Two-dimensional Maximum Local Variation based on Image Euclidean Distance for Face Recognition

Quanxue Gao[1] Feifei Gao [1] Hailin Zhang [1] Xiu-juan Hao [1] Xiaogang Wang[3]
[1]State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071 China
[2]State Key Laboratory of CAD & CG, ZHE JIANG University, HangZhou, 310058 China
[3]Department of Electrical Engineering, The Chinese University of Hong Kong, Hong Kong

[*]**Abstract:** Manifold learning concerns the local manifold structure of high dimensional data and many related algorithms have been developed to improve image classification performance. None of them, however, takes the relationships among pixels in images and the geometrical properties of diversity of images together into account during learning the reduced space. In this paper, we propose a linear approach, called two-dimensional maximum local variation (2DMLV), for face recognition. In 2DMLV, we encode the relationships among pixels in image by using the image Euclidean distance instead of conventional Euclidean distance in estimating the variation of values of images, and then incorporate the local variation, which characterizes the diversity of images and discriminating information, into the objective function of dimensionality reduction. Extensive experiments demonstrate the effectiveness of our approach.

Keywords: Dimensionality reduction, Local variation, Image Euclidean distance, Face recognition

1. Introduction

In recent years, statistical pattern recognition research has witnessed a growing interest in linear dimensionality reduction (DR) techniques [1-8].DR aims to seek a low-dimension space that best preserves the intrinsic geometrical structure embedded in the high-dimensional data space. Two of the most popular techniques for this purpose are Principal Component Analysis (PCA) [1,9] and Linear Discriminant Analysis (LDA) [2,9]. PCA [9] seeks to find a projection direction along which the data have the maximum variance and unfolds the global Euclidean geometrical structure of data. LDA, as a supervised approach, searches for the projection axes on which the data points of different classes are far from each other while requiring data points of the same class to be close to each other [9]. Applying PCA and LDA techniques into data representation and classification, many approaches have been developed in the literature [1, 2, 10-17].

However, most real-life data such as face images possibly reside on a linearly inseparable submanifold of the observed data space [3,18,19]. Thus, the above-mentioned approaches fail in discovering the underlying structure of high-dimensional data space because they only capture the global Euclidean structure of data. Kernel learning is usually considered effective in discovering the geometrical structure of the data manifold [20-22]. The basic idea of kernel techniques is to implicitly map the observed data into potentially much higher dimensional feature space by using a kernel trick and perform linear dimensionality reduction techniques, such as PCA and LDA, in feature space. However, an open question of these approaches is how to preserve the geometrical relationship embedded in the training data.

Recently, geometrically motivated approaches (also called manifold learning approaches), which are straightforward in discovering the geometric structure of the data manifold, have been shown to be successful in improving the recognition performance of image classification [3-5, 23-36]. Two of the most prevalent linear approaches are Locality Preserving Projection (LPP) [3] and Isometric Projection (IsoP) [36]. LPP maps nearby points in the high-dimensional data space to nearby points with the low-dimensional representation. Motivated by LPP, many linear approaches have been developed in many areas [23-36]. Although their motivations are different, both of them can be unified within a graph-embedding framework [4] and preserve the intrinsic geometrical structure by minimizing the sum of distances, i.e. variance between nearby data points. In the ideal case, the nearby data points are mapped to a single point in the reduced space. Thus, these methods ignore the variation, which characterizes the different geometrical properties, i.e. diversity of data, and do not unfold the manifold structure of data [29, 37-38]. Another limitation is that they may impair the local topology of data, leading to unstable intrinsic structure representation [39-40].

IsoP, which is a linear approximation of Isomap [18], discovers the intrinsic geometrical structure of data by preserving all the pairwise distances. Motivated by Isop and PCA, Weinberger and Saul proposed maximum variance unfolding (MVU) method for nonlinear DR [38]. MVU unfolds the manifold structure and preserves the diversity of patterns by maximizing the sum total of their pairwise distances with preserving the distances between nearby points. However, some limitations are exposed when MVU is applied to image recognition. The first limitation is the out-of-sample problem. The second limitation is that it may impair the diversity among nearby images and does not well unfold the local manifold structure that is important for image recognition, because the objective function of MVU emphasizes the large distance pair points and deemphasizes the small distance pair points. The third limitation is that MVU does not well detect the local discriminating information. The fourth limitation is that MVU does not take into account the spatial relationships among pixels in images. Therefore, it is sensitive to the perturbation of images [41-42].

Combining the aforementioned insight into the dimensionality reduction, and motivated by MVU and two-dimensional techniques, such as 2DPCA [11] that avoids transforming image matrix into image vector and has better recognition accuracy, we propose a novel linear approach, namely two-dimensional maximum local variation (2DMLV), to detect the geometrical properties of diversity of images in this paper. In 2DMLV, we first define the local variation of vector-valued variables, which may characterizes both the diversity and discriminating information of images, and then encode the spatial relationships among pixels by employing image Euclidean distance instead of conventional Euclidean distance in measuring the local variation. Thus, by maximizing local variation, we obtain a low-dimensional space that well encodes the discriminating information of images and preserves the intrinsic geometrical structure that characterizes the diversity of images and is not sensitive to the perturbation of images. Extensive experiments demonstrate the effectiveness of our approach.

The rest of this paper is organized as follows. Section 2 introduces some related work. Section 3 presents the motivation and formulation of 2DMLV. Section 4 presents 2DMLV+2DPCA approach to further reduce the dimensionality. Section 5 provides a discussion of 2DMLV in detail. Section 6 shows the experimental results. Some conclusions are drawn in Section 7.

2. Related work

During the past decade, lots of manifold learning approaches have been developed to preserve the local intrinsic structure of data, among which two of the popular techniques are LPP and MVU. LPP seeks to preserve the intrinsic geometry of the data and the local structures. Given training data matrix $X = [\mathbf{x}_1\ \mathbf{x}_2\ \ \ \mathbf{x}_N]$ and a k nearest neighbor graph $G = \{X, S\}$ with weight matrix S, the objective function of LPP is as follows[3]:

$$\min_{\mathbf{y}} \sum_{i,j} \left(y_i - y_j\right)^2 S_{ij} \tag{1}$$

where y_i is the low dimensional representation of $\mathbf{x}_i$, the elements S_{ij} are defined as in [3].

In the ideal case, the objective function (1) maps nearby data points to be a single point in the reduced space. Thus, Eq. (1) ignores the variation, which characterizes the diversity of data, and does not unfold the manifold structure. Moreover, Eq. (1) emphasizes the pair points with large distance and does not guarantee that the larger the distance between nearby two points is, the further they are embedded in the low-dimensional space. Thus, Eq.(1) may impairs the local topology of data [29].

MVU aims at finding low-dimensional representations that can pull the data points apart with the locally distance-preserving constraint [37, 38]. Let y_i is the low dimensional representation of $\mathbf{x}_i$. The optimal maps can be obtained by solving the following maximization problem.

$$\begin{aligned} &\max \quad \sum_{ij} \left\| y_i - y_j \right\|^2 \\ &S.t \quad \left\| y_i - y_j \right\|^2 = \left\| \mathbf{x}_i - \mathbf{x}_j \right\|^2 \ with\, \eta_{ij} = 1 \\ &\qquad \sum_i y_i = 0 \end{aligned} \tag{2}$$

where $\eta_{ij} = 1$ means that $\mathbf{x}_i$ and $\mathbf{x}_j$ are k-nearest neighbors.

Comparing the objective function (2) with the objective function (1), we can see that, different from LPP, which preserves the local intrinsic structure of data by minimizing the variance of data, MVU maximizes the variance of data to preserve the intrinsic structure and helps to unfolding the manifold structure and characterizing the diversity of data [38]. However, the objective function (2) has the aforementioned limitations due to the facts that it ignores the relationships among pixels in faces,and faces in a local neighborhood usually come from different classes.

3. Two-dimensional Maximum local variance

3.1 Motivation

Given an arbitrary image matrix $A \in R^{m \times n}$, our objective is to find a lower dimensional feature vector $\mathbf{y} \in R^m$ by applying linear transformation $\mathbf{y} = A\boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ is the projection vector. The problem addressed in the paper is how to estimate $\boldsymbol{\alpha}$ such that both the local intrinsic structure, which characterizes the diversity of images, and discriminating information are preserved well. As aforementioned analysis, local diversity and intrinsic structure can be preserved by maximizing the variance of data in the local neighborhoods [35, 37, 38, 40]. Moreover, image Euclidean distance considers the relationships among pixels in images and is insensitive to the perturbation of images [41, 42]. Figure 1 shows 3 nearby images, where (a) and (b) belong to the same person, and (c) belongs to another person. Computing the Euclidean distance yields $d(a,b) = 1489.2$ and $d(a,c) = 1334.5$ while the image Euclidean distance between them are respectively $d_{IED}(a,b) = 4.9583$ and $d_{IED}(a,c) = 5.9269$. Therefore, maximizing the image Euclidean distance helps to encoding the local discriminating information. In the following section, we first present the definition of variation of the vector-valued variables and then encode the relationships among pixels in images into the variation of a vector-valued variables.

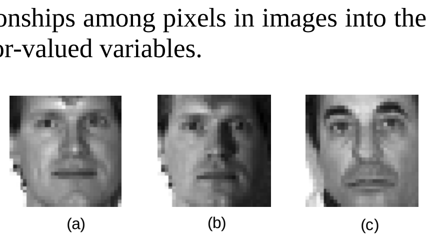

Figure 1. Three nearby images from different persons.

3.2 Variation of vector-valued variable

Variance is the mean of the squared deviation of that variable from its mean and characterizes the amount of variation of the values of that variable. For a vector-valued random variable $\mathbf{y} \in R^{\mathrm{m}}$, each element y_j in $\mathbf{y}$ can be viewed as a random variable, thus the sum total of the amount of variation of the values of each element y_j in

$\mathbf{y}$ effectively characterizes the amount of variation of the values of vector $\mathbf{y}$. Thus, the variation of random vector $\mathbf{y} \in R^n$ is

$$Var(\mathbf{y}) = \sum_{j=1}^{m} Var(y_j) = \sum_{j=1}^{m} E\left((y_j - \overline{y_j})^2\right) = E\left(\left(\mathbf{y} - \overline{\mathbf{y}}\right)^T \left(\mathbf{y} - \overline{\mathbf{y}}\right)\right) \quad (3)$$

where y_j is a random variable and denotes the j th element in $\mathbf{y}$. $\overline{y_j}$ is the mean of random variable y_j. $\overline{\mathbf{y}}$ denotes the mean of random variable $\mathbf{y}$. $\left(\mathbf{y} - \overline{\mathbf{y}}\right)^T$ is the transpose of $\left(\mathbf{y} - \overline{\mathbf{y}}\right)$.

Given N random vectors $\mathbf{y}_i \in R^m$, Eq. (3) becomes:

$$Var(\mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} \left(\mathbf{y}_i - \overline{\mathbf{y}}\right)^T \left(\mathbf{y}_i - \overline{\mathbf{y}}\right) \quad (4)$$

where $\overline{\mathbf{y}} = 1/N \sum_{i=1}^{N} \mathbf{y}_i$ is the mean of N vectors.

By simple algebraic formulation, we see that

$$\begin{aligned} Var(\mathbf{y}) &= \frac{1}{N} \sum_{i=1}^{N} \left(\mathbf{y}_i - \overline{\mathbf{y}}\right)^T \left(\mathbf{y}_i - \overline{\mathbf{y}}\right) \\ &= \frac{1}{N} \sum_{i=1}^{N} \left(\mathbf{y}_i^T \mathbf{y}_i - \mathbf{y}_i^T \overline{\mathbf{y}} - \overline{\mathbf{y}}^T \mathbf{y}_i + \overline{\mathbf{y}}^T \overline{\mathbf{y}}\right) \\ &= \frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\mathbf{y}_i^T \mathbf{y}_i - \mathbf{y}_j^T \mathbf{y}_i + \mathbf{y}_i^T \mathbf{y}_j - \mathbf{y}_i^T \mathbf{y}_j\right) \\ &= \frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\mathbf{y}_i - \mathbf{y}_j\right)^T \left(\mathbf{y}_i - \mathbf{y}_j\right) \end{aligned} \quad (5)$$

3.3 Local variation

Given training images $A_i \in R^{m \times n}$ ($i = 1, \quad, N$). Let $\mathbf{y}_i \in R^m$ be the low-dimensional representations of A_i. According to Eq. (5), the variation of the low- dimensional representations in the local neighborhood Ω_i^k can be defined as follows:

$$Var\left(\Omega_i^k\right) = \frac{1}{2k^2} \sum_{\substack{\mathbf{y}_l \in \Omega_i^k \\ \mathbf{y}_j \in \Omega_i^k}} \left(\mathbf{y}_l - \mathbf{y}_j\right)^T \left(\mathbf{y}_l - \mathbf{y}_j\right) \quad (6)$$

where $\Omega_i^k = \{\mathbf{y}_j \mid A_j (or\ A_i)\ is among k\ nearest neighbors of\ A_i (or\ A_j)\}$ Thus, the total local variation of the low-dimensional representations is

$$\sum_{i=1}^{N} Var\left(\Omega_i^k\right) = \sum_{i=1}^{N} \left(\frac{1}{2k^2} \sum_{\substack{\mathbf{y}_l \in \Omega_i^k \\ \mathbf{y}_j \in \Omega_i^k}} \left(\mathbf{y}_l - \mathbf{y}_j\right)^T \left(\mathbf{y}_l - \mathbf{y}_j\right) \right) \quad (7)$$

Let $V_{jl} = 1$ if A_j is among k nearest neighbors of A_l or A_l is among k nearest neighbors of A_j, otherwise $V_{jl} = 0$. Substituting V_{jl} into Eq. (7), Eq. (7) becomes

$$\sum_{i=1}^{N} Var\left(\Omega_i^k\right) = \frac{1}{2k^2} \sum_{l=1}^{N} \sum_{j=1}^{N} \left(\left(\mathbf{y}_l - \mathbf{y}_j\right)^T \left(\mathbf{y}_l - \mathbf{y}_j\right) V_{jl}\right) \quad (8)$$

In real-world applications, images in the neighborhood may come from different classes, and the variation of images from the same class reflects the geometrical properties of diversity of images, while the variation of images from different classes characterizes the discriminating information of images. Thus, we divide Eq. (8) into two parts. One is called the local within-class variation S^w that characterizes the diversity of data; another is called the local between-class variation S^b that characterizes the discriminating information. They can be defined respectively as follows:

$$S^w = \frac{1}{2k^2} \sum_{l=1}^{N} \sum_{j=1}^{N} \left(\left(\mathbf{y}_l - \mathbf{y}_j\right)^T \left(\mathbf{y}_l - \mathbf{y}_j\right) V_{jl}^w\right) \quad (9)$$

$$S^b = \frac{1}{2k^2} \sum_{l=1}^{N} \sum_{j=1}^{N} \left(\left(\mathbf{y}_l - \mathbf{y}_j\right)^T \left(\mathbf{y}_l - \mathbf{y}_j\right) V_{jl}^b\right) \quad (10)$$

where V_{jl}^w and V_{jl}^b are defined as follows

$$V_{jl}^w = \begin{cases} V_{jl}, if\ \tau_j = \tau_l \\ 0 \quad , Otherwise \end{cases}, \quad V_{jl}^b = \begin{cases} V_{jl}, if\ \tau_j \neq \tau_l \\ 0 \quad , Otherwise \end{cases} \quad (11)$$

where τ_j denotes the class label of image A_j.

Eq. (9) and Eq. (10) implicitly considers that all low-dimensional representations are equally important in estimating the amount of variation of images. This impairs the recognition accuracy of the algorithm. In order to well preserve the diversity of images, we should guarantee that the larger the distance between nearby two images is, the further they are embedded in the low-dimensional space. If they are embedded close to each other, we should assign larger penalty weight to separate them. It means that the weight is proportional to the distance between them. Thus, the elements V_{jl}^w in V^w and elements V_{jl}^b in V^b are respectively defined

$$V_{jl}^w = \begin{cases} V_{jl} \times \exp\left(-t / \left\|A_j - A_l\right\|_F^2\right), if\ \tau_j = \tau_l \\ 0 \qquad , Otherwise \end{cases} \quad (12)$$

$$V_{jl}^b = \begin{cases} V_{jl} \times \exp\left(-t / \left\|A_j - A_l\right\|_F^2\right), if\ \tau_j \neq \tau_l \\ 0 \qquad , Otherwise \end{cases} \quad (13)$$

where $t \geq 0$ is a parameter, $\left\|A_j - A_l\right\|_F$ denotes the Frobenius norm of matrix $A_j - A_l$.

Now consider the problem of mapping training images to a line so that, if the variation among nearby images in the original data space is large, then the variation among the corresponding low-dimensional representations should be large. As it happens, two reasonable objective functions are as follows:

$$\arg\max \sum_{l=1}^{N} \sum_{j=1}^{N} \left(\left(\mathbf{y}_l - \mathbf{y}_j\right)^T \left(\mathbf{y}_l - \mathbf{y}_j\right) V_{jl}^w\right) \quad (14)$$

$$\arg\max \sum_{l=1}^{N} \sum_{j=1}^{N} \left(\left(\mathbf{y}_l - \mathbf{y}_j\right)^T \left(\mathbf{y}_l - \mathbf{y}_j\right) V_{jl}^b\right) \quad (15)$$

The objective functions (14) and (15) are formally similar, but their motivations are different. Eq. (14) characterizes the diversity of data and helps to unfolding the manifold structure while Eq. (15) encodes the local discriminating information of data.

3.4 Encode relationships among pixels

Motivated by [41] and [42], we integrate the relationships among pixels into the objective functions (14) and (15) by employing image Euclidean distance instead of the conventional Euclidean distance.

Definition 1 [41, 42] (Image Euclidean Distance, IED) Given two images $A_j \in R^{m\times n}$ and $A_l \in R^{m\times n}$, the image Euclidean distance between them is

$$d_{IED}\left(A_j, A_l\right) = \overline{Vec}\left(A_j - A_l\right)^T G\, \overline{Vec}\left(A_j - A_l\right)$$

where $\overline{Vec}\left(A_j - A_l\right) = \left[\left(\mathbf{a}_1^{jl}\right)^T \left(\mathbf{a}_2^{jl}\right)^T \quad \left(\mathbf{a}_m^{jl}\right)^T\right]^T$, $\mathbf{a}_i^{jl}$ denotes the i th row of $\left(A_j - A_l\right)$. The elements $G_{ii,jj}$ in symmetric matrix $G \in R^{LL\times LL}$ ($LL = m\times n$) can be defined as follows:

$$G_{ii,jj} = \frac{1}{2\pi}\exp(-[\left(p-p^{'}\right)^2 + \left(q-q^{'}\right)^2]/2)$$

where $ii = \left(p-1\right)n + q$, $jj = \left(p'-1\right)n + q'$, $p, p' = 1,2, \quad ,m$, $q, q' = 1,2, \quad ,n$.

Figure 2 shows an 2×2 image A that denotes the difference between two images and matrix G. The ith row in G shows the relationships between the ith pixel in A and all pixels. From G, we can see that the closer the location between two pixels is, the larger the value of the corresponding location in G is, it means that they have a large relationship. Obviously, the pixel and itself have the largest relationship than other pixels, so the value of diagonal elements in G is the largest. Thus, G characterizes the relationships among pixels in images. If G becomes identity matrix I , then IED becomes Euclidean distance.

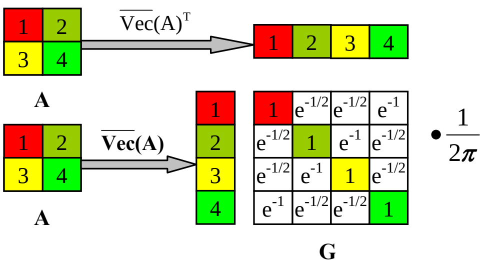

Figure 2. Difference images and symmetrical matrix G.

In image recognition, it is very difficult to directly calculate the matrix G due to the large size. Motivated by theorem 1, we give the matrix version of IED without calculating G.

Theorem 1 [43]. Suppose B_1 and B_2 are $m\times p$ and $n\times q$ matrices, $\mathbf{h}_i^T$ ($i = 1,2, \quad ,p$) denotes the ith row of $H \in R^{p\times q}$, and $\overline{Vec}(H) = \left[\mathbf{h}_1^T\ \mathbf{h}_2^T \quad \mathbf{h}_p^T\right]^T$. If $\mathbf{z} = \overline{Vec}(Z) = \left(B_1 \otimes B_2\right)\overline{Vec}(H)$, then $Z = B_1 H B_2^T$.

According to theorem 1, we have the corollary 1.

Corollary 1 (Matrix version of IED) For two images A_j and A_l , the matrix version of image Euclidean distance between two images is

$$d_{IED}\left(A_j, A_l\right) = \left\|\left(\hat{A}_j - \hat{A}_l\right)\right\|_F \tag{16}$$

where $\hat{A}_j = G_1^{1/2} A_j \left(G_2^{1/2}\right)^T$. The elements $G_1\left(p, p'\right)$ and $G_2\left(q, q'\right)$ in symmetric matrixes $G_1 \in R^{m\times m}$ and $G_2 \in R^{n\times n}$ can be respectively calculated as follows:

$$G_1(p, p') = \left(1/2\pi\right)^{1/2}\exp\{-(p-p^{'})^2/2\} \quad q, q' = 1, \quad ,n$$

$$G_2(q, q^{'}) = \left(1/2\pi\right)^{1/2}\exp\{-(q-q^{'})^2/2\} \quad p, p' = 1, \quad ,m$$

The proof of Corollary 1 is given in the appendix.

Let $\hat{\mathbf{y}}_i \in R^m$ be the low-dimensional representations of $\hat{A}_i$. We can incorporate the relationships among pixels in images into the objective functions (14) and (15) by using $\hat{\mathbf{y}}_i$ instead of $\mathbf{y}_i$. Thus, the objective functions (14) and (15) become respectively

$$\arg\max \sum_{l=1}^{N}\sum_{j=1}^{N}\left(\left(\hat{\mathbf{y}}_l - \hat{\mathbf{y}}_j\right)^T\left(\hat{\mathbf{y}}_l - \hat{\mathbf{y}}_j\right)V_{jl}^w\right) \tag{17}$$

$$\arg\max \sum_{l=1}^{N}\sum_{j=1}^{N}\left(\left(\hat{\mathbf{y}}_l - \hat{\mathbf{y}}_j\right)^T\left(\hat{\mathbf{y}}_l - \hat{\mathbf{y}}_j\right)V_{jl}^b\right) \tag{18}$$

3.5 Optimal linear embedding

Denote by $\boldsymbol{\alpha}$ the projection direction, substituting $\hat{\mathbf{y}}_i = \hat{A}_i\boldsymbol{\alpha}$ into Eq. (17). By the simple algebra formulation, we see that

$$\begin{aligned}
&\frac{1}{2}\left(\sum_{l=1}^{N}\sum_{j=1}^{N} V_{jl}^{w}\left(\hat{\mathbf{y}}_j - \hat{\mathbf{y}}_l\right)^T\left(\hat{\mathbf{y}}_j - \hat{\mathbf{y}}_l\right)\right) \\
&= \frac{1}{2}\boldsymbol{\alpha}^T\left[\sum_{j=1}^{N}\sum_{l=1}^{N} V_{jl}^{w}\left(\hat{A}_j - \hat{A}_l\right)^T\left(\hat{A}_j - \hat{A}_l\right)\right]\boldsymbol{\alpha} \\
&= \boldsymbol{\alpha}^T\left[\sum_{j=1}^{N}\left(\hat{A}_j^T\hat{A}_j\sum_{l=1}^{N}V_{jl}^{w}\right) - \sum_{j=1}^{N}\hat{A}_j^T\left(\sum_{l=1}^{N}V_{jl}^{w}\hat{A}_l\right)\right]\boldsymbol{\alpha} \\
&= \boldsymbol{\alpha}^T\left[\hat{A}^T\left(D\otimes I_m\right)\hat{A} - \hat{A}^T\left(V^w\otimes I_m\right)\hat{A}\right]\boldsymbol{\alpha} \\
&= \boldsymbol{\alpha}^T\hat{A}^T\left(L_w\otimes I_m\right)\hat{A}\boldsymbol{\alpha}
\end{aligned} \tag{19}$$

where $L_w = D - V^w$, D is a diagonal matrix whose entries $D_{j,j}$ are row (or column since V^w is symmetric) sum of V^w, i.e. $D_{j,j} = \sum_l V_{jl}^w \ or \ \sum_l V_{lj}^w$. $\otimes$ denotes the Kronecker product. I_m is an $m\times m$ identity matrix. $\hat{A}^T = \left[\hat{A}_1^T\ \hat{A}_2^T \quad \hat{A}_N^T\right]$.

Likewise, substituting $\hat{\mathbf{y}}_i = \hat{A}_i\boldsymbol{\alpha}$ into Eq. (18), we see that

$$\begin{aligned}
&\frac{1}{2}\left(\sum_{l=1}^{N}\sum_{j=1}^{N} V_{jl}^{b}\left(\hat{\mathbf{y}}_j - \hat{\mathbf{y}}_l\right)^T\left(\hat{\mathbf{y}}_j - \hat{\mathbf{y}}_l\right)\right) \\
&= \frac{1}{2}\boldsymbol{\alpha}^T\left[\sum_{j=1}^{N}\sum_{l=1}^{N} V_{jl}^{b}\left(\hat{A}_j - \hat{A}_l\right)^T\left(\hat{A}_j - \hat{A}_l\right)\right]\boldsymbol{\alpha} \\
&= \boldsymbol{\alpha}^T\left[\sum_{j=1}^{N}\left(\hat{A}_j^T\hat{A}_j\sum_{l=1}^{N}V_{jl}^{b}\right) - \sum_{j=1}^{N}\hat{A}_j^T\left(\sum_{l=1}^{N}V_{jl}^{b}\hat{A}_l\right)\right]\boldsymbol{\alpha} \\
&= \boldsymbol{\alpha}^T\left[\hat{A}^T\left(F\otimes I_m\right)\hat{A} - \hat{A}^T\left(V^b\otimes I_m\right)\hat{A}\right]\boldsymbol{\alpha} \\
&= \boldsymbol{\alpha}^T\hat{A}^T\left(L_b\otimes I_m\right)\hat{A}\boldsymbol{\alpha}
\end{aligned} \tag{20}$$

where $L_b = F - V^b$, F is a diagonal matrix whose entries $F_{j,j}$ are row (or column since V^b is symmetric) sum of V^b, i.e. $F_{j,j} = \sum_l V_{jl}^b \, or \sum_l V_{lj}^b$.

Finally, the optimal problem reduces to finding

$$\begin{aligned}\boldsymbol{\alpha}^* &= \arg\max_{\alpha^T\alpha=1} \boldsymbol{\alpha}^T \hat{A}^T \left(\left(L_b + aL_w\right) \otimes I_m\right) \hat{A}\boldsymbol{\alpha} \\ &= \arg\max_{\alpha^T\alpha=1} \boldsymbol{\alpha}^T \hat{A}^T \left(L_d \otimes I_m\right) \hat{A}\boldsymbol{\alpha}\end{aligned} \quad (21)$$

where $L_d = L_b + aL_w$, parameter a controls the balance between diversity and discriminating information.

The projection vector $\boldsymbol{\alpha}$ that maximizes (21) is given by the maximum eigenvalue λ solution to the generalized eigenvalue problem:

$$\hat{A}^T \left(L_d \otimes I_m\right) \hat{A}\boldsymbol{\alpha} = \lambda \boldsymbol{\alpha} \quad (22)$$

Let the column vector $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_d$ be the solutions of equation (22), ordered according to their eigenvalues, $\lambda_1 > \lambda_2 > \cdots > \lambda_d$. Thus, the embedding is as follows:

$$\begin{aligned} &A_i \to Y_i = A_i W \\ &W = \left[\boldsymbol{\alpha}_1 \; \boldsymbol{\alpha}_2 \; \cdots \; \boldsymbol{\alpha}_d\right]\end{aligned} \quad (23)$$

4. 2DMLV+2DPCA

Previous works have demonstrated that the low-dimensional representations $Y_i \in R^{m \times d}$ ($i = 1, \cdots, N$) contain many redundancies [44,45]. Moreover, the size of Y_i is still large, thus, the large storage space is required to save these low-dimensional representations. In this section, we introduce an effective approach, called 2DMLV+2DPCA, to further reduce the dimensionality of features. To be specific, we perform 2DPCA [10, 11] on the column vectors of Y_i.

Denote by $\overline{Y} = 1/N \sum_{j=1}^N Y_j$ the global mean of Y_i ($i = 1, 2, \cdots, N$), the projective matrix $V = \left[v_1 \; v_2 \; \cdots \; v_p\right]$ of 2DPCA is found by computing the p eigenvectors corresponding to the p largest eigenvalues of the covariance matrix S_p that can be defined as follows:

$$S_p = \frac{1}{N}\sum_{i=1}^N \left(Y_i - \overline{Y}\right)\left(Y_i - \overline{Y}\right)^T \quad (24)$$

Suppose W is the $n \times d$ projection matrix of 2DMLV. Projecting training faces A_i (i=1,…,n) onto W and V together, yielding the p by d feature matrices

$$Z_i = V^T \times A_i \times W \quad i = 1, \cdots, N \quad (25)$$

Given a test face image A^*, first using Eq. (25) to get the feature matrix $Z^* = V^T A^* W$, then a nearest neighbor classifier can be used for classification, i.e., if $d\left(Z^*, Z_k\right) = \min_i \left\{d\left(Z^*, Z_i\right)\right\}$, then the probe image A^* can be classified as the class to which training sample A_k belongs. $d\left(Z^*, Z_i\right)$, which denotes the Euclidean distance between Z^* and Z_i, and can be defined as

$$d\left(Z^*, Z_i\right) = \sum_{j=1}^d \left\|Z^*(:, j) - Z_i(:, j)\right\|_2 \quad i = 1, \cdots, N \quad (26)$$

where $Z^*(:, j)$ and $Z_i(:, j)$ denote the j th column of Z^* and Z_i, respectively.

5. Theoretical analysis of 2DMLV

5.1 Advantages of 2DMLV

Although our approach and LPP mainly preserve the local geometrical structure of data, their motivations are essentially different. LPP preserves the local geometrical structure of data by minimizing the distance among nearby data, i.e. variation of data. Thus, LPP mainly characterizes the similarity of data and does not unfold the structure of data. Moreover, LPP neglects the local discriminating information of data. In contrast to LPP, our approach preserves the local geometrical structure by maximizing the variation among nearby data. Thus, it mainly characterizes the diversity of data and helps to unfold the manifold structure [37-40]. Moreover, data points in a local neighborhood usually come from different classes, so we can encode the local discriminating information embedded in nearby data by our approach.

Different from LPP and our approach, PCA preserves the global geometrical structure of data by maximizing the global variation of data. Although PCA characterizes the diversity of data, it emphasizes the data points with large distance and deemphasizes the data points with small distance, resulting in the impairment of local topology and local discriminant structure of data. Taking the two-dimensional data points, which are randomly selected, in Figure 3 as an example, we show the projection directions of LPP, PCA and our approach, and their one-dimensional embedded results. For a reasonable comparison, we set the parameter a as 1 in our approach. From Figure 3, it is easy to see that our approach effectively preserves both the local diversity and local discriminating information of data, and separates these points. Note that, the points with the same shape belong to the same class.

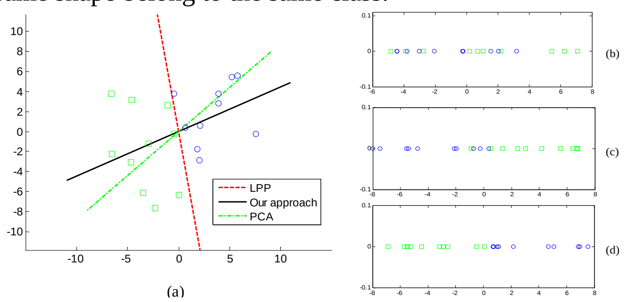

Figure 3. Difference between LPP, PCA and our approach.(a) one-dimensional embedding spaces obtained by LPP, PCA and Our approach, respectively; (b) One-dimensional embedded results obtained by LPP; (c) One-dimensional embedded results obtained by PCA; (d) One-dimensional embedded results obtained by our approach.

5.2 Parameter analysis

In this section, we discuss the effect of parameters *k* and t on the recognition accuracy of our approach, respectively. In the Yale database (please see Section 6.1 for details), we select the first 6 images per person for training and the remaining images for testing.

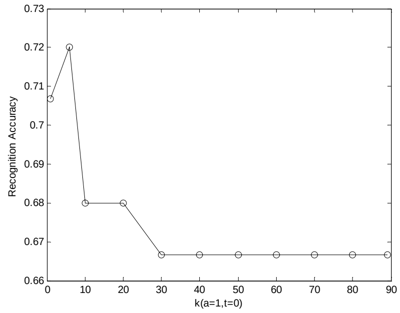

Figure 4. 2DMLV vs. parameter k (t=0 and a=1) on the Yale database.

Parameter k determines the size of neighborhood, if the value of k is very small, then the variation of data in the local neighborhood can not effectively characterize the diversity of data. In this case, nearby data points mainly characterize the geometrical properties of similarity. If the value of k is very large, then data points in the local neighborhood do not satisfy the Gaussian distribution, thus the variation of data may impair both the local intrinsic structure of data and local discriminating information. Thus, our approach is not good when k is very small or very large. Figure 4 shows the recognition accuracy of 2DMLV vs. parameter k ($t=0$ and $a=1$). From Figure 4, it can be seen that the above-mentioned analysis is correct.

Parameter t determines the role of data in characterizing the variation. If the value of t is very small, then all points are equally important in estimating the variation of data. If the value of t is very large, our approach will emphasize the variation between few data points with the large distance, and may impair the intrinsic structure of data and discriminating information. Thus, the diversity and discriminating information can not be detected well when t is very large or very small. Figure 5 plots the recognition accuracy of 2DMLV *vs.* parameter t when k ($k=6$) and $a=1$ are fixed. From Figure 5, it can be seen that 2DMLV is obviously deteriorating when the value of t becomes very large or small. It is consistent with above analysis.

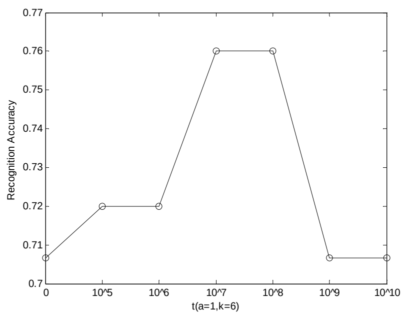

Figure 5. 2DMLV vs. parameter t (k=6 and a=1) on the Yale database.

6. Experiments and analysis

In this Section, we evaluate the performance of 2DMLV and 2DMLV+2DPCA on face recognition and compare with some classical linear dimensionality reduction approaches including 2DPCA [11], 2DLDA[12], 2DEFM (Two-dimensional form of EFM [14]), 2DLPP[31], DVPE[17], 2DMFA [4], and 2DSLSDP [35]. In 2DLPP, 2DMFA, 2DSLSDP, 2DMLV, and 2DMLV+2DPCA, the open problem is how to select the suitable parameters for dimension reduction. We therefore empirically set the parameters of these methods in the following experiments. To be specific, we sampled several values of parameters and chose the values with the best performance for all approaches. In our experiments, parameter a was set 1.

6.1 Face recognition

In this section, we used four face databases (Yale, AR, PIE and FERET) to show the effectiveness of our approach. The Yale face database [46] contains 165 grayscale images of 15 individuals (each person providing 11 different images) under various in facial expressions, lighting conditions, and with/without glasses. In the experiments, each image was manually cropped and resized to 32×32 pixels [3]. We respectively selected 3, 6, and 9 images per person as training, and the corresponding remaining images for testing. Thus, we have three training subsets, which include 45, 90, and 135, and the corresponding three testing subsets including 120, 75, and 30 images, respectively. Table 1 lists the top recognition accuracy of these approaches and the corresponding number of features. Figure 6 plots the recognition accuracy of these nine methods versus the number of projected vectors when 9 images per person are selected as training images.

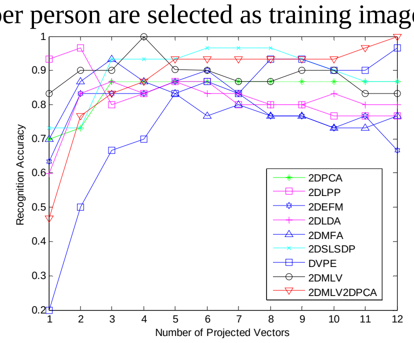

Figure 6. Recognition accuracy *vs.* the number of projected vectors on the Yale database.

Figure 7. Some sample images of one subject in the AR database.

The AR face database[47] is established by Purdue University, which contains over 4000 color face images of 126 people (70 men and 56 women) including frontal views of faces with different facial expressions, lighting conditions, and occlusions. The pictures of most persons were taken in two sessions (separated by two weeks). Each session contained 13 color images and 120 individuals (65men and 55 women) participate in both sessions. In the experiments, the facial portion of each image was manually cropped and then normalized to the size of 50×40 [11]. The images from the first session with (a) “neutral expression”, (b) “smile”, (c) “anger”, (d) “scream”, (e) “left light on”, (f) “right light on” and (g) “both side light on” were selected for training images, and the corresponding remain images from the second session, i.e. (n)-(t), for testing images. Figure 7 shows the training images and the corresponding testing images of one subject, respectively. Table 2 lists the top recognition accuracy and the corresponding dimension. Figure 8 plots the curves of recognition accuracy of nine approaches versus the number of projected vectors.

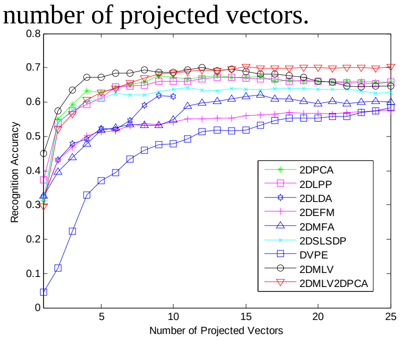

Figure 8. Recognition accuracy versus the number of projected *vs*. on the AR database.

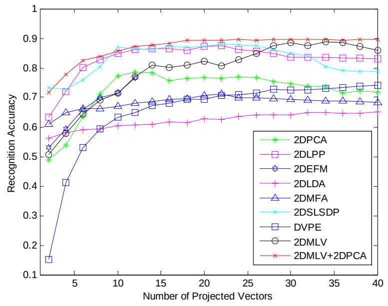

Figure 9. Recognition accuracy versus the number of projected vectors on the PIE database.

The CMU-PIE database [48] contains 68 subjects with 41368 face images as a whole. The face images were captured by 13 synchronized cameras and 21 flashes, under varying pose, illumination and expression. We selected pose-29 images as gallery, which includes 24 samples for each individual in the experiments. Each image was manually cropped and resized to 64×64 pixels [3]. The first 12 samples per person were used for training and the remaining 12 images for testing. Table 3 lists the top recognition accuracy and the corresponding dimension. Figure 9 shows a plot of recognition accuracy of nine approaches versus the number of projected vectors.

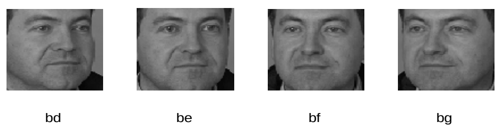

Figure 10. Some sample images of one subject in the FERET database.

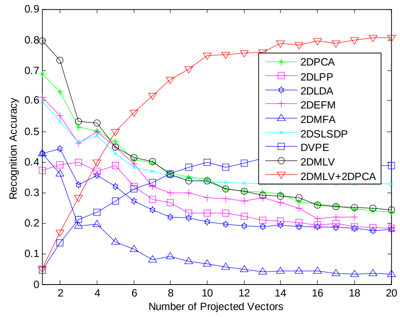

Figure 11. Recognition accuracy *vs*. the number of projected vectors on FERET database.

The proposed approach was also tested on a subset of the FERET database [49]. This subset includes 800 images of 200 individuals (each with four images). All the images were cropped and resized to 80×80 pixels based on the location of the two eyes. The images of one person are shown in Figure 10. In the experiment, we chose two images per person, which are labeled by be and bd in database, for training, and used the remaining two images, which are labeled bf and bg, for testing. Table 4 shows the experimental results of different algorithms, and Figure 11 shows the recognition accuracy versus the number of projected vectors.

In order to effectively evaluate the performance of our approach, we randomly selected 12 images per person for training and the corresponding remaining images for testing on the PIE database, and then repeated this process 10 times. Table 5 list the average recognition accuracy of these approaches and the corresponding standard deviation on PIE database. Moreover, in order to demonstrate the effectiveness of the image Euclidean distance and evaluate the role of parameter a in our approach, we compared 2DPCA, 2DMLV with Euclidean distance (2DMLV-ED) and 2DMLV with $a=1$ and $a \neq 1$ on the above-mentioned four databases. The top recognition accuracy and the corresponding dimension are listed in Table 6. Note that, $a=1$ means that our approach belongs to unsupervised approach. $a \neq 1$ means that our approach belongs to supervised approach. From Table 6, it is easy to see that 2DMLV (a=1) overall outperforms 2DMLV-ED (a=1) in all cases, and 2DMLV achieves the best recognition accuracy under $a \neq 1$.

We also compared the computation time of these approaches on the PIE database. Table 7 lists the training time and classification time of all approaches when they achieved the top recognition accuracy as in table 3. It can be seen that our approach needs the more time for training.

6.2 Analysis

Experiments have been performed on four face databases. These experiments reveal a number of interesting points:

(1) Our 2DMLV approach markedly outperforms 2DPCA. This is probably because that 2DPCA weakens or even impairs the local variation of face images. Moreover, 2DPCA does not encode the discriminating information. Another limitation of 2DPCA is that it is sensitive to the perturbation of face images.

(2) 2DMLV is superior to 2DLPP. This is probably because that 2DLPP ignores the most important geometrical properties of diversity of images. This may result in

unstable intrinsic structure representation of images. Another reason may be that 2DLPP does not well encode the discriminating information of face images.

(3) 2DMLV is overall superior to discriminant approaches DVPE, 2DLDA, 2DEFM, and 2DMFA. This is probably because that these discriminant approaches ignore the geometrical properties of diversity of faces from the same class in learning the within-class compactness. This leads to instability of geometrical structure representation and over-fitting.

(4) The top recognition of 2DMLV outperforms 2DSLSDP. This is probably due to the fact that 2DSLSDP considers within-class variation and between-class variation equally important, while the distance between nearby data from the same class is usually larger than that of data points from different classes in face recognition. Thus, 2DSLSDP can not well encode the local discriminating information. Another reason may be that 2DSLSDP ignores the relationships among pixels in images. It reduces the flexibility of algorithm.

(5) The top recognition accuracy of 2DMLV+2DPCA overall outperforms that of 2DMLV. This is probably because that 2DPCA can effectively reduce the redundancy embedded in 2DMLV. The top recognition accuracy of 2DMLV and 2DMLV+ 2DPCA is overall superior to other methods. This is probably because that they well unfold the local manifold structure and detect both the diversity of images and discriminating information.

(6) 2DMLV is superior to 2DMLV-ED when a is 1. This is probably because that image Euclidean distance considers the relationship among pixels in image, which helps to improving the stability of the algorithm to the perturbation of images. 2DMLV with a=1 overall superiors to 2DMLV-ED with a≠1. And, compared with the performance with a=1, the top recognition accuracy of 2DMLV can not be improved significantly with a≠1. This is probably because that image Euclidean distance helps to well encoding the local discriminating information of images.

(7) Compared with the other approaches, 2DMLV needs the more time for calculating the projection matrix. This is probably because that 2DMLV spends more time to calculate the image Euclidean distance, but it has no effect for the classification time. All approaches need different time for classifying all testing image. This is due to the fact that each approach achieves the top recognition accuracy with different dimensionality of features.

7. Conclusion

We introduce a novel linear dimensionality reduction approach called two-dimensional maximum local variation (2DMLV). Our approach 2DMLV preserves intrinsic structure by maximizing the variation of the images. Our method differs in that we measure the variation of images, which characterizes the diversity of images, by using image Euclidean distance instead of the traditional Euclidean distance. Image Euclidean distance considers the relationships among pixels in images and helps to well encoding discriminating information of images. Experimental results on four face databases demonstrate that our approach 2DMLV is overall superior to other two-dimensional dimensionality reduction approaches. Experiments also indicate that the accuracy can be further improved by combining 2DMLV and 2DPCA together.

Acknowledgements

The authors would like to thank the anonymous reviewers and associate editor for their invaluable comments and suggestions, which have helped improve the presentation of this paper greatly. We would also like to thank Prof. Xiaofei He for invaluable suggestions. This work is supported by the National Science Foundation of China under Grant 60802075, 61271296 and 61172068, Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2012JM8002, Fundamental Research Funds for the Central Universities of China, China Postdoctoral Science Foundation under Grant 2012M521747, and the Open Project Program of the State Key Laboratory of CAD&CG, Zhejiang University, China. This work is also supported partially by 111 Project of China (B08038) and the Natural Science Foundation of Zhejiang Province under Grant LY12F03004.

Appendix

Proof of Corollary 1

By simple algebra formulation, symmetric matrix G can be rewritten as

$$G = G_1 \otimes G_2 \tag{a-1}$$

where G_1 and G_2 are of the size $m \times m$ and $n \times n$ matrices, respectively. The elements $G_1(i,i')$ and $G_2(j,j')$ can be defined as follows

$$G_1(i,i') = (1/2\pi)^{1/2} \exp\{-(i-i')^2/2\} \qquad i,i'=1, \quad ,m$$

$$G_2(j,j') = (1/2\pi)^{1/2} \exp\{-(j-j')^2/2\} \qquad j,j'=1, \quad ,n$$

Image Euclidean distance $d_{IED}(A_j, A_l) = \overline{Vec}(A_j - A_l)^T G\, \overline{Vec}(A_j - A_l)$ can be rewritten as

$$\begin{aligned} d_{IED}(A_j, A_l) &= \overline{Vec}(A_j - A_l)^T G^{1/2} G^{1/2}\, \overline{Vec}(A_j - A_l) \\ &= \left(G^{1/2}\overline{Vec}(A_j - A_l)\right)^T \left(G^{1/2}\overline{Vec}(A_j - A_l)\right) \end{aligned} \tag{a-2}$$

Denote by $z = \overline{Vec}(Z) = (G_1 \otimes G_2)^{1/2}\, \overline{Vec}(A_j - A_l)$, according to theorem 1, we have

$$Z = \left(G_1^{1/2}(A_j - A_l)(G_2^{1/2})^T\right) \tag{a-3}$$

Substituting (a-3) into (a-2), we see that

$$d_{IED}(A_j, A_l) = z^T z = \|Z\|_F = \left\|\hat{A}_j - \hat{A}_l\right\|_F \tag{a-4}$$

Where $\hat{A}_j = G_1^{1/2} A_j (G_2^{1/2})^T$, $\|\bullet\|_F$ denotes the Frobenius norm of matrix.

References

[1] M. Turk and A. P. Pentland, “Face recognition using eigenfaces,” *In Proc. Int. Conf. Computer Vision and Pattern Recognition*, 1991.

[2] P. Belhumeur, P. Hespanha, and D. Kriegman, “Eigenfaces vs fisherfaces: recognition using class specific linear projection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19,

pp. 711-720, 1997.

[3] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, "Face recognition using laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 27, no. 3, pp. 328-340, 2005.

[4] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: a general framework for dimensionality reduction,". *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 29, no. 1, pp. 40-51, 2007.

[5] T. Zhang, D. Tao, X. Li, and J. Yang, "Patch alignment for dimensionality reduction," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp.1299–1313, 2009.

[6] X. D. Jiang, B. Mandal, and A. Kot, "Eigenfeature regularization and extraction in face recognition", *IEEE Trans. Pattern Anal. Machine Intel*l. Vol. 30, no. 3, pp. 383-394, 2008.

[7] R. He, B. g. Hu, W. S. Zheng, and X. W. Kong, "Robust principal component analysis based on maximum correntropy criterion," *IEEE Trans. Image Processing.*, vol. 20, no. 6, pp. 1485-1494, 2011.

[8] S. J. Wang, J. Yang, N. Zhang, and C. G. Zhou, "Tensor discriminant color space for face recognition", *IEEE Trans. Image Processing*, vol. 20, no. 9, pp. 2490-2501, 2011.

[9] K. Fukunaga, Introduction to statistical Pattern Recognition, Second e.d., Academic Press, 1990.

[10] J. Yang and J. Y. Yang, "From image vector to matrix: a straightforward image projection technique-IMPCA vs. PCA", *Pattern Recognition*, vol. 35, pp. 1997-1999, 2002.

[11] J. Yang, D. Zhang, A. F. Frangi, and J. Y. Yang, "Two-Dimensional PCA: a new approach to appearance-based face representation and recognition," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 26, no. 1, pp. 131-137, 2004.

[12] J. Ye, R. Janardan, and Q. Li, "Two-dimensional linear discriminant analysis", In Proceedings of Neural Information Processing System (NIPS), 2004.

[13] X. D. Jiang, "Asymmetric principal component and discriminant analyses for pattern classification," *IEEE Trans. Pattern Anal. Machine Intell.* Vol. 31, no.5, pp.931-937, 2009.

[14] C. Liu and H. Wechsler, "Gabor Feature Based Classification Using the Enhanced Fisher Linear Discriminant Analysis for Face Recognition," *IEEE Trans. Image Processing*, vol. 11, no. 4, pp. 467-476, 2002.

[15] D.L. Swets, and J. Weng, "Using discriminant eigenfeatures for image retrieval", *IEEE Trans. Pattern Anal. Mach. Intell*. Vol. 18, no. 8, pp. 831–836, 1996.

[16] D. Tao, X. Li, X. Wu, and S. J. Maybank, "General tensor discriminant analysis and gabor features for gait recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1700–1715, Oct. 2007.

[17] B. Mandal, X. Jiang, H.L. Eng, and A. Kot, "Prediction of eigenvalues and regularization of eigenfeatures for human face verification", Pattern Recognition Letters, vol. 31, pp. 717-724, 2010.

[18] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction", *Science*, vol. 290, pp. 2320-2323, 2000.

[19] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no, 6, pp. 1373-1396, 2003.

[20] B. Schölkopf, A. Smola, and K. Mü ller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput*., vol. 15, no. 5, pp. 1299-1319, 1998.

[21] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Computation*, vol. 12, no. 10, pp. 2385-2404, 2000.

[22] J. Ham, D.D. Lee, S. Mika, and B. Schölkopf, "A kernel view of the dimensionality reduction of manifolds," *Technical Report*, *University of Pennsylvania*, 2003.

[23] D. Cai, X. He, J. Han, and H. Zhang, "Orthogonal Laplacianfaces for Face Recognition," *IEEE Trans. Image Processing*, vol. 15, no. 11, pp. 3608-3614, Nov. 2006.

[24] D. Xu, S. Yan, D. Tao, S. Lin, and H. Zhang, "Marginal fisher analysis for human gait recognition and content based image retrieval," *IEEE Trans. Image Processing*, vol. 16, no. 11, pp. 2811–2821, 2007.

[25] H. Wang, H. Chang, and T. Liu, "Local discriminant embedding and its variant", *IEEE International Conf. Computer Vision (ICCV)*,vol.2,pp.846-853, 2005.

[26] M. Sugiyama, "Local fisher discriminant analysis for supervised dimensionality reduction," *In Proc. Int. Conf. Machine Learning (ICML)*, pp. 905-912, 2006.

[27] B. Li, C. Zheng, and D. S. Huang, "Locally linear discriminant embedding: an efficient method for face recognition", *Pattern Recognition*, vol. 41, pp. 3813-3821 2008.

[28] H. Wang, S. Che, Z. Hu, and W. Zheng, "Locality-preserved maximum information projection", *IEEE Trans. Neural Networks*, vol. 19, no. 4, pp. 571-585, 2008.

[29] C. Hou, C. Zhang, Y. Wu, and Y. Jiao, "Stable local dimensionality reduction approaches"，*Pattern Recognition*, vol. 42, pp. 2054-2066, 2009.

[30] M. Cheng, B. Fang, Y. Y. Tang, T. Zhang, and J. Wen, "Incremental embedding and learning in the local discriminant subspace with application to face recognition", *IEEE Trans. Systems, Man, and Cybernetics, Part C: Applications and Reviews, pp. 580-591*, 2010.

[31] X. He, D. Cai, and P. Niyogi, "Tensor Subspace Analysis," In Advances in *Neural Information Processing Systems (NIPS'05), Vancouver, Canada*, 2005.

[32] D. Hu, G. Feng, and Z. Zhou, "Two-dimensional locality preserving projections (2DLPP) with its application to palmprint recognition", *Pattern Recognition*, vol. 40, pp. 339-342, 2007.

[33] N. Guan, D. Tao, Z. Luo, and B. Yuan, "Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent", *IEEE Trans. Image Processing*, vol. 20, no. 7, pp. 2030-2048, 2011.

[34] Y. Xu, G. Feng, and Y. Zhao, "One improvement to two-dimensional locality preserving projection method for use with face recognition", *Neurocomputing*, vol. 73, pp: 245-249, 2009.

[35] Q. Gao, H. Xu, Y. Li, and D. Xie, "Two-dimensional supervised local similarity and diversity projection", *Pattern recognition*, vol. 43, no. 10, pp. 3359-3363, 2010.

[36] D .Cai, X. He, and J. Han, "Isometric Projection," *In Proc. Association for the Advancement of Artificial Intelligence (AAAI)*, pp. 528-533, 2007.

[37] J. Sun, S. Boyd, L. Xiao, and P. Diaconis, "The fastest mixing markov process on a graph and a connection to a maximum variance unfolding problem", SIAM Review, vol. 48, no. 4, pp. 681-699, 2006.

[38] K.Q. Weinberger, and L.K. Saul, "An introduction to nonlinear dimensionality reduction by maximum variance unfolding", In *Proc. the 21^{th} AAAI*, pp. 1683-1686, 2006.

[39] D. Luo, C. Ding, F. Nie, and H. Huang, "Cauchy graph embedding", In *Proc. the 28th International Conference on Machine Learning (ICML)*, 2011.

[40] Q. Gao, J. Liu, H. Zhang, J. Hou, and J. Yang, "Enhanced fisher discriminant criterion for image recognition", *Pattern Recognition*, vol. 45, pp. 3717-3724, 2012.

[41] L. Wang, Y. Zhang, and J. Feng, " On the Euclidean distance of images", *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 8, pp. 1334-1339, 2005.

[42] M. Postnikov, Lectures in Geometry, Semester I, Analytic Geometry, translated by V. Shokurov, Moscow: Mir

Publishers, 1982.
[43] Joel N. Franklin. Matrix Theory. Dover Publications, *Inc, New York, NY*,2000.
[44] Q. Gao, L. Zhang, and D. Zhang, “Sequential row-column independent component analysis for face recognition”, *Neurocomputing*, vol. 72, pp. 1152-1159, 2009.
[45] W. Zuo, D. Zhang, and K. Wang, "Bidirectional PCA with assembled matrix distance metric for image recognition”, *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 36, no.4, pp 863-872, 2006.
[46] Yale Univ. Face Database, http://cvc.yale.edu/projects /yalefaces/yalefaces.html, 2002.
[47] The AR Face Database, http://rvll.ecn.purdue.edu /~aleix/aleix_face_DB.html.
[48] T. Sim, S. Baker, and M. Bsat. The CMU Pose, Illumination, and Expression (PIE) Database. *In Proc. IEEE Int’l Conf. Automatic Face and Gesture Recognition*, 2002.
[49] P. Phillips, H. Moon, S. Rizvi, and P. Rauss, “The FERET evaluation methodology for face- recognition algorithms,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090-1104, 2000.

Quanxue Gao received the B. Eng. degree from Xi’an Highway University, China, in 1998 and the Ph.D. degree in control theory and control engineering from Northwestern Polytechnical University, China, in 2005. From 2006 to 2007, he was a research associate in the Department of Computing at The Hong Kong Polytechnic University. He is currently an associate Professor in School of Telecommunications Engineering, Xidian University. His research interests include pattern recognition, dimensionality reduction, manifold learning, semi-supervised learning, and sparse representation.

Feifei Gao received her bachelor degree from Xi’an University of Post and Telecommunications in 2009. She is currently working toward the M.S. degree in traffic information engineering and control from Xidian University, China. Her research interests include pattern recognition, Semi-supervised learning, and face recognition.

Hailin Zhang received the B.S. and M.S. degrees from Northwestern Polytechnical University, China, in 1985 and 1988 respectively, and the Ph.D. form Xidian University, China, in 1991, all in electronic information engineering. He is now a senior Professor and a Ph.D. adviser with School of Telecommunications Engineering at Xidian University. He is currently the Dean of this school, the Director of Key Laboratory in Wireless Communications Sponsored by China Ministry of Information Technology, a key member of State Key Laboratory of Integrated Services Networks, a field leader in Telecommunications and Information Systems in Xidian University, and an associate Director for National 111 Project (a National Base for International Exchange of Talented Researchers in Fundamental Theory and Technologies of Wireless Communications and Networking)

His research interests include key transmission technologies and standards on broadband wireless communications for B3G, 4G and next generation broadband wireless access systems. He has recently published 78 papers in core journals and conferences.

Xiujuan Hao Master candidate at Xidian University.Her research interest cover statistical patter recognition， machine learning and face recognition.

Xiaogang Wang (S'03-M'10) received the B.S. degree from University of Science and Technology of China in Electrical Engineering and Information Science in 2001, and the M.S. degree from Chinese University of Hong Kong in Information Engineering in 2004. He received the PhD degree in Computer Science from the Massachusetts Institute of Technology. He is currently an assistant professor in the Department of Electronic Engineering at the Chinese University of Hong Kong. He was the Area Chair of IEEE International Conference on Computer Vision (ICCV) 2011. He is the associate editor of Image and Visual Computing Journal. He received the Outstanding Young Researcher in Automatic Human Behavior Analysis award in 2011 and Hong Kong RGC Early Career Award in 2012. His research interests include computer vision and machine learning.

Table 1. The top classification accuracy (%) on the Yale database and the corresponding dimension (Shown in parentheses).

Training/Testing number	DVPE	2DPCA	2DLPP	2DLDA	2DEFM	2DMFA	2DSLSDP	2DMLV	2DMLV+2DPCA
45/120	68.85 (10)	61.67 (32*4)	65.00 (32*4)	56.67 (32*3)	61.67 (32*3)	52.50 (32*3)	68.33 (32*4)	**71.67 (32*2)**	**71.67 (28*2)**
90/75	**82.67 (9)**	68.00 (32*1)	77.33 (32*2)	73.33 (32*4)	76.00 (32*4)	74.67 (32*2)	80.00 (32*2)	**80.00 (32*2)**	**81.33 (11*2)**
135/30	96.67 (12)	86.67 (32*3)	96.67 (32*2)	86.67 (32*3)	90.00 (32*3)	93.33 (32*3)	96.67 (32*6)	**100.00 (32*4)**	**100.00 (11*4)**

Table 2. Top recognition accuracy (%) on the AR database and the corresponding dimension

Methods	DVPE	2DPCA	2DLPP	2DLDA	2DEFM	2DMFA	2DSLSDP	2DMLV	2DMLV+2DPCA
Recognition accuracy	66.67	67.74	67.14	58.57	61.90	62.14	64.29	**70.00**	**70.36**
Dimension	117	50*13	50*13	50*26	50*9	50*16	50*11	**50*12**	**15*12**

Table 3. Top recognition accuracy (%) on the PIE database and the corresponding dimension

Methods	DVPE	2DPCA	2DLPP	2DLDA	2DEFM	2DMFA	2DSLSDP	**2DMLV**	**2DMLV+2DPCA**
Recognition Accuracy	77.70	81.62	87.87	65.93	76.84	71.08	88.11	**89.34**	**89.71**
Dimension	51	64*13	64*21	64*47	64*12	64*22	64*22	**64*35**	**24*35**

Table 4. Top recognition accuracy (%) on the FERET database and the corresponding dimension

Methods	DVPE	2DPCA	2DLPP	2DLDA	2DEFM	2DMFA	2DSLSDP	**2DMLV**	**2DMLV+2DPCA**
Recognition Accuracy	41.25	68.75	40.00	44.25	61.25	42.75	60.00	**79. 50**	**80.75**
Dimension	13	80*1	80*3	80*2	80*1	80*1	80*1	**80*1**	**19*1**

Table 5. The average accuracy (%) of nine approaches on the PIE database and the corresponding standard deviation.

Methods	DVPE	2DPCA	2DLPP	2DLDA	2DEFM	2DMFA	2DSLSDP	**2DMLV**	**2DMLV+2DPCA**
Average Accuracy	93.81	94.30	90.36	91.26	93.31	90.22	94.03	**95.22**	**95.40**
Standard Deviation	±6.24	±5.16	±5.41	±9.65	±9.59	±8.02	±4.09	**±3.49**	**±3.29**

Table 6. Top recognition accuracy (%) of four approaches and the corresponding dimension (shown in parentheses)

Database		Yale			AR	PIE	FERET
		45/120	90/75	135/30			
2DMLV-ED	a=1	70.83(32*3)	78.67(32*2)	100(32*4)	69.52(50*19)	87.87(64*36)	76.25(80*1)
	a≠1	71.67(32*3)	**80.00(32*2)**	100(32*4)	69.64(50*7)	89.22(64*46)	**79.00(80*1)**
2DMLV	a=1	**71.67(32*3)**	**80.00(32*2)**	**100(32*4)**	**70.00(50*12)**	**89.34(64*35)**	**79.50(80*1)**
	a≠1	**71.67(32*2)**	**81.33(32*2)**	**100(32*4)**	**70.00(50*12)**	**89.34(64*35)**	**79.75(80*1)**

Table 7. Training time and testing time (seconds) of nine approaches on the PIE database.

Approaches	DVPE	2DPCA	2DMFA	2DEFM	2DLPP	2DLDA	2DSLISP	2DMLV	2DMLV+2DPCA
Training	8.8594	10.0625	37.8125	37.8281	21.7813	11.2188	21.7188	560.9375	595.578
Testing	10.2344	13.7813	33.7813	12.8906	33.2188	112.8438	33.7813	90.875	89.980