Published in IET Image Processing
Received on 20th November 2012
Revised on 2nd October 2013
Accepted on 26th October 2013
doi: 10.1049/iet-ipr.2013.0539

ISSN 1751-9659

Neighbourhood sensitive preserving embedding for pattern classification

Bing-Hui Wang[1], *Chuang Lin*[1,2], *Xue-Feng Zhao*[3], *Zhe-Ming Lu*[4]

[1]*School of Software, Dalian University of Technology, Dalian, People's Republic of China*
[2]*Department of Neurorehabilitation Engineering, University Medical Center Göttingen, Georg-August University, Göttingen, Germany*
[3]*School of Civil Engineering, Dalian University of Technology, Dalian, People's Republic of China*
[4]*School of Aeronautics and Astronautics, Zhejiang University, Hangzhou, People's Republic of China*
E-mail: linchuang_78@dlut.edu.cn

Abstract: Recently, a large family of supervised or unsupervised manifold learning algorithms that stem from statistical or geometrical theory has been designed to solve the problem of pattern classification. In this study, consider the fact that the data are usually sampled from a low-dimensional manifold space which resides in a high-dimensional Euclidean space, the authors propose a novel two-graph-based supervised linear classification algorithm called neighbourhood sensitive preserving embedding (NSPE). Different from local linear embedding (LLE) (or neighbourhood preserving embedding (NPE)) which preserves the local neighbourhood structure with one graph, NSPE can discover both the intrinsic and discriminant structure of the data manifold by constructing two graphs, that is, the within-class graph and the between-class graph. Thus, the data are mapped into a subspace where the nearby points with the same label are close to each other, whereas the nearby points with different labels are far apart. As a classification method, besides being defined on training samples, NSPE is also defined on testing samples. Experiments carried on the real-world face databases demonstrate that the results of all two-graph-based spectral methods are comparable and better than that of one-graph-based methods.

1 Introduction

In many real-world applications such as information retrieval, face recognition and data mining, one is often confronted with the high-dimensional data. However, it might be reasonable to assume that the naturally generated high-dimensional data probably lie on or near a relative lower-dimensional manifold. This motivates us to develop methods for dimensionality reduction (DR) which represent the data in a lower-dimensional space.

Classical techniques for DR are based on the assumption that the manifold is embedded linearly in the ambient space. For example, principal component analysis (PCA) [1] projects the data points along the directions of maximal variances and multi-dimensional scaling (MDS) [2] yields the same outputs as PCA by preserving the inner products between input data. When the class information is available, linear discriminant analysis (LDA) [3] can be applied to find out a linear subspace which is optimal for discrimination. However, these methods fail to discover the underlying structures of data when the original data lie on or close to a low-dimensional submanifold.

Recently, a large family of one graph-based non-linear DR algorithms has emerged for analysing high-dimensional data that lies on or near a low-dimensional manifold. The typical algorithms include isometric mapping (ISOMAP) [4], LLE [5], Laplacian eigenmap (LE) [6] and maximum variance unfolding (MVU) [7]. Although they are effective in discovering the geometrical structure of the underlying manifold, they are unsupervised in nature and fail to discover the discriminant structure of the data. Moreover, they generate mappings defined only on training samples, which make it unclear how to naturally evaluate the maps on new testing samples. Based on this fact, several methods, such as local preserving projection (LPP) [8, 9], NPE [10] and IsoP [11], have been proposed to define the maps everywhere such that they can locate any new testing sample in the reduced representation space. Although they can find the intrinsic dimensionality of the original data and perform classification for testing samples, it is necessary to point out that all these methods are unsupervised. On above basis, Maaten *et al.* [12] presents a review and systematic comparison of these DR techniques and concludes that, in the general case, non-linear methods perform better than linear ones. Furthermore, Maaten proposes a framework which can identify the weaknesses of current non-linear techniques, and suggests how the performance of non-linear DR methods may be improved. In addition, Hou *et al.* [13] indicates that the local-based approaches are often in lack of robustness and the eigenproblems that they encounter are hard to solve. Thus, he proposes a unified framework that reformulates local approaches as the semi-definite programming.

Besides one graph-based unsupervised methods, several two-graph-based supervised algorithms have also been proposed, such as LDE [14], marginal Fisher analysis (MFA) [15] (LDE and MFA are the same in essence) and local sensitive discriminant analysis (LSDA) [16]. One important thing to be noted is that these three methods inherit the key idea of LE [6] (or LPP [8]) and the neighbourhood graphs are artificially predefined. Afterwards, Raducanu and Dornaika [17] proposes an adaptive neighbourhood graphs building method named supervised Laplacian eigenmaps (S-LE), in which the neighbourhood size of each sample is adaptively learned according to data density and similarity. There is a need to illustrate that S-LE is also on the basis of LE and is a variant of LSDA. The only difference is LSDA chooses the neighbourhood size of each sample manually, whereas it is adaptively selected in S-LE.

In addition to non-orthogonal methods, a series of orthogonal methods [18–23] are also emerged when taking advantage of the orthogonality of eigenvectors.

In this paper, we propose a novel two-graph-based supervised linear classification algorithm called neighbourhood sensitive preserving embedding (NSPE). As an extension of LLE (or NPE) and a two-graph-based method, NSPE benefits from LLE for the local linear reconstruction, and LSDA (or S-LE) for the two-graph construction. NSPE shares some similar properties with MFA, LSDA and S-LE: first, all of them construct two graphs, that is, the within-class graph and the between-class graph, to discover both the local geometric structure and the discriminant structure of the data manifold. Second, all of them are performed in the supervised mode, that is, they utilise the label information of training samples during the two-graph construction. Thus, the objective function can maximise the margin of between-class samples and push the within-class samples as close as possible.

However, we should also note the main differences between NSPE and other three methods: first, different from MFA, LSDA and S-LE which inherit the key idea of one-graph-based LE (or LPP), NSPE inherits the key idea of one-graph-based LLE (or NPE). Second, the construction of the weight matrix representing affinity relationship with respect to the specific graphs is different. In MFA, LSDA and S-LE, the weight matrix is constructed by '0 – 1 weighting' or 'Gaussian kernel weighting' which is related to the previous built adjacency graph (the difference is that MFA and LSDA build the adjacency graph using parametric neighbourhood size, whereas S-LE learns the neighbourhood size of each sample adaptively). However, in NSPE, the weight matrix is accurately calculated by least-square method in the local linear reconstruction stage.

Table 1 displays the relationship among these involved graph-based spectral methods. The algorithmic procedure of NSPE is as follows: given a set of training samples, we firstly build a weight graph which describes the affinity relationship between these samples. By this way, each sample can be represented as a linear combination of its local neighbourhoods, and the combination coefficients are solved by least-square method and stored in a weight matrix. Then, the weight graph is split into two parts, that is, within-class graph and between-class graph, by using the label information of these samples. Thus, the intrinsic and discriminant structure of the samples can be accurately characterised by two graphs. Finally, we find out a projection matrix which maps the samples to a low-dimensional representation subspace and effectively preserves the local neighbourhood information, as well as the discriminant information. From the steps of NSPE below, as a whole, it can be considered as the combination of LLE for the local linear reconstruction and LSDA (or S-LE) for the two-graph construction.

Table 1 Relationship among the mentioned graph-based methods

DR	One-graph	Two-graph
LLE [5]	NPE [10]	NSPE
LE [6]	LPP [8]	MFA [15], LSDA [16], S-LE [17]

The bold with italic corresponds to our proposed method.

2 Neighbourhood sensitive preserving embedding

In this section, we will give the detail of our NSPE method. We begin with a description of the most related work, that is, LLE [5] (or NPE [10]). Then, the objective function of NSPE is given and analysed. Finally, the linear embedding process of NSPE through generalising the eigenfunction is shown.

2.1 Review of LLE

As we know, naturally generated data lie on or close to a submanifold of the ambient space. One hopes to estimate the geometrical and discriminant properties of the submanifold from random points lying on this unknown submanifold. In this paper, we concentrate on the particular problem of maximising the local margin between different classes.

Given N training samples $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\} \subset \mathbb{R}^m$ sampled from an underlying submanifold $\mathcal{M}$. Let G denote the k nearest-neighbour graph with N nodes. The ith node corresponds to sample $\boldsymbol{x}_i$. For each $\boldsymbol{x}_i$, we find its k nearest neighbours. In many cases, the samples may reside on non-linear submanifold, and it might be reasonable to assume that each local neighbourhood is linear. Thus, we can characterise the local geometry of these patches by reconstructing each sample from its neighbours with linear coefficients. Reconstruction errors are measured by the cost function [5, 10]

$$\phi(\boldsymbol{W}) = \sum_i \left\| \boldsymbol{x}_i - \sum_j W_{ij}\boldsymbol{x}_j \right\|^2 \tag{1}$$

Here, the weights W_{ij} summarise the contribution of the jth sample to the ith reconstruction. To compute the weights W_{ij}, we minimise (1) subject to the following two constraints: first, each sample $\boldsymbol{x}_i$ is reconstructed only from its neighbours (not include itself), enforcing W_{ij} be zero if $\boldsymbol{x}_j$ does not belong to this set; second, the rows of the weight matrix sum-to-one: $\sum_j W_{ij} = 1$. Obviously, the reason for the sum-to-one constraint is clear. The optimal $\boldsymbol{W}$ can then be found by solving a least-square problem using Lagrangian multiplier [5].

Now, consider the problem of mapping the samples to a line so that each mapped sample can be represented as a linear combination of its neighbours with the solved weights W_{ij} in (1). Let $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)^{\mathrm{T}}$ be such a map. In previous work [5, 10], the criterion is to minimise the following cost function

$$\phi(\boldsymbol{y}) = \sum_i \left\| y_i - \sum_j W_{ij}y_j \right\|^2 \tag{2}$$

under appropriate constraints. This cost function, like (1), is based on locally linear reconstruction errors, but here we fix W_{ij} while optimising y_i.

2.2 Objective function of NSPE

As mentioned above, in order to discover both geometrical and discriminant structure of the data, we construct two graphs, that is, within-class graph G_{w} and between-class graph G_{b}. Note here that the graph construction of NSPE is similar to that of LSDA [16]. Let $l(\boldsymbol{x}_i)$ be the class label of sample $\boldsymbol{x}_i$. For each $\boldsymbol{x}_i$, the k nearest neighbourhood set $N(\boldsymbol{x}_i)$ can be split into two parts, that is, $N_{\mathrm{w}}(\boldsymbol{x}_i)$ and $N_{\mathrm{b}}(\boldsymbol{x}_i)$. $N_{\mathrm{w}}(\boldsymbol{x}_i)$ contains k_1 neighbours sharing the same label with $\boldsymbol{x}_i$, whereas $N_{\mathrm{b}}(\boldsymbol{x}_i)$ contains k_2 neighbours having different labels with $\boldsymbol{x}_i$. Specifically

$$N_{\mathrm{w}}(\boldsymbol{x}_i) = \left\{\boldsymbol{x}_i^{(m)} | l\left(\boldsymbol{x}_i^{(m)}\right) = l(\boldsymbol{x}_i), \quad 1 \le m \le k\right\}$$

$$N_{\mathrm{b}}(\boldsymbol{x}_i) = \left\{\boldsymbol{x}_i^{(n)} | l\left(\boldsymbol{x}_i^{(n)}\right) \ne l(\boldsymbol{x}_i), \quad 1 \le n \le k\right\}$$

where $N_{\mathrm{w}}(\boldsymbol{x}_i) \cup N_{\mathrm{b}}(\boldsymbol{x}_i) = N(\boldsymbol{x}_i)$, $|N_{\mathrm{w}}(\boldsymbol{x}_i)| + |N_{\mathrm{b}}(\boldsymbol{x}_i)| = k$.

Now, think over the locally linear reconstruction errors in (2), with the purpose of handling two types of the nearest neighbourhood information, for each $\boldsymbol{x}_i$, we split its mapping y_i into two parts, that is, one part is reconstructed by $N_{\mathrm{w}}(y_i)$, which are mapped by corresponding $N_{\mathrm{w}}(\boldsymbol{x}_i)$ with coefficient $d_{\mathrm{w},i}$; the other part by $N_{\mathrm{b}}(\boldsymbol{x}_i)$, mapped by corresponding $N_{\mathrm{b}}(\boldsymbol{x}_i)$ with coefficient $d_{\mathrm{b},i}$. Thus, we have the following two reconstruction cost functions

$$\phi_{\mathrm{w}}(\mathrm{y}) = \sum_i \left\| d_{\mathrm{w},i} y_i - \sum_j W_{\mathrm{w},ij} y_j \right\|^2 \tag{3}$$

$$\phi_{\mathrm{b}}(\mathrm{y}) = \sum_i \left\| d_{\mathrm{b},i} y_i - \sum_j W_{\mathrm{b},ij} y_j \right\|^2 \tag{4}$$

where $W_{\mathrm{w},ij}$ and $W_{\mathrm{b},ij}$ are the weights of $N_{\mathrm{w}}(\boldsymbol{x}_i)$ and $N_{\mathrm{b}}(\boldsymbol{x}_i)$, satisfying $W_{\mathrm{w},ij} + W_{\mathrm{b},ij} = W_{ij}$.

Since $d_{\mathrm{w},i}$ and $d_{\mathrm{b},i}$ represent the reconstruction coefficients of y_i corresponds to $N_{\mathrm{w}}(\boldsymbol{x}_i)$ and $N_{\mathrm{b}}(\boldsymbol{x}_i)$, respectively, they obviously satisfy $d_{\mathrm{w},i} + d_{\mathrm{b},i} = 1$.

Suppose the mapping is linear, that is, $\boldsymbol{y}^{\mathrm{T}} = \boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}$, where $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{m \times N}$, then we define

$$\mathrm{z}_{\mathrm{w},i} = d_{\mathrm{w},i} y_i - \sum\nolimits_j W_{\mathrm{w},ij} y_j, \quad \mathrm{z}_{\mathrm{b},i} = d_{\mathrm{b},i} y_i - \sum\nolimits_j W_{\mathrm{b},ij} y_j$$

Above equations can then be rewritten in the vector form as

$$\mathrm{z}_{\mathrm{b}} = \boldsymbol{D}_{\mathrm{b}}\mathrm{y} - \boldsymbol{W}_{\mathrm{b}}\mathrm{y} = (\boldsymbol{D}_{\mathrm{b}} - \boldsymbol{W}_{\mathrm{b}})\mathrm{y},$$
$$\mathrm{z}_{\mathrm{w}} = \boldsymbol{D}_{\mathrm{w}} y - \boldsymbol{W}_{\mathrm{w}} y = (\boldsymbol{D}_{\mathrm{w}} - \boldsymbol{W}_{\mathrm{w}}) y$$

where $\boldsymbol{D}_{\mathrm{w}}$ and $\boldsymbol{D}_{\mathrm{b}}$ are diagonal matrices whose elements are $d_{\mathrm{w},i}$ and $d_{\mathrm{b},i}$.

Considering that $W_{\mathrm{w},ij} + W_{\mathrm{b},ij} = W_{ij}$, $\sum_j W_{ij} = 1$ and $d_{\mathrm{w},i} + d_{\mathrm{b},i} = 1$, and if we denote $d_{\mathrm{w},i} = \sum_j W_{\mathrm{w},ij}$ and $d_{\mathrm{b},i} = \sum_j W_{\mathrm{b},ij}$, then we have $D_{\mathrm{w},ii} = \sum_j W_{\mathrm{w},ij}$ and $D_{\mathrm{b},ii} = \sum_j W_{\mathrm{b},ij}$. Let $\boldsymbol{L}_{\mathrm{w}} = \boldsymbol{D}_{\mathrm{w}} - \boldsymbol{W}_{\mathrm{w}}$ and $\boldsymbol{L}_{\mathrm{b}} = \boldsymbol{D}_{\mathrm{b}} - \boldsymbol{W}_{\mathrm{b}}$, then $\boldsymbol{L}_{\mathrm{w}}$ and $\boldsymbol{L}_{\mathrm{b}}$ can be regarded as the Laplacian matrices in spectral graph theory [24]. In this paper, we name them the Laplacian matrix of the within-class graph G_{w} and Laplacian matrix of the between-class graph G_{b}, respectively.

Following some algebraic derivations, (3) reduces to

$$\begin{aligned}\phi_{\mathrm{w}}(\mathrm{y}) &= \sum_i \left\| d_{\mathrm{w},i} y_i - \sum_j W_{\mathrm{w},ij} y_j \right\|^2 = \sum_i (z_{\mathrm{w},i})^2 = \mathrm{z}_{\mathrm{w}}^T \mathrm{z}_{\mathrm{w}} \\ &= \mathrm{y}^T (\boldsymbol{D}_{\mathrm{w}} - \boldsymbol{W}_{\mathrm{w}})^T (\boldsymbol{D}_{\mathrm{w}} - \boldsymbol{W}_{\mathrm{w}}) \mathrm{y} \\ &= \mathrm{v}^T \boldsymbol{X} \boldsymbol{L}_{\boldsymbol{w}}^T \boldsymbol{L}_{\mathrm{w}} \boldsymbol{X}^T \mathrm{v} = \mathrm{v}^T \boldsymbol{X} \boldsymbol{M}_{\mathrm{w}} \boldsymbol{X}^{\boldsymbol{T}} \mathrm{v}\end{aligned}$$

where $\boldsymbol{M}_{\mathrm{w}} = \boldsymbol{L}_{\mathrm{w}}^{\mathrm{T}} \boldsymbol{L}_{\mathrm{w}}$. Similarly, we have $\phi_{\mathrm{b}}(\boldsymbol{y}) = \boldsymbol{v}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{M}_{\mathrm{b}} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{v}$ with $\boldsymbol{M}_{\mathrm{b}} = \boldsymbol{L}_{\mathrm{b}}^{\mathrm{T}} \boldsymbol{L}_{\mathrm{b}}$.

Now, we turn to the problem of mapping G_{w} and G_{b} to a line so that each sample reconstructed by the connected points in G_{w} stay as close as possible, whereas reconstructed by the connected points in G_{b} stay as far as possible. Therefore the objective function of NSPE can be finally given as

$$\begin{aligned} v^* &= \arg\min_v \frac{\phi_{\mathrm{w}}(\mathrm{y})}{\phi_{\mathrm{b}}(\mathrm{y})} = \arg\min_v \frac{v^T \boldsymbol{X} \boldsymbol{M}_{\mathrm{w}} \boldsymbol{X}^T v}{v^T \boldsymbol{X} \boldsymbol{M}_{\mathrm{b}} \boldsymbol{X}^T v} \\ &\Leftrightarrow \min \ v^T \boldsymbol{X} \boldsymbol{M}_{\mathrm{w}} \boldsymbol{X}^T \mathrm{v} \quad \text{s.t. } v^T \boldsymbol{X} \boldsymbol{M}_{\mathrm{b}} \boldsymbol{X}^T v = 1 \end{aligned} \tag{5}$$

2.3 Linear embedding of NSPE

Obviously, the numerator of (5) on G_{w} incurs a heavy penalty if neighbouring points are mapped far apart while they are actually in the same class. Likewise, the denominator of (5) on G_{b} incurs a heavy penalty if neighbouring points are mapped close together while they actually belong to different classes. Therefore the consequence of minimising (5) is to ensure that for a given sample, the nearest neighbours sharing the same label are close; meanwhile, the nearest neighbours having different labels are far apart. The vector $\boldsymbol{v}$ that minimises (5) can be solved by using Lagrange multiplier. Specifically, we first define a function $f(\boldsymbol{v})$ on $\boldsymbol{v}$ as

$$f(v) = v^T \boldsymbol{X} \boldsymbol{M}_{\mathrm{w}} \boldsymbol{X}^T v - \lambda (v^T \boldsymbol{X} \boldsymbol{M}_{\mathrm{b}} \boldsymbol{X}^T v - 1) \tag{6 − 1}$$

Then, we set the gradient with respect to $\boldsymbol{v}$ be zero and obtain the minimum eigenvalue to the following generalised eigenvector function

$$\boldsymbol{X} \boldsymbol{M} \boldsymbol{X}^T v = \lambda \boldsymbol{X} \boldsymbol{M}_{\mathrm{b}} \mathbf{X}^{\mathrm{T}} v \tag{6 − 2}$$

Let $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_d]$ be the projection matrix composed of $\boldsymbol{d}$ eigenvectors which corresponds to the first d smallest eigenvalues $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_d$ solved by (6). Finally, the optimal linear embedding is

$$\boldsymbol{x}_i \to \boldsymbol{u}_i = \boldsymbol{V}^T \boldsymbol{x}_i \tag{7}$$

3 Connections to MFA

MFA [15] is a recently proposed supervised DR method in which the intrinsic graph (similar to G_{w}) characterises the intraclass compactness, whereas the penalty graph (similar to G_{b}) connects the marginal points and characterises the interclass separability. Experimental results demonstrate its effectiveness in real-world applications. In this section, we

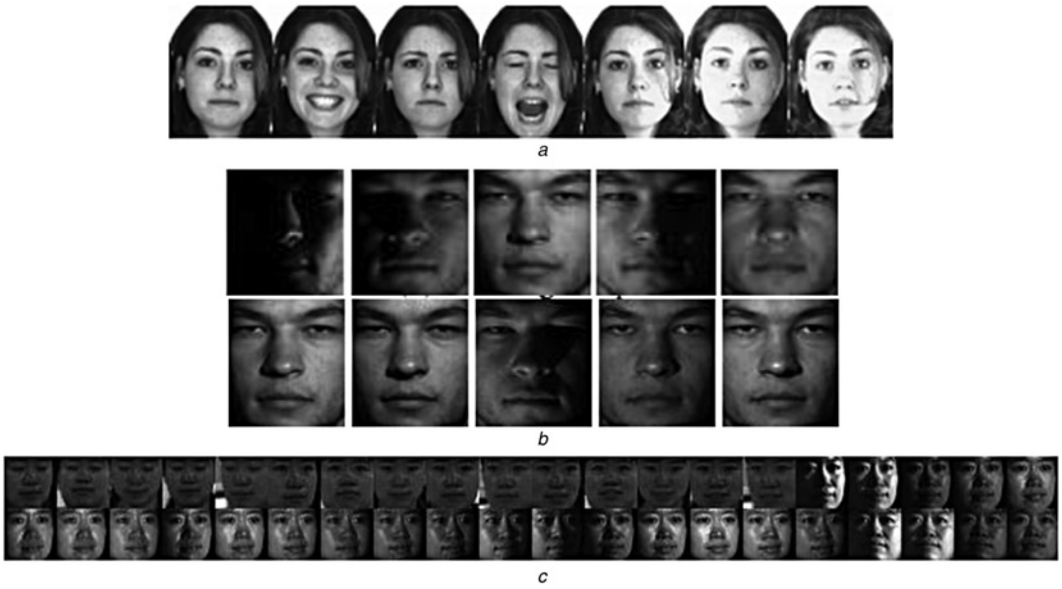

Fig. 1 *Randomly selected training samples in the database*

a AR
b E-YaleB
c PIE

discuss the connection and difference between MFA and NSPE.

Compared with (5), the objective function of MFA in [15] is given by

$$v^* = \arg\min_{v} v^T \boldsymbol{X}\boldsymbol{L}_{\mathrm{w}}\boldsymbol{X}^T v \quad \text{s.t. } v^T \boldsymbol{X}\boldsymbol{L}_{\mathrm{b}}\boldsymbol{X}^T v = 1 \tag{8}$$

where $\boldsymbol{L}_{\mathrm{w}}$ and $\boldsymbol{L}_{\mathrm{b}}$ are the Laplacian matrices of the intrinsic graph and the penalty graph, respectively. As can be seen, MFA tries to minimise $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}_{\mathrm{w}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{v}$, whereas NSPE tries to minimise $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{M}_{\mathrm{w}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{v}$ where $\boldsymbol{M}_{\mathrm{w}} = \boldsymbol{L}_{\mathrm{w}}^{\mathrm{T}}\boldsymbol{L}_{\mathrm{w}}$.

Considering that a 'vector' $\boldsymbol{f} = (f_1, f_2, \ldots, f_N)$ as a function defined on the graph such that f_i is the map of the ith node, a matrix can be considered as an operator performed on some functions defined on the graph. Under certain conditions, the following equation satisfies [6]

$$\boldsymbol{M}_{\mathrm{w}} f \simeq \frac{1}{2}\boldsymbol{L}_{\mathrm{w}}^2 f \tag{9}$$

Note that $\boldsymbol{L}_{\mathrm{w}}$ provides a discrete approximation to the 'Laplace–Beltrami operator' $\mathcal{L}$ on the manifold. Thus, the matrix $\boldsymbol{M}_{\mathrm{w}}$ can be identified as a discrete approximation to $\mathcal{L}^2$, indicating that NSPE tries to find the linear approximation to the eigenfunction of the iterated Laplacian $\mathcal{L}^2$. The eigenfunction of $\mathcal{L}^2$ is in accord with that of $\mathcal{L}$; therefore NSPE and MFA provide two different ways to approximate the eigenfunction of the 'Laplace–Beltrami operator'.

4 Experimental results

In this section, several experiments are carried out to show the performance of NSPE and related methods for pattern classification. We compare it with PCA [1], LDA [3], NPE [10], sNPE [25], MFA [15], LSDA [16] and S-LE [17] to execute face recognition on three representative facial databases: AR [26], extended YaleB (E-YaleB) [27, 28] and pose illumination, and expression (PIE) [29]. To reduce computational complexity, face images are resized to 32 × 32. Some sample images from the same individual are shown in Fig. 1. In each database, the image set is partitioned into several gallery and probe sets. For ease of representation, *Gm/Pn* indicates *m* images per person are selected for training and the remaining *n* images for testing.

In all experimental results, we firstly report the optimal average recognition rates over ten random splits (we record ten optimal results and then average them) to see all methods' ability dealing with training sets with different *Gm*s. Secondly, in order to analyse the numerical stability of these methods, we show the mean, as well as the standard deviation (mean ± std.–dev.) of the recognition rates. Thirdly, to gain an insight into the intrinsic dimension, we plot the 'average recognition rates' against feature dimensions with different *Gm*s. Since the dimension of the image is much larger than the number of training samples, all methods involve a PCA transformation. In this paper, we select the number of principal axes as 200 in the PCA phase in all databases. Then, the results are obtained cover the PCA transformed subspace. Here, the feature dimension is chosen from 0 to 200 with a gap 10. We choose the number 200 is because we observe that after a specific dimension (<200) the result associated with each method becomes stable. We conclude that this specific dimension is the intrinsic dimension with respect to the corresponding method. Furthermore, we apply the nearest-neighbour classifier for its simplicity. Euclidean metric is used as the distance measure.

4.1 Experiments on AR database

AR dataset [26] consists of over 3000 frontal face images of 126 individuals. There are 26 images of each individual, taken at two different occasions. The faces in AR contain variations such as illumination change, expressions and facial disguises. We randomly selected 100 subjects (50 male and 50 female) for our experiments. For each subject, we randomly permute 14 images. Firstly, to evaluate all methods' ability to dealing with small training set with different *Gm*s, a random subset with $Gm(=G3, G5, G7)$ is taken with labels to form the training set, respectively. The corresponding remaining part $Pn(=P11, P9, P7)$ with labels is the testing set. Here, I would like to explain that learning from small training set is a relative hard problem when compared with big training set. Therefore, in each experiment, we choose the maximum *Gm* as half of the number of each person.

He *et al.* [16, 30] indicate that the classification performance is stable with respect to the neighbourhood size k only if its value is not so small when compared with Gm. Here, the parameter k is chosen as 5 in NPE, sNPE and 10 in LSDA, MFA and NSPE as they construct two graphs (k_1 and k_2 are automatically chosen via pairwise affinity weights).

Table 2 shows the 'optimal average recognition rates' obtained by PCA, LDA, NPE, sNPE, LSDA, MFA, S-LE and NSPE with the corresponding feature dimensions in different *Gm*s over ten random splits ('Note that the dimension of the optimal result in each running is almost the same, and we finally set the optimal dimension of the optimal average result using the voting rule. We exploit the same technique on E-YaleB and PIE'.). From it, we see that NSPE, together with LSDA, S-LE and MFA are comparable when obtaining the optimal result no matter what *Gm* is. It is the fact that all of them are two-graph based, that is, not only do they take into account the discriminant power, but also the locality preserving power of the original data. What is more, in this case, LSDA and S-LE have almost the same results. The reason is that S-LE is a variant of LSDA by selecting the parameter k adaptively. However, as illustrated by He and Cai [30], k is not so sensitive to affecting the performance. Incorporated into the 'supervised' property, sNPE is naturally better than NPE. Even, sNPE can be comparable with MFA in the 'G7/P7' case (Note that sNPE just considers the label of the original data, but not maximise the margin between data with different labels, so it is one graph-based method in essence.). On the other hand, seen in the 'G5/P9' and 'G7/P7' cases, we observe that LDA can also gain very high performance. A proper explanation is that LDA can

Table 2 Optimal average recognition rates (dimensions) on AR database with different '*Gm/Pn*'s of 100 individuals

Methods	G3/P11	G5/P9	G7/P7
PCA	0.620 (150)	0.630 (180)	0.710 (120)
LDA	0.620 (190)	0.731 (100)	0.870 (150)
NPE	0.611 (170)	0.630 (200)	0.710 (150)
sNPE	0.631 (170)	0.632 (100)	0.866 (150)
LSDA	0.691 (150)	0.740 (130)	0.878 (200)
S-LE	0.691 (150)	0.750 (100)	0.878 (170)
MFA	0.687 (190)	0.731 (150)	0.864 (130)
NSPE	0.707 (170)	0.740 (180)	0.873 (150)

Numbers in parentheses are the corresponding feature dimensions with the optimal results. The bold with italic and the bold only, respectively, mean the best and the second best results among all methods. Tables 3–7 use the same way.

Table 3 Recognition rates% on AR database with the mean and standard deviation (mean ± std. – dev.) with different '*Gm/Pn*'s of 100 individuals

Methods	G3/P11	G5/P9	G7/P7
PCA	55.30 ± 0.64	57.65 ± 0.30	67.85 ± 0.36
LDA	55.75 ± 0.58	64.30 ± 2.00	76.90 ± 2.80
NPE	50.35 ± 1.10	56.85 ± 0.39	67.85 ± 0.42
sNPE	51.40 ± 1.26	58.25 ± 0.73	76.55 ± 3.02
LSDA	61.30 ± 1.38	67.30 ± 1.31	78.50 ± 2.68
S-LE	61.20 ± 1.22	67.70 ± 1.15	77.26 ± 2.62
MFA	60.95 ± 1.18	66.85 ± 0.43	76.77 ± 2.62
NSPE	61.35 ± 1.15	67.35 ± 0.41	77.71 ± 2.81

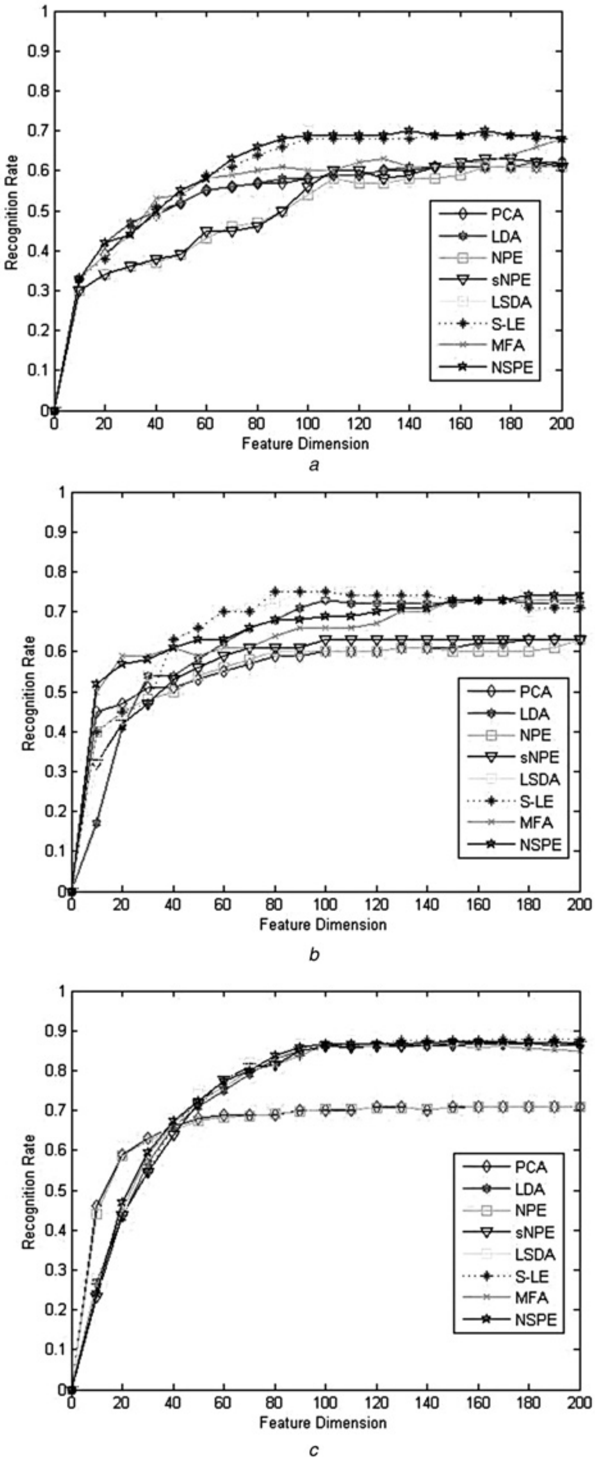

Fig. 2 *Average recognition results in AR database against feature dimensions from 0 to 200 with a gap 10 given different 'Gm/Pn's of 100 individuals*

From Figs. 2*a–c* are the results of *G*3/*P*11, *G*5/*P*9 and *G*7/*P*7 own better discriminant power when the data distribution satisfies its demand to some extent. As a whole, PCA performs the worst because neither does it consider the discriminant power nor the locality preserving power of the original data.

The results over numerical stability are shown in Table 3. Seen from it, NSPE has high mean values and relative low standard deviations, revealing that it also has a very stable solution. Besides, MFA has very similar standard deviations with NSPE allowing for the close relationship between them. Furthermore, benefitting from the adaptivity, S-LE is more stable than LSDA. Although PCA is the most stable, it achieves the worst mean values.

The average recognition rates against feature dimensions of all methods with different *Gm*s are plotted in Fig. 2. From Figs. 2*a*–*c*, we learn that NSPE can fast get close to the intrinsic dimension (about 110), which demonstrates that it finds the intrinsic structure of the data. Again, as mentioned above, since LSDA, S-LE, MFA and NSPE are two-graph-based methods, they are comparable in the whole process whatever *Gm* is and better than other one graph-based methods.

4.2 Experiments on E-YaleB database

E-YaleB dataset [18, 19] consists of 2414 frontal face images of 38 subjects under 9 poses and 64 illumination conditions. They are captured under various lighting conditions and cropped and normalised to 32 × 32 pixels. For our experiment, we randomly take 1262 images of 20 subjects. A random subset with $Gm(=G16, G24, G32)$ is taken with labels to form the training set, and the remaining part $Pn(=P48, P40, P32)$ as the testing set. As *Gm* is bigger than that in AR database, so the neighbourhood size *k* is set bigger accordingly. Here, *k* is selected as 10 in NPE, sNPE, and 15 in LSDA, MFA and NSPE.

Table 4 Optimal average recognition rates (dimensions) on E-Yale database with different '*Gm/Pn*'s of 20 individuals

Methods	G16/P48	G24/P40	G32/P32
PCA	0.630 (160)	0.691 (190)	0.740 (170)
LDA	0.710 (180)	0.740 (190)	0.761 (140)
NPE	0.690 (180)	0.751 (170)	0.780 (190)
sNPE	0.720 (190)	0.780 (140)	0.830 (190)
LSDA	0.710 (170)	0.801 (150)	0.830 (200)
S-LE	0.710 (120)	0.791 (180)	0.830 (200)
MFA	0.740 (170)	0.780 (150)	0.841 (200)
NSPE	0.731 (170)	0.801 (160)	0.841 (200)

Table 5 Recognition rates% on E-YaleB database with the mean and standard deviation (mean ± std.–dev.) with different '*Gm/Pn*'s of 20 individuals

Methods	G16/P48	G24/P40	G32/P32
PCA	$62.25 \pm 1 \times 10^{-6}$	$67.15 \pm 1 \times 10^{-6}$	$71.35 \pm 1 \times 10^{-6}$
LDA	62.80 ± 1.21	67.20 ± 1.04	74.90 ± 0.22
NPE	67.45 ± 0.12	69.75 ± 0.21	71.50 ± 0.62
sNPE	66.80 ± 0.34	71.20 ± 0.66	76.55 ± 0.86
LSDA	67.60 ± 0.24	71.60 ± 1.06	71.85 ± 1.42
S-LE	67.50 ± 0.32	72.45 ± 0.87	72.85 ± 1.60
MFA	68.95 ± 0.46	70.20 ± 0.82	76.70 ± 0.74
NSPE	68.70 ± 0.44	71.85 ± 0.67	76.65 ± 0.78

Tables 4 and 5, respectively, show the 'optimal average recognition rates' and results over 'numerical stability' with different *Gm*s. The average recognition rates against feature dimensions are shown in Fig. 3. Similar to the results in AR database, no matter what *Gm* is, the best and second best results are achieved among two-graph-based LSDA, S-LE, MFA and our NSPE. Testing on this database, sNPE

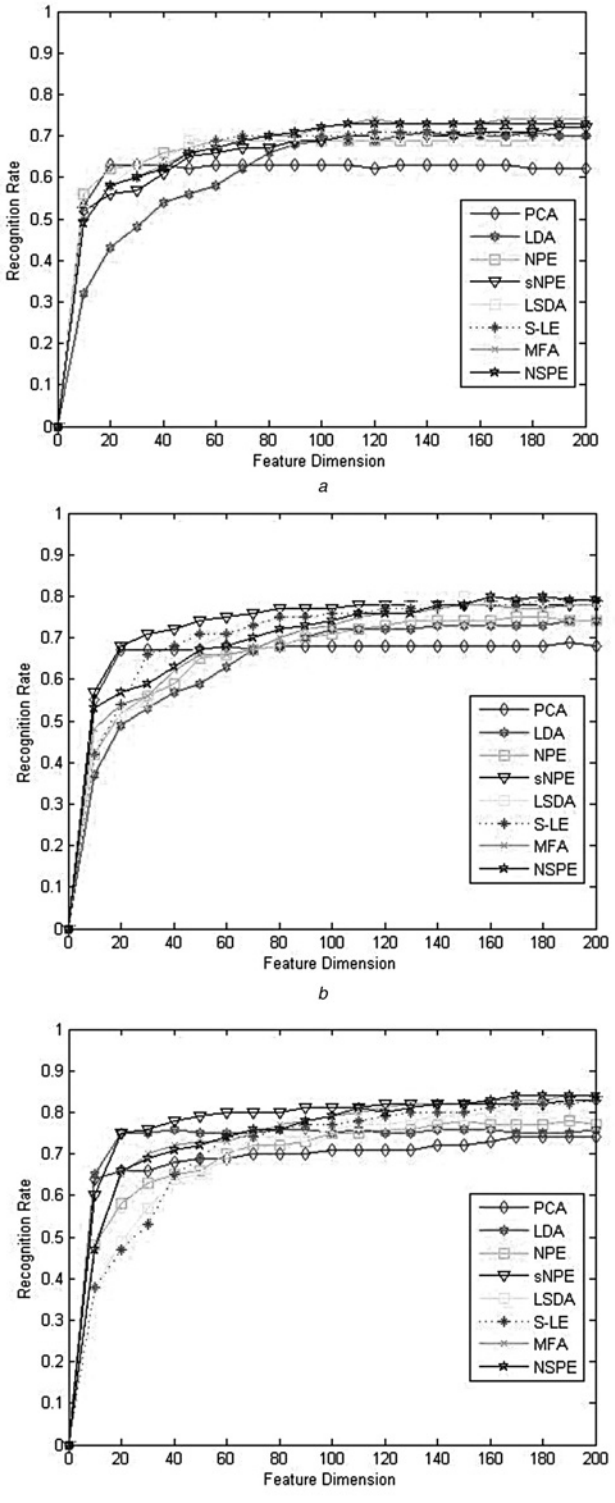

Fig. 3 *Average recognition results on E-YaleB database against feature dimensions from 0 to 200 with a gap 10 given different 'Gm/Pn's of 20 individuals*

From Figs. 3*a*–*c* are the results of *G6*/*P48*, *G24*/*P40* and *G32*/*P32* can also gain a very high performance. Seen from Table 4, when taking all cases into account, NSPE has relatively more stable results than LSDA and S-LE. The performance of MFA is the closest to that of NSPE. Meanwhile, LSDA and S-LE also have very similar results, further demonstrating that S-LE is nearly the same as LSDA, except for its adaptivity for choosing k. Note that although PCA and NPE have top 2 stable results, their recognition results are not so good.

In addition, different from results in AR database, one can find some interesting things from Fig. 3: firstly, when Gm is small (‘G16/P48’ and ‘G24/P40’) with feature dimension < 80 from Figs. 3*a* and *b*, PCA performs better than LDA, which is in line with what Martinez proved in [31]. Secondly, all these methods are able to discover the intrinsic dimension quickly whatever Gm is. In this case, we can arrive at the conclusion that the intrinsic dimension is about 110.

4.3 Experiments on PIE database

PIE dataset [20] consists of 41 368 images of 68 people, each person under 13 different poses, 43 different illumination conditions and with 4 different expressions. We select a subset of 1700 images of ten subjects that only contains five near frontal poses (C05, C07, C09, C27, C29). Therefore, there are nearly 170 images for each individual. In the experiment, a random subset with $Gm(=G45, G65, G85)$ is taken with labels to form the training set, and the remaining part nearly $Pn(=P125, P105, P85)$ as the testing set. The parameter k is chosen as 10 in NPE and sNPE. In LSDA, MFA and NSPE, it is 15.

Table 6 Optimal average recognition rates (dimensions) on PIE database with different ‘*Gm/Pn*’s of ten individuals

Methods	G45/P125	G65/P105	G85/P85
PCA	0.788 (190)	0.825 (200)	0.855 (200)
LDA	0.900 (150)	0.911 (150)	0.914 (140)
NPE	0.839 (150)	0.839 (150)	0.885 (200)
sNPE	0.896 (180)	0.899 (140)	0.899 (140)
LSDA	0.892 (200)	0.930 (170)	0.929 (200)
S-LE	0.892 (200)	0.930 (180)	0.929 (190)
MFA	0.884 (150)	0.930 (200)	0.930 (170)
NSPE	0.883 (180)	0.930 (190)	0.930 (170)

Table 7 Recognition rates% on PIE database with the mean and standard deviation (mean ± std.–dev.) with different ‘*Gm/Pn*’s of ten individuals

Methods	G45/P125	G65/P105	G85/P85
PCA	73.06 ± 0.24	80.85 ± 0.14	81.81 ± 0.28
LDA	86.53 ± 0.11	88.25 ± 0.15	89.34 ± 0.12
NPE	81.34 ± 0.23	81.34 ± 0.23	84.42 ± 0.24
sNPE	85.68 ± 0.33	87.70 ± 0.32	86.64 ± 0.22
LSDA	84.89 ± 0.22	90.53 ± 0.20	89.62 ± 0.26
S-LE	85.23 ± 0.21	89.38 ± 0.32	90.05 ± 0.27
MFA	85.09 ± 0.36	89.72 ± 0.31	90.21 ± 0.16
NSPE	84.99 ± 0.36	89.73 ± 0.31	90.18 ± 0.16

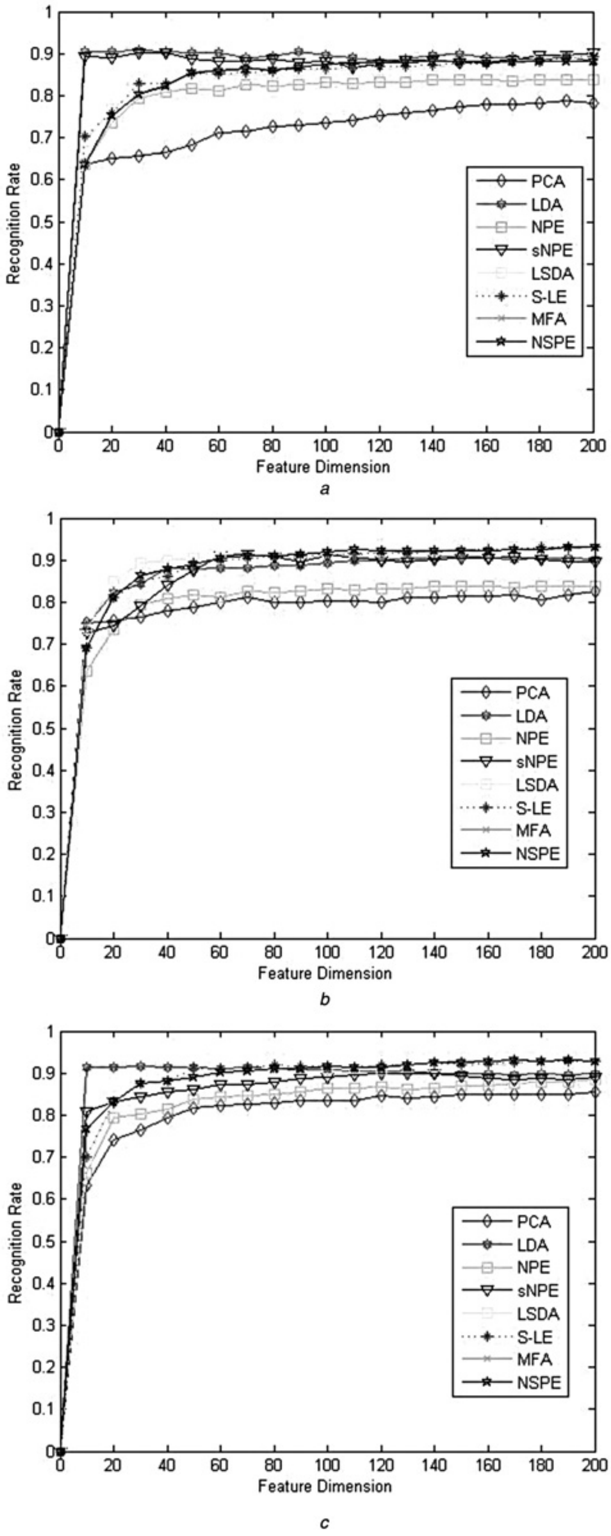

Fig. 4 *Average recognition results on PIE database against feature dimensions from 0 to 200 with a gap 10 given different ‘Gm/Pn’s of ten individuals*

From Figs. 4*a–c* are the results of *G*45/*P*125, *G*65/*P*105 and *G*85/*P*85

The optimal average recognition rates and numerical stability results are displayed in Tables 6 and 7. Fig. 4 reports the ‘average recognition rates’. Same as the results on AR and E-YaleB databases, LSDA, S-LE, MFA and NSPE are comparable and achieve very high optimal results. However, there is an exception in the ‘G45/P125’ case where the best and the second best results are LDA and sNPE, respectively. As mentioned above, a reasonable explanation is that LDA can own better discriminant power when the data distribution satisfies its demand. Moreover, high performance of sNPE is owing to its incorporated ‘supervised’ property. An interesting thing is that the stability is in line with the optimal results to a large extent. Moreover, we should note some other important and interesting points: firstly, the optimal results, the stability and the average results of MFA and NSPE are almost the same in all cases in this database. This just in time interprets the relationship between MFA and NSPE, that is, they provide two different ways to approximate the eigenfunction of the Laplace–Beltrami operator. Secondly, from Figs. 4*a*–*c* and Table 6, we see that the optimal results of all methods do not change much as *Gm* increases. This is due to the fact that our chosen subset has a relative small number of classes when compared with the number of training samples of each class. Thirdly, once again, the results of LSDA and S-LE are very close.

5 Conclusions and future work

In this paper, we have proposed a new two-graph-based supervised linear classification method called NSPE for face recognition. Based on spectral techniques, NSPE construct a within-class graph and a between-class graph to discover the local geometric structure and the discriminant structure of the data manifold, respectively. With the two-graph, NSPE can maximise the margin of between-class samples and minimise the margin of within-class samples. Moreover, as a classification method, NSPE is defined everywhere, so testing samples can also be mapped into the low-dimensional space.

As LSDA, S-LE, MFA and NSPE all construct two-graph, they share some similar properties in discovering the intrinsic structure of the face manifold. Experimental results demonstrate that they have comparable recognition rates and gain better results than one-graph-based PCA, LDA, NPE and sNPE. Furthermore, we can come to the following conclusions: firstly, the almost same results and stability between NSPE and MFA verify that they just provide two different ways to approximate the eigenfunction of the Laplace–Beltrami operator. Secondly, the very similar results between LSDA and its variant S-LE prove that the neighbourhood size parameter is not so sensitive to affecting the performance.

However, one should also note the differences among them: firstly, MFA, LSDA and S-LE inherit from one-graph-based LE (or LPP), whereas NSPE inherits the key idea of one-graph-based LLE (or NPE); theoretically, it can preserve the intrinsic structure better. Secondly, the construction of the weight matrix representing affinity relationship with respect to the specific graphs among them is different. As a whole, NSPE can be considered as the combination of LLE and LSDA.

Several problems still remain unclear and should be investigated in future work. Firstly, as the discriminative information and the intrinsic property are equally important, how can we distinguish and represent the difference between the discriminating power and locality preserving power theoretically, and then integrate them? Secondly, since both NSPE and MFA try to linearly approximate the eigenfunction of the Laplace–Beltrami operator on the manifold, it is still unclear how to evaluate them in theory. More precisely, it is unclear how to define the optimal graph structure which can provide the best discrete approximation to $\mathcal{L}$. Thirdly, the face data spaces are assumed on the same manifold. However, the data space may be disconnected and different manifolds may have different dimensionality. It is still unknown how often such a case may occur and how to deal with it. Finally, NSPE is linear, and it can also be performed in the reproducing kernel Hilbert space, orthogonal space and tensor space, giving rise to non-linear maps, orthogonal maps and multi-linear maps, respectively. Thus, the performance of kernel NSPE, orthogonal NSPE and tensor NSPE need to be further evaluated.

6 Acknowledgments

Our work is supported by the Natural Science Foundation of China under grant nos. 61033012, 611003177 and 61272371, Fundamental Research Funds for the Central Universities, no. DUT12JR07, and specialised research fund for the Doctoral Program of Higher Education, no. 20120041120046. We especially thank for the support of the European Research Council through the Advanced Research Grant DEMOVE (contract #267888).

7 References

1 Turk, M.A., Pentland, A.P.: 'Face recognition using eigenfaces'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, 1991, pp. 586–591

2 Cox, T., Cox, M.: 'Multidimensional scaling' (Chapman & Hall, 1994)

3 Belhumeur, P., Hespanha, J., Kriegman, D.: 'Eigenfaces vs. Fisherfaces: recognition using class specific linear projection', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1997, **19**, (7), pp. 711–720

4 Tenenbaum, J., de Silva, V.: 'A global geometric framework for nonlinear dimensionality reduction', *Science*, 2000, **290**, (5500), pp. 2319–2323

5 Roweis, S., Saul, L.: 'Nonlinear dimensionality reduction by locally linear embedding', *Science*, 2000, **290**, (5500), pp. 2323–2326

6 Belkin, M., Niyogi, P.: 'Laplacian eigenmaps for dimensionality reduction and data representation', *Neural Comput.*, 2003, **15**, (6), pp. 1373–1396

7 Weinberger, K., Saul, L.: 'An introduction to nonlinear dimensionality reduction by maximum variance unfolding'. AAAI Conf. Artificial Intelligence, 2006, vol. 2, pp. 1683–1686

8 He, X., Niyogi, P.: 'Locality preserving projections', *Adv. Neural Inf. Process. Syst., Camb.*, 2003, **16**, pp. 153–160

9 He, X., Yan, S., Hu, Y.: 'Face recognition using Laplacianfaces', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2005, **27**, (3), pp. 328–340

10 He, X., Cai, D., Yan, S.: 'Neighborhood preserving embedding'. IEEE Int. Conf. Computer Vision, Beijing, 2005, vol. 2, pp. 1208–1213

11 Cai, D., He, X., Han, J.: 'Isometric projection'. AAAI Conf. Artificial Intelligence, Vancouver, 2007, vol. 1, pp. 528–533

12 Maaten, L., Postma, E., Herik, H.: 'Dimensionality reduction: a comparative review', *J. Mach. Learn. Res.*, 2009, **20**, (1), pp. 1–41

13 Hou, C., Zhang, C., Wu, Y.: 'Stable local dimensionality reduction approaches', *Pattern Recognit.*, 2009, **42**, (6), pp. 2054–2066

14 Chen, H., Chang, H., Liu, T.: 'Local discriminant embedding and its variants'. IEEE Conf. Computer Vision and Pattern Recognition, 2005

15 Yan, S., Xu, D., Zhang, B., Zhang, H.: 'Graph embedding and extensions: a general framework for dimensionality reduction', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2007, **29**, (1), pp. 40–51

16 Cai, D., He, X., Zhou, K., *et al.* 'Locality sensitive discriminant analysis'. Int. Joint Conf. Artificial Intelligence, Hyderabad, 2007, pp. 708–715

17 Raducanu, B., Dornaika, F.: 'A supervised non-linear dimensionality reduction approach for manifold learning', *Pattern Recognit.*, 2012, **45**, (6), pp. 2432–2444

18 Cai, D., He, X.: 'Orthogonal laplacianfaces for face recognition', *IEEE Trans. Image Process.*, 2006, **15**, (11), pp. 3608–3614

19 Kokiopoulou, E., Saad, Y.: 'Orthogonal neighborhood preserving projections: a projection-based dimensionality reduction technique', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2007, **29**, (12), pp. 2143–2156

20 Liu, X., Yin, J., Feng, Z.: 'Orthogonal neighborhood preserving embedding for face recognition'. Int. Conf. Image Processing, 2007

21 Jin, Y., Ruan, Q., Wu, J.: 'Gabor-based orthogonal locality sensitive discriminant analysis for face recognition'. Int. Conf. Signal Processing, 2009, pp. 1625–1628

doi: 10.1049/iet-ipr.2013.0539

22 Khushaba, R.: ‘Orthogonal fuzzy neighborhood discriminant analysis for multifunction myoelectric hand control’, *IEEE Trans. Biomed. Eng.*, 2010, **25**, (6), pp. 1410–1419

23 Gui, J., Sun, Z., Jia, W., *et al.*: ‘Discriminant sparse neighborhood preserving embedding for face recognition’, *Pattern Recognit.*, 2012, **45**, (8), pp. 2884–2893

24 Chung, F.: ‘Spectral graph theory’, *Am. Math. Soc.*, 1997, (92), pp. 1–207

25 Zeng, X., Luo, S.: ‘A supervised subspace learning algorithm: supervised neighborhood preserving embedding’. Int. Conf. Advanced Data Mining and Applications, 2007, pp. 81–88

26 Martinez, A., Benavente, R.: ‘The AR face database’. Technical Report, no. 24, 1998

27 Georghiades, A., Belhumeur, P.: ‘From few to many: illumination cone models for face recognition under variable lighting and pose’, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2001, **23**, (6), pp. 643–660

28 Lee, K., Ho, J., Kriegman, D.: ‘Acquiring linear subspaces for face recognition under variable lighting’, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2005, **27**, (5), pp. 684–698

29 Sim, T., Baker, S., Bsat, M.: ‘The CMU pose illumination, and expression (PIE) database’. Int. Conf. Automatic Face and Gesture Recognition, Washington, DC, USA, 2002, pp. 46–51

30 He, X., Cai, D.: ‘Laplacian regularized gaussian mixture model for data clustering’, *IEEE Trans. Knowl. Data Eng.*, 2011, **23**, (9), pp. 1406–1418

31 Martinez, A., Kak, A.: ‘PCA versus LDA’, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2001, **23**, (2), pp. 228–233

doi: 10.1049/iet-ipr.2013.0539