



Local maximal margin discriminant embedding for face recognition



Pu Huang^{a,*}, Zhenmin Tang^a, Caikou Chen^b, Zhangjing Yang^a

^a School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

^b College of Information Engineering, Yangzhou University, Yangzhou 225009, China

ARTICLE INFO

Article history:

Received 10 November 2012

Accepted 24 November 2013

Available online 7 December 2013

Keywords:

Local maximal margin discriminant embedding

Locality preserving projection

Maximum margin criterion

Small sample size problem

Local structure

Appearance-based

Dimensionality reduction

Manifold learning

Face recognition

ABSTRACT

In this paper, a manifold learning based method named local maximal margin discriminant embedding (LMMDE) is developed for feature extraction. The proposed algorithm LMMDE and other manifold learning based approaches have a point in common that the locality is preserved. Moreover, LMMDE takes consideration of intra-class compactness and inter-class separability of samples lying in each manifold. More concretely, for each data point, it pulls its neighboring data points with the same class label towards it as near as possible, while simultaneously pushing its neighboring data points with different class labels away from it as far as possible under the constraint of locality preserving. Compared to most of the up-to-date manifold learning based methods, this trick makes contribution to pattern classification from two aspects. On the one hand, the local structure in each manifold is still kept in the embedding space; one the other hand, the discriminant information in each manifold can be explored. Experimental results on the ORL, Yale and FERET face databases show the effectiveness of the proposed method.

© 2013 Elsevier Inc. All rights reserved.

1. Introduction

Face recognition has attracted wide attention of the researchers in the fields of pattern recognition and computer vision because of its immense application potential. Many face recognition methods have been developed over the past few decades. One of the most successful and well-studied techniques to face recognition is the appearance-based method. In an appearance-based technique, a two-dimensional face image of size w by h pixels is represented by a vector in a $w \times h$ -dimensional space. In practice, however, these $w \times h$ -dimensional spaces are too large to allow robust and fast recognition. A common way to attempt to resolve this problem is to use dimensionality reduction techniques. Two of the most popular dimensionality reduction methods are principal component analysis (PCA) [1] and linear discriminant analysis (LDA) [2].

PCA is a classical dimensionality reduction and data representation technique widely used in pattern classification and visualization tasks. PCA is an unsupervised method, which aims to find a linear mapping that preserves the total variance by maximizing the trace of feature variance. The optimal mapping is the leading eigenvectors corresponding to the largest eigenvalues of the covariance matrix for data of all classes.

LDA produces an optimally discriminative projection for certain cases. LDA searches for the transformation that maximizes the between-class scatter and at the same time minimizes the within-class scatter. Different from PCA which is completely unsupervised with regard to the class information of the data, LDA takes full consideration of the class labels and it is generally believed that LDA is able to enhance class separability. Despite the success of the LDA algorithm in many applications, its effectiveness is still limited since, in theory, the number of available projection directions is lower than the class number. Furthermore, class discrimination in LDA is based upon within-class and between-class scatters, which is optimal only in cases where the data of each class is approximately Gaussian distributed, a property that cannot always be satisfied in real-world applications. At the same time, LDA cannot be applied directly to small sample size problem [3] because the within-class scatter matrix is singular [2]. To avoid the singularity problem of LDA, Li et al. [4] used the difference of both between-class scatter and within-class scatter as discriminant criterion, called maximum margin criterion (MMC). MMC has the advantages of effectiveness and simplicity.

Recent studies [5–7] have shown that the high-dimensional data possibly resides on a nonlinear sub-manifold. However, both PCA and LDA effectively see only the global Euclidean structure. When they are applied to face recognition, they fail to discover the underlying structure, if the face images lie on a nonlinear sub-manifold hidden in the image space. Some nonlinear

* Corresponding author.

E-mail address: huangpu3355@163.com (P. Huang).

techniques have been proposed to discover the nonlinear structure of the manifold. The basic assumption of manifold learning is that the input data lie on a smooth low-dimensional manifold. Each manifold learning based method attempts to preserve a different geometrical property of the underlying manifold. The representative ones include Isomap [5], LLE [6], Laplacian Eigenmap [7] and local tangent space alignment (LSTA) [8]. These nonlinear methods do yield impressive results on some benchmark artificial data sets. However, they yield maps that defined only on the training data points and how to evaluate the maps on novel test data points remains unclear. To overcome this limitation, He et al. extended Laplacian Eigenmap to its linearized version, i.e. locality preserving projection (LPP) [9–13] for an explicit map. LPP attempts to construct a nearest neighbor graph and then evaluate the low-dimensional embedding to best preserve local structure of the data set.

Although LPP is effective in many domains, it is unsupervised and its unsupervised nature restricts its discriminating capability. To consider class label information in LPP, several supervised LPP methods [14–21] have been developed. Local discriminant embedding (LDE) [15] and marginal fisher analysis (MFA) [16], whose objective functions are very similar, can also be viewed as supervised LPP methods. This is because their training phases both exploit the class label information of samples. They are derived by using a motivation partially similar to LPP and each of them is based on an eigen-equation formally similar to the eigen-equation of LPP. On the other hand, since LDE and MFA partially borrow the idea of discriminant analysis and try to produce satisfactory linear separability, their ideas are also somewhat different from the idea of preserving the local structure of LPP. LDE and MFA can be viewed as two combinations of the locality preserving technique and the linear discriminant analysis [22]. Compared with LDA, both LDE and MFA do not depend on the assumption that the data of each class is Gaussian distributed and can obtain more available projection directions and better characterize the separability of different classes.

The purpose of LPP is to preserve the proximity relationship of the input data. In LPP, by applying k nearest neighbor (k -NN) criterion, any point and its k nearest neighbors are viewed as located on a super-plane, where all the descriptions in linear space can be performed. A common problem with the classical LPP and several supervised LPP methods [14,17,18] is that they might not necessarily discover the most discriminative manifold for pattern classification tasks because the manifold learning is originally modeled based on a characterization of “locality”, a model that has no direct connection to classification. This is unproblematic for existing LPP algorithms as they seek to model a simple manifold, for example, to recover an embedding of one person’s face images. In face recognition each person forms his or her own manifold in the feature space [23]. If one person’s face images do exist on a manifold, different persons’ face images could lie on different manifolds. If the images needed to be classified reside on multi-manifolds and two or more models have a common axis, then the locality preserving algorithms of manifold learning may result in overlapped embedding belonging to different classes because to recognize faces it would be necessary to distinguish between images from different manifolds. This problem is referred to as “overlearning of locality” [24].

In order to solve the problem of “overlearning of locality”, Yang et al. proposed an unsupervised discriminant projection (UDP) [25] method, which can be viewed as simplified LPP on the assumption that the local density is uniform [26]. In the proposed method, locality and non-locality are discussed in detail, where locality means the sum of the squared distance between the points in k nearest neighbors, and the non-locality denotes the sum of the squared distance between two points not belonging to any k nearest neighbors. In order to achieve a discriminative map, UDP aims to find a linear transformation that maximizes the ratio of the

non-locality to the locality. In the literature [27], there is another algorithm named locally preserving and globally discriminant projection with prior information (LPGDP) introduced to address this problem. The LPGDP method utilizes prior misclassification rate of between-class in the training data for the global discriminant measure while using class labels for preserving locality. Besides, Li et al. proposed a linear multi-manifolds learning based approach called constrained maximum variance mapping (CMVM) [28]. CMVM aims at globally maximizing the distances between different manifolds. After the local scatters have been characterized, the CMVM algorithm focuses on developing a linear transformation that maximizes the dissimilarities between all the manifolds under the constraint of locality preserving.

As discussed above, when LPP is used to map the high-dimensional data into a low-dimensional feature space, it may produce high between-class overlaps because of the “overlearning of locality”. To solve this problem, the methods including UDP, LPGDP and CMVM seek to find a transformation that separates different manifolds after the local structure has been characterized. It is unproblematic for these methods to effectively separate different classes when the data distributed on a manifold have the same label. However, in practice, the local scatter is usually constructed according to the k -NN criterion, which will bring another problem. It is that, when there is large variation within the same class, the within-class variation may be larger than the between-class variation, which means that the neighbor relationship measured by the k -NN criterion may be distorted. In other words, data samples residing on a manifold possibly have different labels. In this case, these methods may not work well because of their common assumption that the data distributed on a manifold have the same label.

In this paper, we propose an effective supervised manifold learning algorithm, called local maximal margin discriminant embedding (LMMDE) for feature extraction and recognition. The proposed algorithm LMMDE incorporates LPP and MMC for data analysis. Similar to MFA, LMMDE characterizes intra-class compactness and inter-class separability to maximize the margins between different classes. One difference between MFA and the proposed method lies that MFA neglects the local structure based on the overall samples which may be helpful for classification. In addition, both CMVM and LMMDE have the common purpose that is to take class label information into account based on the property of locality preserving, but they are essentially different because: (1) CMVM is originally designed to separate different manifolds based on the assumption that the data distributed on a manifold have the same label, while LMMDE is designed to reduce the between-class overlaps based on the assumption that the data distributed on a manifold may have different labels and (2) CMVM characterizes only the inter-class separability in a global way, while LMMDE measures both the inter-class separability and the intra-class compactness in a local way like MFA.

The rest of this paper is structured as follows: In Section 2, the PCA, LDA, LPP are briefly reviewed. Section 3 describes the proposed algorithm in detail. In Section 4 the proposed algorithm is examined on three data sets and the experimental results are offered. Section 5 finishes this paper with some conclusions.

2. Outline of PCA, LDA, LPP

Let us consider a set of n samples $\{x_1, \dots, x_n\}$ takes values in an N -dimensional image space, and assume that each image belongs to one of C classes. Let us also consider a linear transformation that maps the original N -dimensional space into a d -dimensional feature space, where $N > d$. The new feature vectors in the d -dimensional space are defined by the following linear transformation:

$$y_k = A^T x_k, \quad k = 1, \dots, n \quad (1)$$

where $A \in R^{N \times d}$ is a transformation matrix.

2.1. Principal component analysis (PCA)

PCA seeks to find a transformation matrix such that the global scatter is maximized after the projection of samples. Let S_T be the total scatter matrix:

$$S_T = \sum_{i=1}^n (x_i - m)(x_i - m)^T \quad (2)$$

where m is the mean of total training samples. The PCA transformation matrix is defined as:

$$A_{PCA} = \arg \max_A [\text{tr}(A^T S_T A)] \quad (3)$$

where $\text{tr}(\cdot)$ denotes the trace of a matrix.

Then the transformation matrix A that maximizes the objective function is obtained by solving the following generalized eigenvalue problem,

$$S_T A = \lambda A \quad (4)$$

2.2. Linear discriminant analysis (LDA)

LDA is a supervised algorithm, which seeks to find a transformation matrix such that the fisher criterion (i.e. the ratio of the between-class scatter to the within-class scatter) is maximized after projection of samples. The between-class and within-class scatter matrices S_B and S_W are defined by:

$$S_B = \sum_{i=1}^C n_i (m_i - m)(m_i - m)^T \quad (5)$$

$$S_W = \sum_{i=1}^C \sum_{j=1}^{n_i} (x_j^i - m_i)(x_j^i - m_i)^T \quad (6)$$

where C denotes the total class number and n_i denotes the number of training samples in the i th class; m_i is the mean vector of the i th class samples and m is the mean vector of total training samples; x_j^i is the j th sample in the i th class.

The LDA transformation matrix is defined as:

$$A_{LDA} = \arg \max_A \frac{\text{tr}(A^T S_B A)}{\text{tr}(A^T S_W A)} \quad (7)$$

The optimal transformation matrix that maximizes the objective function is composed of eigenvectors associated with d top eigenvalues of the following generalized eigenvalue equation,

$$S_B A = \lambda S_W A \quad (8)$$

Note that there are at most $C - 1$ non-zero (or available) generalized eigenvalues.

2.3. Locality preserving projection (LPP)

LPP aims at finding a transformation that preserves local structure of the samples, i.e. the neighbor relationship between samples so that samples that were originally in close proximity in the original space remain so in the new space. Firstly an adjacency graph $G = \{V, E\}$ is constructed using the k -NN criterion, where G denotes the graph, V is the node set and E is the edge set. Then an adjacency matrix W is defined, whose elements used to characterize the likelihood of two points are given by using the heat kernel weight below:

$$W_{ij} = \begin{cases} \exp(-\|x_i - x_j\|^2/t), & \text{if } j \in N_k(i) \text{ or } i \in N_k(j) \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

or simply 0–1 way,

$$W_{ij} = \begin{cases} 1, & \text{if } j \in N_k(i) \text{ or } i \in N_k(j) \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

where $t > 0$ is an adjustable parameter, $N_k(i)$ is the set of k nearest neighbors of x_i . In fact, the 0–1 way is a special case of (9) when $t = +\infty$.

Due to introducing the adjacency matrix W , the local scatter matrix S_L can be expressed to:

$$S_L = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n W_{ij} (x_i - x_j)(x_i - x_j)^T = X(D - W)X^T = XLX^T \quad (11)$$

where $L = D - W$ is the Laplacian matrix and D is a diagonal matrix whose entries are column (or row, since W is symmetric) sum of W , i.e. $D_{ii} = \sum_{j=1}^n W_{ij}$.

To preserve local scatter of the manifold, LPP seeks an optimal linear subspace to minimize the following constrained objective function:

$$A_{LPP} = \arg \min_{A^T X D X^T A = I} \text{tr}(A^T X L X^T A) \quad (12)$$

The transformation matrix A that minimizes the objective function are given by the minimum eigenvalue solutions to the following generalized eigenvalue problem,

$$X L X^T A = \lambda X D X^T A \quad (13)$$

3. Local maximal margin discriminant embedding (LMMDE)

3.1. Motivation

The k -NN criterion is a common way to construct a local neighborhood graph to model a manifold. Given an appropriate neighborhood size k , define a graph G with the data points as the vertices by the means of k -NN method. For the training data, each point is connected to its nearest neighbors in the training set. Apparently, the nearest neighbor approach cannot guarantee a connected graph. At this step, several disconnected graph components may be obtained and each graph component can be considered as a data manifold [23]. If two points A and B reside on two manifolds respectively, we can get that A and B are not neighbors of each other, i.e. $A \notin N_k(B)$ and $B \notin N_k(A)$. When LPP is used to project the data onto a feature space so that the neighbor relationship of the data set is preserved, it may produce high between-class overlaps. As described above, one reason may be that data points from different classes are evaluated to distribute on a manifold by using the k -NN criterion, and then they get mapped close together in the feature space. Fig. 1 illustrates an example of three classes (class 1 (\square), class 2 (\triangle), class 3 (\circ)).

From Fig. 1a, we can see that: (1) two disconnected graph components (data manifolds) are formed and (2) x_i and its neighbors not only from Class 2 but also from Class 3 reside on a manifold. From Fig. 1b, we can see that: (1) after LPP projection, data points distributed on a manifold are clustered together and (2) Classes 2 and 3 are partially overlapped due to preserving the neighbor relationship of data points in a manifold. This limitation may be overcome by developing a criterion that characterizes intra-class compactness and inter-class separability of data points in the manifold. Motivated by this, we propose a new algorithm, called local maximal margin discriminant embedding (LMMDE). After projection by LMMDE as shown in Fig. 1c, x_i and its neighbors are still living on a manifold, but data points from different classes in the manifold have been well separated.

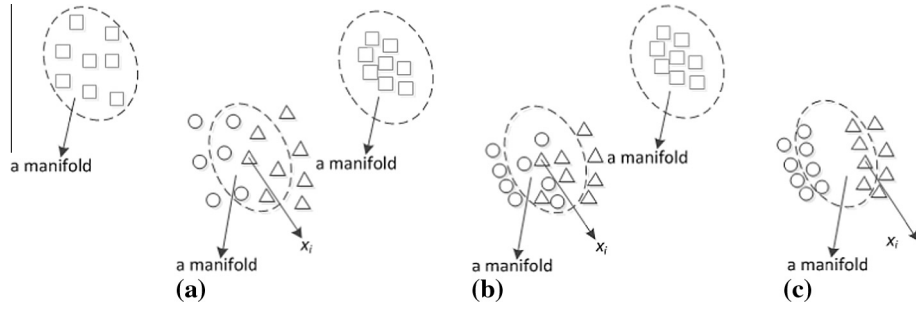


Fig. 1. An illustration of LPP and LMMDE: (a) samples in original space; (b) samples projected by LPP; (c) samples projected by LMMDE.

3.2. Formulation of the between-class neighborhood scatter

In order to characterize the inter-class separability, for each data point, we need to push its neighboring data points from different classes away from it as far as possible. Let $l_i \in \{1, \dots, C\}$ denote the class label of x_i . As mentioned in Section 2.3, if $j \in N_k(i)$ or $i \in N_k(j)$, then x_j is thought to belong to the neighborhood of x_i . Note that, the neighborhood of x_i possibly contains the data points having the same label as x_i or having different class labels from x_i . Thus, the between-class neighborhood $N_k^b(i)$ of x_i is defined as:

$$N_k^b(i) = \{x_j | \text{if } j \in N_k(i) \text{ or } i \in N_k(j), l_i \neq l_j, i, j = 1, \dots, n\} \quad (14)$$

To separate x_i from its neighboring data points with different class labels, we consider enlarging the distance between x_i and the mean of its between-class neighborhood in the projected space, i.e.

$$\left\| y_i - \frac{1}{|N_k^b(i)|} \sum_{j \in N_k^b(i), |N_k^b(i)| \neq 0} y_j \right\|^2 \quad (15)$$

where $|\cdot|$ represents the cardinality of a set.

Then the total between-class neighborhood scatter (see the derivation in Appendix A) can be defined as:

$$S'_b = \sum_i \left\| y_i - \frac{1}{|N_k^b(i)|} \sum_{j \in N_k^b(i), |N_k^b(i)| \neq 0} y_j \right\|^2 = \text{tr}(A^T S_b A) \quad (16)$$

where S_b is called the between-class neighborhood scatter matrix which is calculated as:

$$S_b = \sum_i \left(x_i - \frac{1}{|N_k^b(i)|} \sum_{j \in N_k^b(i), |N_k^b(i)| \neq 0} x_j \right) \left(x_i - \frac{1}{|N_k^b(i)|} \sum_{j \in N_k^b(i), |N_k^b(i)| \neq 0} x_j \right)^T \quad (17)$$

3.3. Formulation of the within-class neighborhood scatter

In order to characterize the intra-class compactness, for each data point, we need to pull its neighboring data points of the same class toward it as near as possible. Similarly, the within-class neighborhood $N_k^w(i)$ can be defined as:

$$N_k^w(i) = \{x_j | \text{if } j \in N_k(i) \text{ or } i \in N_k(j), l_i = l_j, i, j = 1, \dots, n\} \quad (18)$$

To compact x_i and its neighboring data points having the same class label as it, we focus on reducing the distance between x_i and the mean of its within-class neighborhood, i.e.

$$\left\| y_i - \frac{1}{|N_k^w(i)|} \sum_{j \in N_k^w(i), |N_k^w(i)| \neq 0} y_j \right\|^2 \quad (19)$$

Then the total within-class neighborhood scatter (see the derivation in Appendix A) can be formulated as:

$$S'_w = \sum_i \left\| y_i - \frac{1}{|N_k^w(i)|} \sum_{j \in N_k^w(i), |N_k^w(i)| \neq 0} y_j \right\|^2 = \text{tr}(A^T S_w A) \quad (20)$$

where S_w is called the within-class neighborhood scatter matrix which is computed as:

$$S_w = \sum_i \left(x_i - \frac{1}{|N_k^w(i)|} \sum_{j \in N_k^w(i), |N_k^w(i)| \neq 0} x_j \right) \left(x_i - \frac{1}{|N_k^w(i)|} \sum_{j \in N_k^w(i), |N_k^w(i)| \neq 0} x_j \right)^T \quad (21)$$

3.4. The objective function and the algorithm of LMMDE

The objective function of LMMDE is constructed from two aspects: (1) characterizing intra-class compactness and inter-class separability of data points in the manifold and (2) preserving the local scatter. Therefore the transformation matrix A can be obtained by solving the following objective functions:

$$\arg \max \text{tr}(A^T S_b A - A^T S_w A) = \arg \max \text{tr}(A^T (S_b - S_w) A) \quad (22)$$

and

$$\arg \min \text{tr}(A^T S_L^T A) \quad (23)$$

To eliminate the freedom that we can multiply A with some nonzero scalar, we add the constraint,

$$A^T A = I$$

where I is an identity matrix.

Thus the goal of LMMDE algorithm is just to solve the following optimization problem:

$$\begin{aligned} & \arg \max \text{tr}((1 - \mu)A^T (S_b - S_w) A - \mu A^T S_L A) \\ & \text{s.t. } A^T A = I \end{aligned} \quad (24)$$

where $0 < \mu < 1$ is a non-negative constant to balance the two terms of the objective function. Note that, both the formulations (22) and (24) are developed based on the MMC [4] to avoid the singularity problem.

Using the Lagrangian method, we can easily find that the optimal projection vectors a_1, \dots, a_d can be selected as the d eigenvectors corresponding to the first d largest eigenvalues of the following generalized eigenvalue problem:

$$((1 - \mu)(S_b - S_w) - \mu S_L)A = \lambda A \quad (25)$$

As the previous description, the proposed LMMDE algorithmic procedure can be summarized as follows:

1. For a high-dimensional application, we first project a data set $\{x_i\}_{i=1}^n$ into an m dimensional PCA subspace to reduce noise by retaining a certain portion of energy. For simplicity, we still use $\{x_i\}$ to denote the data projected in the PCA subspace in the following steps, let $W_{PCA} \in R^{N \times m}$ denote the transformation matrix of PCA.
2. Construct the adjacency matrix W using Eq. (10) and then compute the local scatter matrix S_L using Eq. (11), i.e.

$$S_L = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n W_{ij} (x_i - x_j)(x_i - x_j)^T = X(D - W)X^T = XLX^T$$

3. Use Eq. (14), i.e. $N_k^b(i) = \{x_j | \text{if } j \in N_k(i) \text{ or } i \in N_k(j), l_i \neq l_j, i, j = 1, \dots, n\}$, to determine the between-class neighborhood of each data point, and use Eq. (17), i.e.

$$S_b = \sum_i \left(x_i - \frac{1}{|N_k^b(i)|} \sum_{j \in N_k^b(i), |N_k^b(i)| \neq 0} x_j \right) \left(x_i - \frac{1}{|N_k^b(i)|} \sum_{j \in N_k^b(i), |N_k^b(i)| \neq 0} x_j \right)^T$$

to compute the between-class neighborhood scatter matrix S_b ,

4. Use Eq. (18), i.e. $N_k^w(i) = \{x_j | \text{if } j \in N_k(i) \text{ or } i \in N_k(j), l_i = l_j, i \neq j, i, j = 1, \dots, n\}$, to determine the within-class neighborhood of each data point, and use Eq. (21), i.e.

$$S_w = \sum_i \left(x_i - \frac{1}{|N_k^w(i)|} \sum_{j \in N_k^w(i), |N_k^w(i)| \neq 0} x_j \right) \left(x_i - \frac{1}{|N_k^w(i)|} \sum_{j \in N_k^w(i), |N_k^w(i)| \neq 0} x_j \right)^T$$

to compute the within-class neighborhood scatter matrix S_w .

5. Solve the eigenvalue problem $((1 - \mu)(S_b - S_w) - \mu S_L)A = \lambda A$. Let $\lambda_1 > \lambda_2 > \dots > \lambda_d$ be the d largest eigenvalues of $(1 - \mu)(S_b - S_w) - \mu S_L$ and a_1, \dots, a_d be the associated eigenvectors.
6. The final projection matrix is $A = A_{PCA}A_{LMMDE}$, where $A_{LMMDE} = [a_1, \dots, a_d]$.

4. Experiments and results

In this section, we will conduct some experiments to systematically evaluate the performance of the proposed algorithm LMMDE and some other algorithms such as MMC [4], LPP [13], MFA [16], UDP [25] and CMVM [28] on the real-work facial databases such as ORL, Yale and FERET face data. It must be noticed that PCA is firstly adopted to preprocess the data before implementing MMC, LPP, MFA, UDP, CMVM and LMMDE for feature extraction. And then, the nearest neighbor classifier is adopted to recognize the extracted feature.

For the manifold learning based methods such as LPP, MFA, UDP, CMVM and LMMDE, the k -NN criterion is used to construct graphs, and the 0–1 way is used to construct the adjacency matrices.

4.1. Experiments on ORL database

The ORL [29] face database contains images from 40 individuals, each providing 10 different images. The facial expressions and facial details (glasses or no glasses) also vary. The images were taken with a tolerance for some tilting and rotation of the face of up to 20°. Moreover, there is also some variation in the scale of up to about 10%. In our experiments, two kinds of ORL databases with different resolutions are used to show the impact of resolution on the performance of the compared methods. Fig. 2 shows sample images of one person from the ORL face database.

On ORL database, the first l ($= 2, 3, 4$) images of each person are selected to form the training sample set, and the rest $10-l$ are used to form the testing set. Note that all compared methods involve a PCA phase for data preprocess. In this phase, nearly 88% image energy is kept. The parameters of each method are set as follows: for LMMDE, the parameter μ is empirically set to 0.1 and the neighborhood size k varies from 1 to 10; for LPP and UDP, the neighborhood size k varies from 1 to 10; for MFA, the two parameters k_1 and k_2 are set to $l-1$ and $(l-1) * C$ (C is the number of classes), respectively.

Figs. 3–5 show the recognition performance of different methods corresponding to dimensions when different trains are used. Shown in Table 1 are the maximal recognition rates of MMC, LPP, MFA, UDP, CMVM and LMMDE and the corresponding dimensions (in the parentheses) when the first 2, 3, 4 images per class are used for training and the remaining for testing. From the experimental results, it can be found that LMMDE performs better than the other methods no matter what the resolution of facial images is, and in particular, when the training sample number is small, the LMMDE algorithm significantly outperforms the other methods.

4.2. Experiments on Yale database

The Yale [30] face database contains 165 grayscale images of 15 individuals. There are 11 images per subject, one per different facial expression or configuration: center/left/right-light, w/wo glasses, happy, normal, sad, sleepy, surprised, and winking. Two kinds of Yale databases are used to observe the impact of the resolution on the performance of different methods. Fig. 6 shows samples images of one person from the Yale face database.

In the experiments, the first 4 images of each person are used for training, and the remaining 7 images are used for testing. Note that, the PCA method is firstly used as a preprocessing, by which the original face images are projected into a subspace where 98% image energy is kept. To find how the neighborhood size k affects the recognition performance, we firstly set μ to 0.1 and then change k from 1 to 10 with step 1. Fig. 7 displays the recognition rates with varied k . From Fig. 7, it can be found that LMMDE obtains the best recognition rate with $k=2$ when the resolution of each image is 32×32 pixels, and achieves the maximal recognition rate with $k=6$ when the resolution of each image is 64×64 pixels.

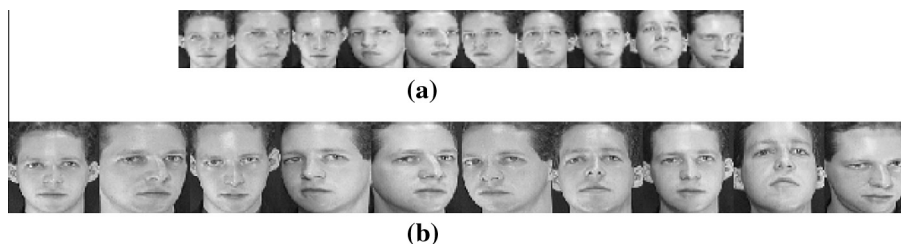


Fig. 2. Sample images of one person from the ORL database, (a) 32×32 pixels; (b) 64×64 pixels.

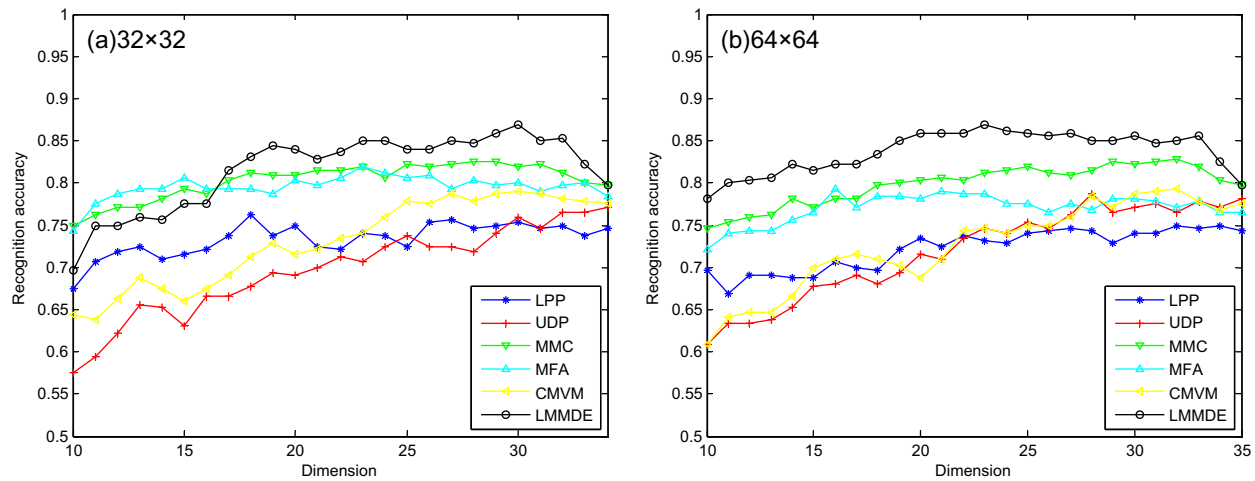


Fig. 3. The recognition rate curves of MMC, LPP, MFA, UDP, CMVM and LMMDE vs. the dimensions when 2 trains are used on ORL database, (a) 32×32 pixels; (b) 64×64 pixels.

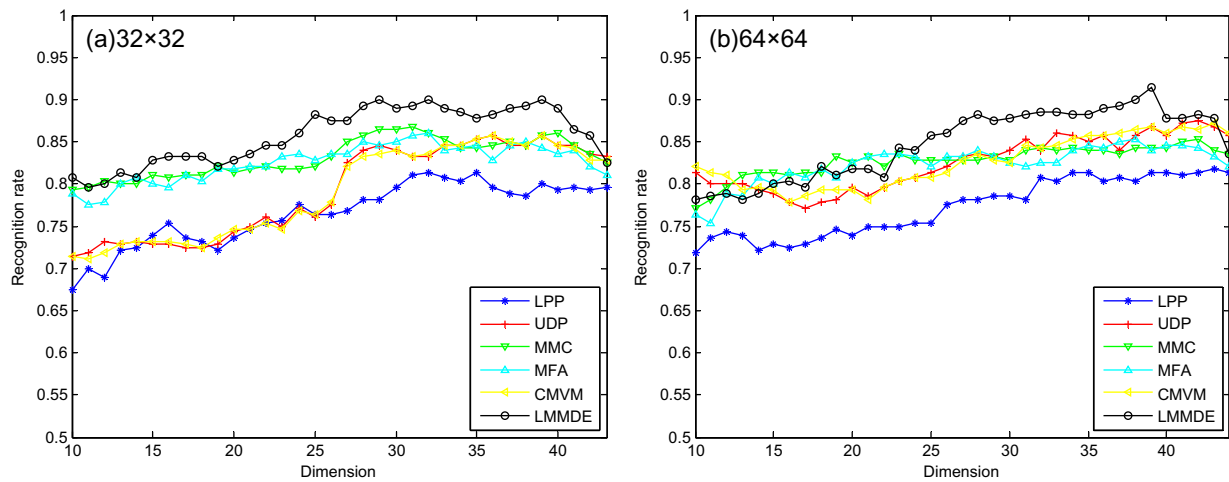


Fig. 4. The recognition rate curves of MMC, LPP, MFA, UDP, CMVM and LMMDE vs. the dimensions when 3 trains are used on ORL database, (a) 32×32 pixels; (b) 64×64 pixels.

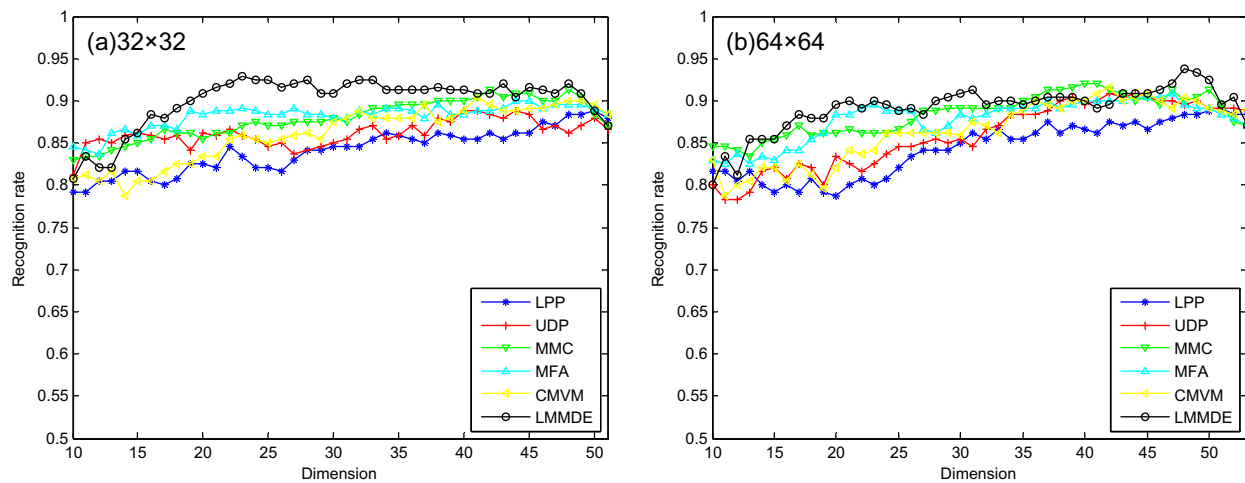
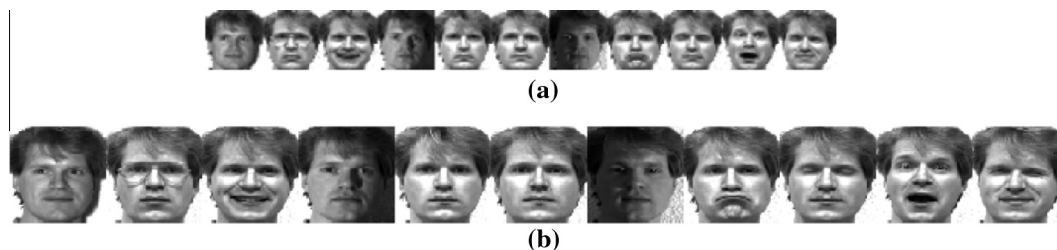
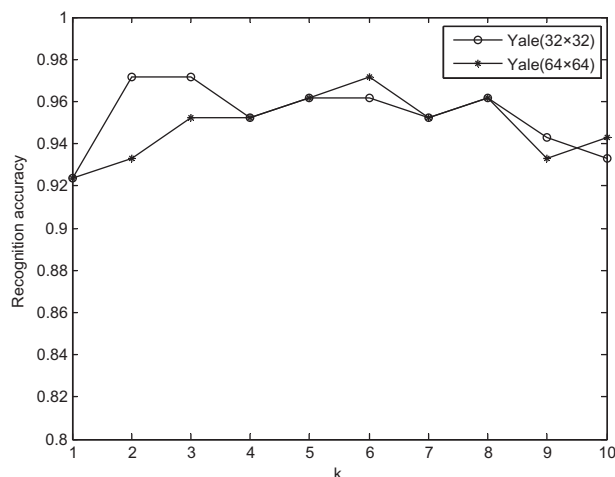
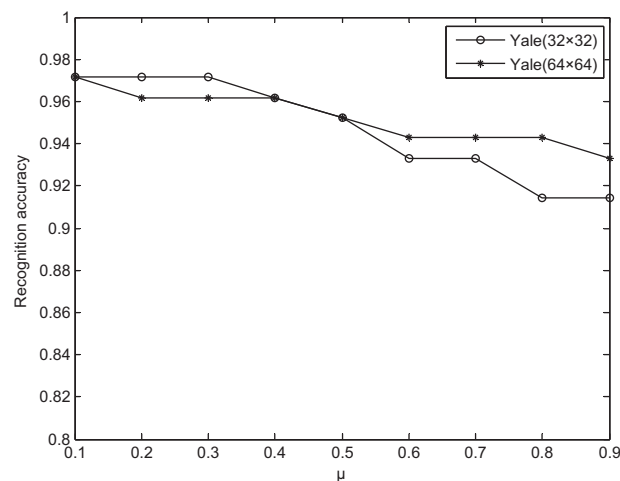


Fig. 5. The recognition rate curves of MMC, LPP, MFA, UDP, CMVM and LMMDE vs. the dimensions when 4 trains are used on ORL database, (a) 32×32 pixels; (b) 64×64 pixels.

Table 1

The maximal recognition rates (%) of MMC, LPP, MFA, UDP, CMVM and LMMDE and the corresponding dimensions (shown in the parentheses) on the ORL database when the first 2, 3, 4 images per class are selected for training and the rest for testing.

Method	Resolution = 32×32			Resolution = 64×64		
	$l = 2$	$l = 3$	$l = 4$	$l = 2$	$l = 3$	$l = 4$
MMC	82.50(28)	86.79(31)	91.25(42)	82.81(32)	85.36(42)	92.08(40)
LPP	76.25(18)	81.43(32)	88.75(50)	75.00(32)	81.79(43)	89.17(51)
MFA	81.87(23)	86.07(32)	90.00(44)	79.37(16)	85.36(38)	90.83(47)
UDP	77.19(34)	85.71(36)	88.75(40)	78.75(28)	87.50(42)	90.83(42)
CMVM	79.06(30)	85.71(36)	90.42(41)	79.37(32)	87.14(43)	91.67(42)
LMMDE	86.88(30)	90.00(29)	92.92(23)	86.88(23)	91.43(39)	93.75(48)

**Fig. 6.** Sample images of one person from the Yale database, (a) 32×32 pixels; (b) 64×64 pixels.**Fig. 7.** Recognition rates with varied k on Yale database.**Fig. 8.** Recognition rates with varied μ on Yale database.

To find how the parameter μ affects the recognition performance, we set k to 2 and k to 6 for the two kinds of resolutions of facial images respectively, and then change μ from 0.1 to 0.9 with step 0.1. Fig. 8 shows the recognition rates with varied μ . It can be found that the recognition rates show the decreasing trend with the increasing of μ , which means that maximizing the dissimilarities of data samples from different classes residing on a manifold is more important than preserving the local scatter. Shown in Table 2 are the maximal recognition rates of all the methods and corresponding dimensions. From Table 2, we can find that the LMMDE method performs best among all the methods.

Table 3 reports the computational cost of the proposed LMMDE and other compared methods including MMC, LPP, MFA, UDP and CMVM when they achieve the best recognition rates. Our hardware configuration comprises a 2.2-GHz CPU and 2 GB RAM. From Table 3, we can see that the computational complexity of the proposed algorithms for training is generally larger than other methods and all the methods consume more training time when the resolution of each image becomes larger. In practical

applications, however, training is usually an offline process and only recognition needs to be performed in real time. Thus, the recognition time is usually more of our concern than the training time. As shown in Table 3, when the resolution is 32×32 , the recognition time of LMMDE is less than those of the other approaches, but when the resolution is 64×64 , the recognition time is more than those of MMC, LPP, MFA and CMVM. The reason may be that the recognition time has a close relation to the number of features, and the smaller the number of features is the less time will be cost. From Table 2, we can clearly see that, when the resolution is 32×32 , the number of features learned by LMMDE (which is equal to 13) is the smallest among all the methods, but when the resolution is 64×64 , the number of features learned by LMMDE (which is equal to 43) is larger than the ones learned by MMC, LPP, MFA and CMVM.

4.3. Experiments on FERET database

The subset of FERET [31] face database contains 200 individuals and seven images for each person. It is composed of images whose

Table 2

The maximal recognition rates (%) of MMC, LPP, MFA, UDP, CMVM and LMMDE and the corresponding dimensions on the Yale database when the first 4 images per class are selected for training and the rest for testing.

Resolution	MMC	LPP	MFA	UDP	CMVM	LMMDE
32×32	94.29(14)	94.29(21)	96.19(27)	96.19(41)	95.24(43)	97.14(13)
64×64	95.24(14)	93.33(23)	96.19(31)	94.29(44)	94.29(38)	97.14(42)

Table 3

CPU times (s) used by different methods on the Yale database.

Method	Resolution = 32×32		Resolution = 64×64	
	Training	Recognition	Training	Recognition
MMC	0.1600	0.0320	0.2030	0.0320
LPP	0.2500	0.0360	0.4210	0.0410
MFA	0.2960	0.0470	0.4850	0.0530
UDP	0.3280	0.0650	0.5310	0.0690
CMVM	0.2030	0.0680	0.3750	0.0580
LMMDE	0.7500	0.0310	0.9370	0.0670

names are marked with two-character strings: “bd”, “bj”, “bf”, “be”, “bc”, “ba”, “bk”. This subset involves two facial expression images, two left pose images, two right pose images and an illumination image. All the images in subset are grayscale and cropped. To observe the impact of resolution on the performance of the compared methods, we use two kinds of FERET databases with different resolutions for experiments. Shown in Fig. 9 are two kinds of sample images with different resolutions of one person.

The first 4 images of each person are selected as training samples and the rest 3 images as test set. The PCA method is firstly used as a preprocessing step, whose dimension is set by retaining 98% image energy. To observe the impact of parameters k and μ on the recognition rates of LMMDE, we set the parameters as that on the Yale database. Fig. 10 displays the recognition rates with varied k . It can be found that LMMDE obtains the best recognition rate when k equals 9 in the case that the resolution of each image is 32×32 pixels, and obtains the best recognition rate when k equals 5 in the case that the resolution of each image is 80×80 pixels. Fig. 11 shows the recognition rates with varied μ . It can also be found that the recognition rates decreases with the increasing of μ both in two kinds of FERET databases, which demonstrates the importance of maximizing the dissimilarities of samples from different classes lying on a manifold. Shown in Table 4 are the maximal recognition rates of MMC, LPP, MFA, UDP, CMVM and LMMDE and the corresponding dimensions (in the parentheses). From Table 4, we can see that: (1) LMMDE outperform other methods both on two kinds of FERET databases and (2) the performance of MMC and LMMDE becomes better when the resolution of each image tends to be larger, while the other methods are not.

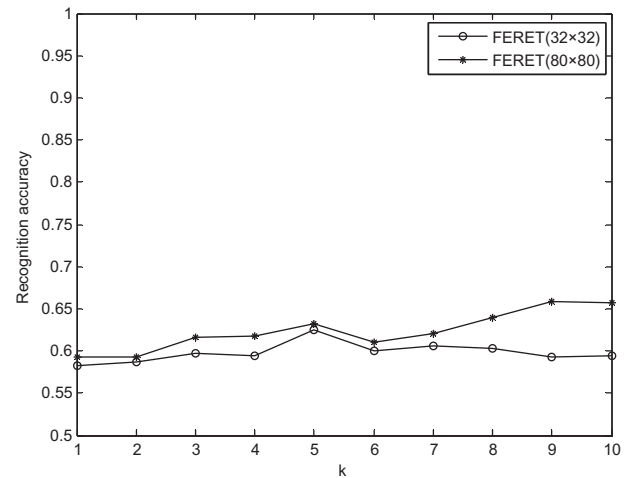
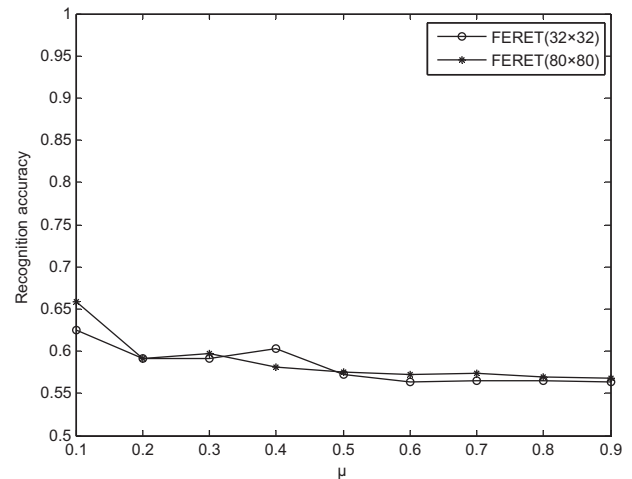
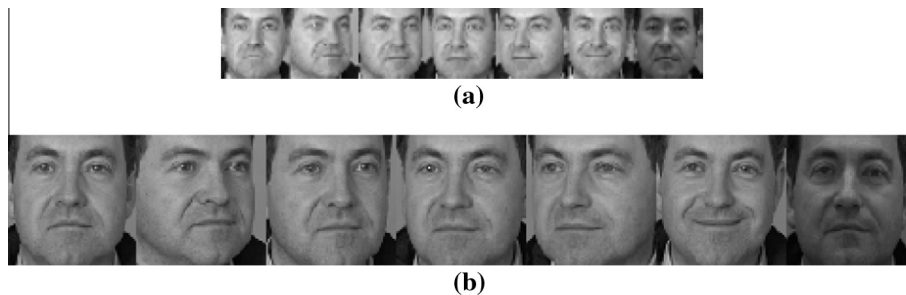
**Fig. 10.** Recognition rates with varied k on FERET database.**Fig. 11.** Recognition rates with varied μ on the FERET database.**Fig. 9.** Samples images of one person from the FERET database, (a) 32×32 pixels; (b) 80×80 pixels.

Table 4

The maximal recognition rates (%) of MMC, LPP, MFA, UDP, CMVM and LMMDE and the corresponding dimensions (in the parentheses) on the FERET database when the first 4 images per class are selected for training and the rest for testing.

Resolution	MMC	LPP	MFA	UDP	CMVM	LMMDE
32 × 32	58.50(26)	40.50(50)	56.67(42)	36.83(262)	36.50(254)	62.50(118)
80 × 80	60.50(35)	34.67(35)	55.50(20)	29.50(365)	30.33(325)	65.83(370)

5. Conclusions

In this paper, a manifold learning based algorithm, namely LMMDE, is proposed for face recognition and classification. The proposed algorithm takes consideration of intra-class compactness and inter-class separability of samples that lies on a manifold after the local structure of the data have been characterized. So the proposed algorithm becomes more suitable for classification, and the experiments results on real-world data sets validate this result.

It must be noted that the pixel based approaches [32–38] are reported to show better performance, but they are essentially different from the methods involving MMC, LPP, MFA, UDP, CMVM and LMMDE in reducing the computational complexity and boosting pattern matching results. MMC, LPP, MFA, UDP, CMVM and LMMDE are all dimensionality reduction methods with the purpose of transforming the original high-dimensional data into a meaningful representation of reduced dimensionality. When the representation is used for classification with a classifier, it will save much time and enhance the pattern recognition results in comparison with that the original data is directly used for classification. Instead of reducing the dimension of the original data, the methods proposed in [32–38] mainly focus on learning different classifiers (or pattern matching functions). In contrast to the methods in [32–37], a fast illumination and deformation insensitive image comparison algorithm [38] shows substantial improvements in recognition accuracy and speed because it can vastly simplify the calculations of matching functions and effectively handle variations in illumination and moderate amounts of deformation.

There are still several drawbacks existed in LMMDE to be considered in the future work. First, the proposed algorithm LMMDE fails to work in the scenario that there is only a single sample per person (SSPP) available during the training phase since the within-class neighborhood scatter cannot be computed due to the lack of samples. To address the SSPP problem in face recognition, there have been some attempts in the literatures [39–48], which can be mainly classified into three categories [44]: generic learning, virtual sample generation, and image partitioning. Using the virtue sample generation [46] and image partitioning [47] techniques to extend LMMDE for solving SSPP problem is one goal in our future work. Second, our method is linear, and extending it to a nonlinear method using the kernel trick [49] is another goal in the future. Third, the neighborhood size k and the parameter μ are very important in improving the performance of LMMDE, so how to effectively set them is an interesting topic in future research.

Acknowledgements

This work was supported by the Grants of the National Science Foundation of China, Nos. 90820306, 61125305 and 60875004; the Grant of college postgraduate research and innovative project in Jiangsu province, No. CXZZ12_0204.

Appendix A. Derivation of Eqs. (16) and (20)

Substituting $y_i = A^T x_i$ into Eq. (16), we can see that:

$$\begin{aligned}
 S_b' &= \sum_i \left\| y_i - \frac{1}{|N_k^b(i)|} \sum_{j \in N_k^b(i), |N_k^b(i)| \neq 0} y_j \right\|^2 \\
 &= \text{tr} \left(\sum_i \left(y_i - \frac{1}{|N_k^b(i)|} \sum_{j \in N_k^b(i), |N_k^b(i)| \neq 0} y_j \right) \left(y_i - \frac{1}{|N_k^b(i)|} \sum_{j \in N_k^b(i), |N_k^b(i)| \neq 0} y_j \right)^T \right) \\
 &= \text{tr} \left(\sum_i \left(A^T x_i - \frac{1}{|N_k^b(i)|} \sum_{j \in N_k^b(i), |N_k^b(i)| \neq 0} A^T x_j \right) \left(A^T x_i - \frac{1}{|N_k^b(i)|} \sum_{j \in N_k^b(i), |N_k^b(i)| \neq 0} A^T x_j \right)^T \right) \\
 &= \text{tr} \left(A^T \left(\sum_i \left(x_i - \frac{1}{|N_k^b(i)|} \sum_{j \in N_k^b(i), |N_k^b(i)| \neq 0} x_j \right) \left(x_i - \frac{1}{|N_k^b(i)|} \sum_{j \in N_k^b(i), |N_k^b(i)| \neq 0} x_j \right)^T \right) A \right) \\
 &= \text{tr}(A^T S_b A)
 \end{aligned} \tag{26}$$

where $\frac{1}{|N_k^b(i)|} \sum_{j \in N_k^b(i), |N_k^b(i)| \neq 0} x_j$ is the mean vector of the within-class neighborhood of x_i .

Similar to Eqs. (16), (20) can be rewritten in a form of the matrix trace as follows:

$$\begin{aligned}
 S_w' &= \sum_i \left\| y_i - \frac{1}{|N_k^w(i)|} \sum_{j \in N_k^w(i), |N_k^w(i)| \neq 0} y_j \right\|^2 \\
 &= \text{tr} \left(\sum_i \left(y_i - \frac{1}{|N_k^w(i)|} \sum_{j \in N_k^w(i), |N_k^w(i)| \neq 0} y_j \right) \left(y_i - \frac{1}{|N_k^w(i)|} \sum_{j \in N_k^w(i), |N_k^w(i)| \neq 0} y_j \right)^T \right) \\
 &= \text{tr} \left(\sum_i \left(A^T x_i - \frac{1}{|N_k^w(i)|} \sum_{j \in N_k^w(i), |N_k^w(i)| \neq 0} A^T x_j \right) \left(A^T x_i - \frac{1}{|N_k^w(i)|} \sum_{j \in N_k^w(i), |N_k^w(i)| \neq 0} A^T x_j \right)^T \right) \\
 &= \text{tr} \left(A^T \left(\sum_i \left(x_i - \frac{1}{|N_k^w(i)|} \sum_{j \in N_k^w(i), |N_k^w(i)| \neq 0} x_j \right) \left(x_i - \frac{1}{|N_k^w(i)|} \sum_{j \in N_k^w(i), |N_k^w(i)| \neq 0} x_j \right)^T \right) A \right) \\
 &= \text{tr}(A^T S_w A)
 \end{aligned} \tag{27}$$

where $\frac{1}{|N_k^w(i)|} \sum_{j \in N_k^w(i), |N_k^w(i)| \neq 0} x_j$ is the mean vector of the within-class neighborhood of x_i .

References

- [1] M. Turk, A. Pentland, Eigenfaces for recognition, *J. Cognitive Neurosci.* 3 (1) (1991) 71–86.
- [2] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, *IEEE Trans. Pattern Anal. Mach. Intell.* 19 (7) (1997) 711–720.
- [3] S.J. Raudys, A.K. Jain, Small sample size effects in statistical pattern recognition: recommendations for practitioners, *IEEE Trans. Pattern Anal. Mach. Intell.* 13 (3) (1991) 252–264.
- [4] H. Li, T. Jiang, K. Zhang, Efficient and robust feature extraction by maximum margin criterion, *IEEE Trans. Neural Networks* 17 (1) (2006) 1157–1165.
- [5] S.T. Roweis, L.K. Saul, Nonlinear dimension reduction by locally linear embedding, *Science* 290 (5500) (2000) 2323–2326.
- [6] J.B. Tenenbaum, V.D. Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, *Science* 290 (2000) 2319–2323.
- [7] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, *Neural Comput.* 15 (6) (2003) 1373–1396.
- [8] Z. Zhang, H. Zha, Principal manifolds and nonlinear dimension reduction via local tangent space alignment, *SIAM J. Sci. Comput.* 26 (1) (2005) 313–338.
- [9] X. He, S. Yan, Y. Hu, H. Zhang, Learning a locality preserving subspace for visual recognition, *ICCV* (2003) 385–393.
- [10] X. He, D. Cai, W. Min, Statistical and computational analysis of locality preserving projection, in: *Proceedings of the 22nd International Conference on Machine Learning*, 2005, pp. 281–288.
- [11] D. Cai, X. He, Orthogonal locality preserving indexing, *SIGIR* (2005) 3–10.
- [12] J.-Z. Wang, B.-X. Zhang, S.-Y. Wang, M. Qi, J. Kong, An adaptively weighted sub-pattern locality preserving projection for face recognition, *J. Network Comput. Appl.* 33 (2010) 323–332.

- [13] X. He, S. Yan, Y. Hu, P. Niyogi, H. Zhang, Face recognition using Laplacian faces, *IEEE Trans. Pattern Anal. Mach. Intell.* 27 (3) (2005) 328–340.
- [14] J. Cheng, Q. Liu, H. Lu, Y.-W. Chen, Supervised kernel locality preserving projections for face recognition, *Neurocomputing* 67 (2005) 443–449.
- [15] H.-T. Chen, H.-W. Chang, T.-L. Liu., Local discriminant embedding and its variants, in: *Proc. CVPR*, 2005.
- [16] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, S. Lin, Graph embedding and extension: a general framework for dimensionality reduction, *IEEE Trans. Pattern Anal. Mach. Intell.* 29 (1) (2007).
- [17] J.-B. Li, J.-S. Pan, Shu-Chuan Chu, Kernel class-wise locality preserving projection, *Inf. Sci.* 178 (2008) 1825–1835.
- [18] X. He, Incremental semi-supervised subspace learning for image retrieval, *ACM Multimedia* (2004) 2–8.
- [19] J. Gui, J. Wei, L. Zhu, S. Wang, D. Huang, Locality preserving discriminant projections for face and palmprint recognition, *Neurocomputing* 73 (2010) 2696–2707.
- [20] W. Yu, X. Teng, C. Liu, Face recognition using discriminant locality preserving projections, *Image Vis. Comput.* 24 (2006) 239–248.
- [21] D. Cai, X. He, K. Zhou, J. Han, H. Bao, Locality sensitive discriminant analysis, in: *Proceeding of International Joint Conference on Artificial Intelligence*, 2007, pp. 708–713.
- [22] Y. Xu, A. Zhong, J. Yang, D. Zhang, LPP solution schemes for use with face recognition, *Pattern Recogn.* 43 (2010) 4165–4175.
- [23] M.-Y. Fan, X.-Q. Zhang, Z.-C. Lin, Z.-F. Zhang, H.-J. Bao, Geodesic based semi-supervised multi-manifold feature extraction, in: *Proceeding of IEEE Conference on Data Mining*, 2012, pp. 852–857.
- [24] N. Vlassis, Y. Motomura, B. Krose, Supervised dimension reduction of intrinsically low dimensional data, *Neural Comput.* 14 (1) (2002) 191–215.
- [25] J. Yang, D. Zhang, J.Y. Yang, B. Niu, Globally maximizing, locally minimizing: unsupervised discriminant projection with application to face and palm biometrics, *IEEE Trans. Pattern Anal. Mach. Intell.* 29 (4) (2007) 650–664.
- [26] W. Deng, J. Hu, J. Guo, H. Zhang, C. Zhang, Comments on globally maximizing, locally minimizing: unsupervised discriminant projection with application to face and palm biometrics, *IEEE Trans. Pattern Anal. Mach. Intell.* 30 (8) (2008) 1503–1504.
- [27] H. Zhang, W. Deng, J. Guo, J. Yang, Locality preserving and global discriminant projection with prior information, *Mach. Vis. Appl.* 21 (2010) 577–585.
- [28] B. Li, D. Huang, C. Wang, K. Liu, Feature extraction using constrained maximum variance mapping, *Pattern Recogn.* 41 (2008) 3287–3294.
- [29] See (<http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html>).
- [30] See (<http://cvc.yale.edu/projects/yalefaces/yalefaces.html>).
- [31] P.J. Phillips, H. Wechsler, J. Huang, P. Rauss, The FERET database and evaluation procedure for face recognition algorithms, *Image Vis. Comput.* 16 (5) (1998) 295–306.
- [32] R. Gopalan, D. Jacobs, Comparing and combining lighting insensitive approaches for face recognition, *CVIU* 114 (1) (2010) 135–145.
- [33] A.P. James, Pixel-level decisions based robust face image recognition, in: M. Oravek (Ed.), *Face Recognition*, INTECH, 2010, pp. 65–86. ch. 5.
- [34] A. Jorstad, D. Jacobs, A. Trounev, A deformation and lighting insensitive metric for face recognition based on dense correspondences, *CVPR* (2011) 2353–2360.
- [35] J.P. Lewis, Fast normalized cross-correlation, *Vision Interface* 10 (1) (1995) 120–123.
- [36] J. Song, B. Chen, W. Wang, X. Ren, Face recognition by fusing binary edge feature and second-order mutual information, in: *The Proceedings of IEEE Conference on Cybernetics and Intelligent Systems*, 2008, pp. 1046–1050.
- [37] S. Zhao, Y. Gao, Significant jet point for facial image representation and recognition, in: *The Proceedings of the International Conference on Image Processing*, 2008, pp. 1664–1667.
- [38] A. Jorstad, D. Jacobs, A. Trounev, A fast illumination and deformation insensitive image comparison algorithm using wavelet-based geodesics, in: *The Proceedings of the European Conference on Computer Vision*, 2012, pp. 71–84.
- [39] T. Kim, J. Kittler, Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image, *IEEE Trans. Pattern Anal. Mach. Intell.* 27 (3) (2005) 318–327.
- [40] C.H. Kuo, J.D. Lee, Face recognition based on a two-view projective transformation using one sample per subject, *IET Comput. Vision* 6 (5) (2012) 489–498.
- [41] Z.L. Sun, K.M. Lam, Z.Y. Dong, H. Wang, A spectral feature based approach for face recognition with one training sample, in: *2012 IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC)*, 2012, pp. 218–212.
- [42] J.-W. Lu, Y.-P. Tan, G. Wang, Discriminative multi-manifold analysis for face recognition from a single training sample per person, in: *2011 IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 1943–1950.
- [43] X. Tan, S. Chen, Z. Zhou, F. Zhang, Recognizing partially occluded expression variant faces from single training image per person with SOM and soft k-NN ensemble, *IEEE Trans. Neural Networks* 16 (4) (2005) 875–886.
- [44] X. Tan, S. Chen, Z. Zhou, F. Zhang, Face recognition from a single image per person: a survey, *Pattern Recogn.* 39 (9) (2006) 1725–1745.
- [45] D. Zhang, S. Chen, Z. Zhou, A new face recognition method based on SVD perturbation for single example image per person, *Appl. Math. Comput.* 163 (2) (2005) 895–907.
- [46] Q. Gao, L. Zhang, D. Zhang, Face recognition using FLDA with single training image per person, *Appl. Math. Comput.* 205 (2) (2008) 726–734.
- [47] S. Chen, J. Liu, Z. Zhou, Making FLDA applicable to face recognition with one sample per person, *Pattern Recogn.* 37 (7) (2004) 1553–1555.
- [48] H. Kanan, K. Faez, Y. Gao, Face recognition using adaptively weighted patch PZM array from a single exemplar image per person, *Pattern Recogn.* 41 (12) (2008) 3799–3812.
- [49] V.N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.