Lu, Z. Lin, Z. Jin, “Face recognition using discriminant locality preserving projections based on maximum margin criterion,” Pattern Recognition, 2010, pp :3572–3579.

[24] A. S. Georghiades, P. N. Belhumeur, et al, “From few to many: Illumination cone models for face recognition under variable lighting and pose,” IEEE Transaction on Pattern Analysis and Machine Intelligence, 2001, pp: 643-660.