Research Article

A Unified Factors Analysis Framework for Discriminative Feature Extraction and Object Recognition

Ningbo Hao,^{1,2} Jie Yang,¹ Haibin Liao,^{3,4} and Wenhua Dai³

¹*School of Information Engineering, Wuhan University of Technology, Wuhan 430070, China*

²*School of International College, Huanghuai University, Zhumadian, Henan 463000, China*

³*School of Computer Science and Technology, Hubei University of Science and Technology, Xianning 437100, China*

⁴*National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China*

Correspondence should be addressed to Haibin Liao; liao_haibing@163.com

Received 22 October 2015; Revised 7 March 2016; Accepted 28 March 2016

Academic Editor: Paolo Crippa

Copyright © 2016 Ningbo Hao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Various methods for feature extraction and dimensionality reduction have been proposed in recent decades, including supervised and unsupervised methods and linear and nonlinear methods. Despite the different motivations of these methods, we present in this paper a general formulation known as factor analysis to unify them within a common framework. During factor analysis, an object can be seen as being comprised of content and style factors, and the objective of feature extraction and dimensionality reduction is to obtain the content factor without style factor. There are two vital steps in factor analysis framework; one is the design of factor separating objective function, including the design of partition and weight matrix, and the other is the design of space mapping function. In this paper, classical Linear Discriminant Analysis (LDA) and Locality Preserving Projection (LPP) algorithms are improved based on factor analysis framework, and LDA based on factor analysis (FA-LDA) and LPP based on factor analysis (FA-LPP) are proposed. Experimental results show the superiority of our proposed approach in classification performance compared to classical LDA and LPP algorithms.

1. Introduction

Feature extraction and selection are both a critical and challenging component of pattern recognition [1, 2]. Some basic principles can be used to guide the design of a feature extractor; however, the design work is essentially an experimental science depending on its field and problem being addressed. The utilization of algorithms for dimensionality reduction in supervised or unsupervised learning tasks has attracted a lot of attention in recent years. The relative simplicity and effectiveness of the linear algorithms for both Principal Component Analysis (PCA) [3] and Linear Discriminant Analysis (LDA) [3, 4] have made these two algorithms very attractive. The Locality Preserving Projections (LPP) [5] and Discriminant Canonical Correlations (DCC) [6] are other linear algorithms that have been proposed for dimensionality reduction. Three other recently developed algorithms—termed ISOMAP [7], LLE [8], and Laplacian Eigenmap [9]—can be utilized when conducting nonlinear dimensionality

on a dataset that lies on or around a lower-dimensional manifold. Additionally, in order to extend the linear dimensionality reduction algorithms to nonlinear ones, the kernel trick approach [10] has also been applied to perform linear operations on higher or infinite dimensional features which are transformed by a kernel mapping function. What is more, a number of algorithms [11–14] have recently been proposed to carry out dimensionality reduction on objects encoded as matrices or tensors of an arbitrary order.

All the aforementioned algorithms have been designed according to specific intuitions. Solutions are given through the optimization of both intuitive and pragmatic objectives, with the aim of extracting low-dimensional, simplified, and robust features from the high-dimensional, redundant, disturbed, and distorted original data. Content and style factors can be seen as the two components of an object [15]. During pattern recognition field, the information explored as base of recognition is defined as content factor and the information disturbed recognition is defined as style factor.

FIGURE 1: Face factors analysis (the content factors which without illumination and poses are what we want to extract and the style factors which with illumination and poses are what we want to remove).

Two examples are given below. One is for face recognition. A front face with neutral expression, no occlusion, and normalized illumination represents the content factor of a face image, whereas a face with variation of poses, illumination, and expressions represents the style factor, as shown in Figure 1. The occurrence of interference between content and style factors necessitates their separation in order to achieve accurate classification capabilities. The other is for speech recognition. In the process of recognizing the text of speech, the style factor includes timbre, tone, and emotion of the speaker. These elements should be regarded as background noise and minimized, as they do not provide any useful information related to speech text. However, the situation is opposite for speaker recognition since the favorable element is timbre information, which is defined as content factor here. The influence of speech text which is defined as style factor should be minimized. Thus the primary objective of the linear, nonlinear, manifold, or tensor decomposition methods is to reduce the style factors of the object.

In this paper, we present a general framework called factors analysis, along with its linearization, kernelization, and tensorization, which offers a unified view for understanding and explaining many of the popular dimensionality reduction algorithms such as the ones mentioned above. The main objective of factor analysis is to find the essential low-dimensional features by simultaneously minimizing the style factors and maximizing the content factors of the object through the method of projection. There are two vital steps in the factor analysis framework. The first is to construct a factor separating objective function which includes the construction of a partition and weight matrix. The second step is to design the space mapping function. Using the factor analysis framework as a platform, we develop two novel dimensionality reduction algorithms, FA-LDA and FA-LPP. FA-LDA is to overcome the limitations of LDA, which solves the problem that the intraclass and interclass scatter matrix in LDA are not optimal through setting proper weights. FA-LPP is to

overcome the limitations of LPP, which improves the design of partition so that the objects not only preserve the locality similarity but also have abundant identification information.

2. Factors Analysis: A General Framework for Feature Extraction

2.1. Factors Analysis. Content and style factors can be seen as two independent elements which influence the object and determine the observation [15], as mentioned before. The information contributed to recognition is defined as content factor and the negative information is style factor during pattern recognition.

For a general classification problem, the sample set is represented as $X = [x_1, x_2, \dots, x_N]$, $x_i \in \mathcal{R}^m$, where N denotes the sample number and m denotes the feature dimension. In practice, the feature dimension m is very high usually, including redundancy and interference information. In order to alleviate the curse of dimensionality, it is necessary to transform the original data X from a high-dimensional space to a low-dimensional space which is represented as $Y = [y_1, y_2, \dots, y_N]$, $y_i \in \mathcal{R}^{m'}$. The essence of dimensionality reduction is to find a mapping function $F: x \rightarrow \hat{y}$ that transforms $x \in \mathcal{R}^m$ into appropriate low-dimensional space $\hat{y} \in \mathcal{R}^{m'}$, where typically $m \gg m'$:

$$\hat{y} = F(x). \quad (1)$$

According to the various cases, the mapping function can be explicit or implicit and linear or nonlinear. A dimensionality reduction algorithm based on factor analysis is proposed in this paper, which is different from traditional methods. The objective of the dimensionality reduction algorithm based on factor analysis is to find a low-dimensional dataset Y from the original dataset X , in which the style factor is suppressed

as much as possible. The objective function is represented as follows:

$$\begin{aligned}
 y^* &= \arg \max \frac{F'_c}{F'_s} \\
 &= \arg \max \frac{\sum_{i=1}^{C-1} \sum_{j=i+1}^C p_i p_j \|\mu'_i - \mu'_j\|^2 W_{ij}^c}{\sum_{i=1}^C p_i \sum_{j=1}^{N_i-1} \sum_{k=j+1}^{N_i} \|y_j^{(i)} - y_k^{(i)}\|^2 W_{jk}^s} \quad (2) \\
 &= \arg \min_{F'_c=d} \sum_{i=1}^C p_i \sum_{j=1}^{N_i-1} \sum_{k=j+1}^{N_i} \|y_j^{(i)} - y_k^{(i)}\|^2 W_{jk}^s,
 \end{aligned}$$

where F'_c , F'_s are the content and style variance matrices after mapping (factors separating), respectively, C is the total number of partitions, N_i is the sample number of the i th partition (including the i th class for supervised methods and the i th local manifold space for unsupervised methods, generally named as the i th partition), μ'_i is the mean vector of the i th partition after factors separating, $y_j^{(i)}$, $y_k^{(i)}$ are the j th and k th sample vectors of the i th partition, respectively, after factors separating, W_{ij}^c , W_{jk}^s are weight matrices of content variance and style variance, respectively, and d is the constraint constant, which is different in different cases. Factor analysis criteria make the content variance of the objects between partitions as high as possible and the style variance of the objects within a partition as low as possible. In order to eliminate the style factor of the objects maximally,

the style factor should be minimized and the content factor should be maximized after mapping transformation.

2.2. General Framework for Feature Extract. There are two vital steps for feature extraction based on factor analysis framework; one is the design of factor separating objective function, mainly including the design of partition and weight matrix, and the other is the design of space mapping function. Partition is designed based on the class information with supervised methods or local neighborhood through unsupervised or semisupervised methods of the samples. Weight matrix is designed according to the classification performance for purpose of finding the optimal distribution scheme, which is biased to the weakly separated samples in the case without sacrificing the strongly separated samples. Space mapping function is designed to transform the samples from input space to feature space. Thus, the research of factor analysis framework can proceed from the following two aspects: the first is to propose different space mapping transformation aimed at different applications, and the second is to find the design scheme of factor separating objective function that matches the pattern recognition best.

2.2.1. Space Mapping Transformation. There are three categories of space mapping transformation algorithms, which are linear space, kernel space, and tensor space mapping transformation, respectively.

(a) Linear Space Mapping Transformation. The relationship of datasets X and Y is supposed to be linear in linear space mapping transformation; namely, $y_i = \psi^T x_i$. Thus the objective function (see (2)) is converted to

$$\begin{aligned}
 J(\psi) &= \arg \max \frac{\text{trace} \left(\sum_{i=1}^{C-1} \sum_{j=i+1}^C p_i p_j \psi^T (\mu_i^0 - \mu_j^0) W_{ij}^c (\mu_i^0 - \mu_j^0)^T \psi \right)}{\text{trace} \left(\sum_{i=1}^C p_i \sum_{j=1}^{N_i-1} \sum_{k=j+1}^{N_i} \psi^T (x_j^{(i)} - x_k^{(i)}) W_{jk}^s (x_j^{(i)} - x_k^{(i)})^T \psi \right)} = \arg \max \frac{\text{trace}(\psi^T F_c \psi)}{\text{trace}(\psi^T F_s \psi)} \quad (3) \\
 &= \arg \min_{\psi^T F_c \psi = d} \text{trace}(\psi^T F_s \psi),
 \end{aligned}$$

where μ_i^0 is the mean vector of the i th partition in original space, $x_j^{(i)}$, $x_k^{(i)}$ are the j th and k th sample vectors of the i th partition samples in original space, and F_c , F_s are the content and style variance matrices in original space, respectively, defined as follows:

$$F_c = \sum_{i=1}^{C-1} \sum_{j=i+1}^C p_i p_j W_{ij}^c (\mu_i^0 - \mu_j^0) (\mu_i^0 - \mu_j^0)^T, \quad (4)$$

$$F_s = \sum_{i=1}^C p_i \sum_{j=1}^{N_i-1} \sum_{k=j+1}^{N_i} W_{jk}^s (x_j^{(i)} - x_k^{(i)}) (x_j^{(i)} - x_k^{(i)})^T.$$

(b) Kernel Space Mapping Transformation. The direct way to extend methods from linear projections to nonlinear cases is

to utilize the kernel trick. The intuition of the kernel methods is to map the data from the original input space to a higher-dimensional Hilbert space (kernel space) as $\phi: x \rightarrow H$. The content and style variance matrices F_c , F_s in kernel space H are defined as follows:

$$\begin{aligned}
 F_c^\phi &= \sum_{i=1}^{C-1} \sum_{j=i+1}^C p_i p_j W_{ij}^c (\mu_{i0}^\phi - \mu_{j0}^\phi) (\mu_{i0}^\phi - \mu_{j0}^\phi)^T, \\
 F_s^\phi &= \sum_{i=1}^C p_i \sum_{j=1}^{N_i-1} \sum_{k=j+1}^{N_i} W_{jk}^s (\phi(x_j^{(i)}) - \phi(x_k^{(i)})) \cdot (\phi(x_j^{(i)}) - \phi(x_k^{(i)}))^T, \quad (5)
 \end{aligned}$$

where $\mu_{i0}^\phi = (1/N_i) \sum_{i=1}^{N_i} \phi(x_i^0)$ is the mean vector of the i th partition samples in kernel space H and $\phi(x_j^{(i)})$ is the kernel mapping to $x_j^{(i)}$.

(c) *Tensor Space Mapping Transformation.* Tensorization is another extension of linear methods, and it is a multiple linear extension. Firstly some terminologies [16] on tensor operations are reviewed before introducing the tensor space mapping transformation.

Tensor is the extension of vector and matrix. Tensor $A \in \mathfrak{R}^{I_1 \times I_2 \times \dots \times I_r}$ represents m^r components in m -dimension space, where every component is a function of coordinates, linearly changing with the coordinates following certain rules. r is the order of the tensor. When $r = 1$ A is simplified to a vector and

when $r = 2$ A is simplified to a matrix. Expanding the tensors can facilitate its matrix representation. The k -order expansion of tensor A is $A_{(k)} \in \mathfrak{R}^{I_k \times (I_{k+1} \dots I_r I_1 I_2 \dots I_{r-1})}$. The k -mode product between a tensor A and a matrix $U \in \mathfrak{R}^{I_k \times I_k}$ is defined as $B = A_{\times k} U$, or $B_{(k)} = U A_{(k)}$ for the form of tensor expansion.

Reform the original high-dimensional samples as the tensor $\{X_i \in \mathfrak{R}^{I_1 \times I_2 \times \dots \times I_r}, i = 1, 2, \dots, N\}$. Corresponding low-dimensional tensors as (6) are obtained from high-dimensional tensors through projecting (multiple linear projection), which is similar to linear methods:

$$y_i = X_i \times_1 \psi^1 \times_2 \psi^2 \dots \times_r \psi^r. \quad (6)$$

Then the objective function (see (2)) is converted to

$$(\psi^1, \dots, \psi^r)^* = \arg \max \frac{\sum_{i=1}^{C-1} \sum_{j=i+1}^C p_i p_j \left\| \mu_i^0 \times_1 \psi^1 \times_2 \psi^2 \dots \times_r \psi^r - \mu_j^0 \times_1 \psi^1 \times_2 \psi^2 \dots \times_r \psi^r \right\|^2 W_{ij}^c}{\sum_{i=1}^C p_i \sum_{j=1}^{N_i-1} \sum_{k=j+1}^{N_i} \left\| X_{j \times 1} \psi^1 \times_2 \psi^2 \dots \times_r \psi^r - X_{k \times 1} \psi^1 \times_2 \psi^2 \dots \times_r \psi^r \right\|^2 W_{jk}^s}. \quad (7)$$

2.2.2. The Design of Factor Separating Objective Function. The design of factor separating objective function is mainly about separating the content and style factors of objects, including the design of partitions and weight matrix.

There are two schemes for partition design recently; one is based on supervised algorithms, such as LDA, in which the samples are partitioned according to the class information of samples; the other is based on unsupervised algorithms, such as LPP, in which the samples are partitioned according to the k -nearest neighborhood of samples.

Variant weighting schemes have been proposed for different cases, which can be classified into four categories according to the weighting algorithms for scatter matrix between classes (content variance matrix F_c) in supervised face recognition methods based on weighted Linear Discriminant Analysis [17].

(1) *Euclidean Distance Algorithm* [17]. Consider

$$w(i, j) = \frac{1}{(\text{ed})^2}, \quad (8)$$

where ed is the Euclidean distance between the mean vectors of two classes.

(2) *Estimating Bayesian Error Rate Algorithm* [18]. Consider

$$w(i, j) = \frac{1}{2\Delta_{ij}^2} \text{erf}\left(\frac{\Delta_{ij}}{2\sqrt{2}}\right), \quad (9)$$

where Δ_{ij} is the Mahalanobis distance between the mean vectors of two classes and $\text{erf}(x) = (2/\sqrt{\pi}) \int_0^x e^{-t^2} dt$ is the error function.

(3) *Confusing Information Algorithm* [19]. Consider

$$w(i, j) = \begin{cases} \frac{E_{ij}}{n_i} & \text{if } i \neq j \\ 0 & \text{if } i = j, \end{cases} \quad (10)$$

where E_{ij} is the number of samples which are misclassified into class j but belong to class i actually and n_i is the sample number of the i th class.

(4) *Grey Correlation Analysis Algorithm Proposed in This Paper.* Consider

$$w(i, j) = \frac{1}{n} \sum_{k=1}^n \gamma(\mu_i(k), \mu_j(k)), \quad (11)$$

where n is the vector dimensionality and $\mu_i(k), \mu_j(k)$ are the k th feature in the mean vectors of the i th and j th partitions. $\gamma(\mu_i(k), \mu_j(k))$ is the correlation coefficients of the k th feature in the mean vectors of the i th and j th partitions, defined as follows:

$$\gamma(\mu_i(k), \mu_j(k)) = \frac{\min_j \min_k |\mu_i(k) - \mu_j(k)| + \rho \max_j \max_k |\mu_i(k) - \mu_j(k)|}{|\mu_i(k) - \mu_j(k)| + \rho \max_j \max_k |\mu_i(k) - \mu_j(k)|}. \quad (12)$$

The weighting scheme about the style variance matrix F_s gets little attention recently. According to the weighting scheme in LPP method, a new weighting scheme for F_s based on thermodynamic similarity is proposed in this paper:

$$W_{ij} = \begin{cases} \exp\left(-\frac{\|x_i - x_j\|^2}{2t}\right) & x_i, x_j \text{ belong to same partition} \\ 0 & \text{otherwise.} \end{cases} \quad (13)$$

It deserves deep research on how to distribute the weights effectively and reasonably. The optimal weighting scheme will never abandon any object that may be classified correctly and never waste any weight value. In other words, it will help all the objects which need help as much as possible in order to achieve a win-win situation.

2.3. Unifying Various Algorithms of Feature Extraction. The relationship between several classical subspace dimensionality reduction algorithms and factor analysis framework is discussed in this section. Actually, most subspace algorithms that are commonly used recently can be represented and explained by factor analysis framework. After unifying current subspace algorithms in framework, it is obvious that these algorithms are just different in the basis of factor construction and constraint of factor separation, though the motivation or objective of these algorithms is varying. Next we will introduce the relationships between factor analysis and some typical methods of dimension reduction based on subspace.

2.3.1. The Relationship with LDA. The objective of LDA is to find the optimal projection direction of classification that maximizes the ratio between the intraclass and interclass scatter matrix:

$$J(\psi) = \arg \max \frac{\text{trace}(\psi^T S_b \psi)}{\text{trace}(\psi^T S_w \psi)}, \quad (14)$$

$$S_b = \sum_{i=1}^C p_i (\mu_i^0 - \mu^0) (\mu_i^0 - \mu^0)^T, \quad (15)$$

$$S_w = \sum_{i=1}^C p_i \sum_{j=1}^{N_i} (x_j^0 - \mu_i^0) (x_j^0 - \mu_i^0)^T, \quad (16)$$

where C is the class number, p_i is the prior probability of class i , μ_i^0 is the mean vector of class i before projection, μ^0 is the overall mean vector before projection, and N_i is the sample number of class i . Refer to (3) and (4), set the weight matrix of F_c and F_s as all 1 matrix, and then

$$F_c = \sum_{i=1}^{C-1} \sum_{j=i+1}^C p_i p_j (\mu_i^0 - \mu_j^0) (\mu_i^0 - \mu_j^0)^T, \quad (17)$$

$$F_s = \sum_{i=1}^C p_i \sum_{j=1}^{N_i-1} \sum_{k=j+1}^{N_i} (x_j^{(i)} - x_k^{(i)}) (x_j^{(i)} - x_k^{(i)})^T.$$

After further conversion,

$$F_c = \sum_{i=1}^C p_i (\mu_i^0 - \mu^0) (\mu_i^0 - \mu^0)^T = S_b, \quad (18)$$

$$F_s = \sum_{i=1}^C p_i \sum_{j=1}^{N_i} (x_j^0 - \mu_i^0) (x_j^0 - \mu_i^0)^T = S_w.$$

Substitute (17) into (14); then

$$J(\psi) = \arg \max \frac{\text{trace}(\psi^T S_b \psi)}{\text{trace}(\psi^T S_w \psi)} = \frac{\text{trace}(\psi^T F_c \psi)}{\text{trace}(\psi^T F_s \psi)}. \quad (19)$$

Thus, LDA is a particular case of factor analysis framework when the entire weight matrix is 1 and the basis of partition is class information. The content factor is the accumulation of mean variance among each class and the style factor is the variance of each sample in the same class.

2.3.2. The Relationship with MFA [20] and DLA [21]. In order to overcome the hypothesis of LDA that the samples should be of Gaussian distribution, MFA redefines the scatter matrix within class and between classes of LDA, and the objective function is as follows:

$$J(\psi) = \arg \max \frac{\sum_c \sum_{(i,j) \in P_{k_2}(c) \text{ or } (j,i) \in P_{k_2}(c')} \|\psi^T x_i^c - \psi^T x_j^{c'}\|^2}{\sum_{c=1}^C \sum_{j \in N_{k_1}^c(i) \text{ or } i \in N_{k_1}^c(j)} \|\psi^T x_i^c - \psi^T x_j^c\|^2}, \quad (20)$$

where c is the class, $P_{k_2}(c)$ is a data pairs set that are the k_2 -nearest pairs between two different classes c and c' , and $N_{k_1}^c(i)$ indicates the index set of the k_1 -nearest neighbors of the sample i within the c th class.

It can be concluded that MFA is also a particular case of factor analysis framework; only it considers both the class information and nearest neighborhood of the samples for the partition design. The numerator of (20) can be rewritten as follows:

$$\sum_c \sum_{(i,j) \in P_{k_2}(c) \text{ or } (j,i) \in P_{k_2}(c')} \|\psi^T x_i^c - \psi^T x_j^{c'}\|^2 = \sum_{i=1}^{C-1} \sum_{j=1}^C \|\psi^T \mu_i^0 - \psi^T \mu_j^0\|^2 = \psi^T F_c \psi, \quad (21)$$

where μ_i^0, μ_j^0 are the mean vector of the k_2 -nearest pairs between the i th and j th classes. Therefore, the interclass separability can be seen as the content factor in factor analysis framework. Though the class information and nearest neighborhood are considered in the design of interclass separability, the weight matrix is ignored.

Similarly, the denominator of (20) can be rewritten as follows:

$$\sum_{c=1}^C \sum_{j \in N_{k_1}^c(i) \text{ or } i \in N_{k_1}^c(j)} \|\psi^T x_i^c - \psi^T x_j^c\|^2 = \sum_{c=1}^C \sum_{i=1}^{N_{k_1}^c(i)} \sum_{j=1}^{N_{k_1}^c(j)} \|\psi^T x_i^c - \psi^T x_j^c\|^2 = \psi^T F_s \psi. \quad (22)$$

Therefore, the intraclass compactness of MFA can be seen as the style factor in factor analysis framework. Interclass separability is conducted based on the k -nearest neighbor relationships of the samples from the same class, disregarding the design of weight matrix.

For DLA, the design criterion of part optimization is that the distance of the nearest neighbor image is minimized within class and maximized among classes. The objective function of DLA is equal to (20), and thus DLA is a particular case of factor analysis framework also.

2.3.3. The Relationship with LPP. Refer to (3); $J(\psi) = \arg \min_{\psi^T F_c \psi = d} \text{trace}(\psi^T F_s \psi)$. Assume that all the prior probability p_i is identical, and then the content and style variance matrix can be represented as follows:

$$\begin{aligned} F_c &= \sum_{i=1}^{C-1} \sum_{j=i+1}^C W_{ij}^c (\mu_i^0 - \mu_j^0) (\mu_i^0 - \mu_j^0)^T \\ &= \sum_{i=1}^C W_{ij}^c (\mu_i^0 - \mu^0) (\mu_i^0 - \mu^0)^T, \end{aligned} \quad (23)$$

$$F_s = \sum_{i=1}^C \sum_{j=1}^{N_i-1} \sum_{k=j+1}^{N_i} W_{jk}^s (x_j^{(i)} - x_k^{(i)}) (x_j^{(i)} - x_k^{(i)})^T,$$

where F_s can be represented as follows:

$$F_s = \sum_{i=1}^C \sum_{j=1}^{N_i-1} \sum_{k=j+1}^{N_i} W_{jk}^s \|x_j^{(i)} - x_k^{(i)}\|^2. \quad (24)$$

Partition the samples according to k -nearest neighbors; then (24) can be written as

$$F_s = \sum_i \sum_j \|x_i - x_j\|^2 w_{ij}, \quad (25)$$

where $W \in \mathfrak{R}^{N \times N}$ is the relation matrix weighted by the heat kernels: $w(i, j) = \exp(-\|x_i - x_j\|^2/t)$ if x_i is one of the k -nearest neighbors of x_j or x_j is one of the k -nearest neighbors of x_i ; otherwise it is 0, and t is a tuning parameter.

Thus (3) can be represented as follows:

$$\begin{aligned} J(\psi) &= \arg \min_{\psi^T F_c \psi = d} \text{trace}(\psi^T F_s \psi) \\ &= \arg \min_{\psi^T F_c \psi = d} \psi^T X L X^T \psi, \end{aligned} \quad (26)$$

where D is diagonal matrix, whose elements are the sum of the column or row elements in symmetric matrix S , represented as $D_{ij} = \sum_j S_{ij}$, and L is Laplace operator; $L = D - S$. It can be seen that LPP is also a particular case of factor analysis framework. The style factor is the weighted accumulation of sample variance in the same neighborhood.

2.3.4. The Relationship with PCA. PCA is a commonly used algorithm for dimensionality reduction in pattern recognition, the objective of which is to find the best representative mapping space of the original data in the minimum mean square sense. The objective function to the dataset $X = \{x_i\}_{i=1}^n \in \mathfrak{R}^m$ is

$$J(\psi) = \arg \min_{\psi^T \psi = 1} \psi^T \Sigma \psi, \quad (27)$$

where Σ is the covariance matrix:

$$\begin{aligned} \Sigma &= \sum_{i=1}^n (x_i - \mu) (x_i - \mu)^T \\ &= \sum_{i=1}^n \left(x_i - \frac{1}{n} \sum_{j=1}^n x_j \right) \left(x_i - \frac{1}{n} \sum_{k=1}^n x_k \right)^T \\ &= \frac{1}{n^2} \sum_{i=1}^n \left[\sum_{j=1}^n \sum_{k=1}^n (x_i - x_j) (x_i - x_k)^T \right], \end{aligned} \quad (28)$$

where μ is the mean value of all the samples; reformulate Σ in different index, and then

$$\Sigma = \frac{1}{n^2} \sum_{j=1}^n \left[\sum_{i=1}^n \sum_{k=1}^n (x_j - x_i) (x_j - x_k)^T \right]. \quad (29)$$

Change the summation order in (29); then

$$\Sigma = \frac{1}{n^2} \sum_{i=1}^n \left[\sum_{j=1}^n \sum_{k=1}^n (x_j - x_i) (x_j - x_k)^T \right]. \quad (30)$$

Add (28) to (30):

$$\begin{aligned} \Sigma &= \frac{1}{2} \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n (x_i - x_j) (x_i - x_j)^T \\ &= \frac{1}{2n} \sum_{i=1}^n \sum_{j=1}^n (x_i - x_j) (x_i - x_j)^T. \end{aligned} \quad (31)$$

Set W_{ij}^s to be an $n \times n$ matrix in which each element is equal to $1/2n$; then style variance matrix F_s can be represented as follows:

$$\begin{aligned} F_s &= \sum_{i=1}^C \sum_{j=1}^{N_i-1} \sum_{k=j+1}^{N_i} W_{jk}^s \|x_j^{(i)} - x_k^{(i)}\|^2 \\ &= \sum_{i=1}^C \sum_{j=1}^{N_i-1} \sum_{k=j+1}^{N_i} \frac{1}{2n} \|x_j^{(i)} - x_k^{(i)}\|^2. \end{aligned} \quad (32)$$

Treat each sample as an independent partition; then

$$F_s = \frac{1}{2n} \sum_{i=1}^n \sum_{j=1}^n (x_i - x_j) (x_i - x_j)^T = \Sigma. \quad (33)$$

Thus (3) can be represented as follows:

$$J(\psi) = \arg \min_{\psi^T F_c \psi = d} \text{trace}(\psi^T F_s \psi) = \arg \min_{\psi^T \psi = 1} \psi^T \Sigma \psi. \quad (34)$$

It can be seen that PCA is also a particular case of factor analysis framework. The style factor of PCA is the accumulation of all the sample variance.

2.4. Related Works and Discussions. Feature extraction methods based on subspace have got a lot of attention. Unifying all the current subspace analysis methods is convenient for comprehending the essence of these methods and comparing the variance of different methods. Various frameworks have been proposed in order to deeply mine the commonality and individuality of current feature extraction methods and provide a novel platform for proposing new methods. In 2007, Yan et al. have proposed the Graph Embedding (GE) framework [20], which describes most current subspace methods and presents that the variance of linear subspace methods is expressed in the graph structure difference, and the variance between linear methods and kernel methods is expressed in different way to embed. In 2009, [21] proposed a Patch Alignment (PA) framework for the motivation of dimensionality reduction. In the same year a universal learning framework of supervised subspace [22] has been proposed, which is based on the fundamental concept of discriminant analysis and evaluates the classification capacity of the features through the ratio between the intraclass compactness and interclass separability after projecting. It proves that LDA, LPP, and NPE algorithms are all particular cases of that framework, and the essence of these algorithms, no matter whether supervised or unsupervised, is to minimize the intraclass compactness while maximizing the interclass separability, only different in the process of sample classes.

We propose the framework based on factor analysis through deep research on subspace methods and current unified framework. Factor analysis framework is superior to other frameworks in the following aspects:

- (1) The classification meaning is clearer and the design is more flexible. The content and style factors in factor analysis are various depending on the different applications, so kinds of design schemes of the two factors can be proposed by researchers. The intraclass compactness and interclass separability in the universal learning framework of supervised subspace proposed in [22] are exactly the content and style factors in factor analysis framework. The part optimization design in [21] is the process of factor separation. Meanwhile, factor analysis criterion provides flexible weights allocation plan to research.
- (2) Factor analysis framework is adapted to feature extraction and dimensionality reduction methods which are commonly used currently, including linear, nonlinear, and tensor decomposition methods. Supervised and unsupervised feature extraction methods are also included.
- (3) The factor analysis framework is designed for classification, so superior feature extraction methods can be designed through this platform. The FA-LDA algorithm presented in Section 3.1 and FA-LPP algorithm presented in Section 3.2 are both improved feature

extraction methods proposed using factor analysis framework.

3. FA-LDA and FA-LPP

3.1. FA-LDA: Weight Design Based on Factor Analysis. In addition to encompassing most popular dimensionality reduction algorithms, the proposed framework can also be used as a general platform to design new algorithms for dimensionality reduction. With this framework, we develop a new dimensionality reduction algorithm to avoid certain limitations of the traditional Linear Discriminant Analysis in terms of suboptimal problem.

Researchers for LDA are mainly focused on the Fisher criterion solution and small sample problems but little on the Fisher criterion itself. Actually, Fisher criterion is suboptimal for the classification problem with K ($K > 2$) classes. Marco Loog is one of the earliest researchers that pointed out this problem and proposed an improved weighting algorithm based on Bayesian error rate function [18]. In the meanwhile, Li et al. proposed an improved weighting algorithm based on Euclidean distance at the same year [17]. After that, fractional-step LDA is proposed in [23] to solve the suboptimal problem of Fisher criterion. Fractional-step LDA is effective but time consuming since it is an iterative algorithm. In this section, we improve the Fisher criterion based on factor analysis in order to solve its suboptimal problem.

Actually, the content factor F_c and style factor F_s in factor analysis criterion are identical to interclass and intraclass scatter matrices S_b , S_w in the case that every weight equals 1. When there is a pair of classes far from each other, the covariance $(\mu_i - \mu_j)(\mu_i - \mu_j)^T$ will be large and totally determine S_b , leading to the fact that the final projecting matrix excessively emphasizes the classes which are already classified well but make less use of the pair of classes (i' , j') near each other. However, the pairs of classes near each other deserve more attention.

According to the above discussion, the weight should be reduced for the class pairs far away from each other while it should be increased for the class pairs near each other. F_c is obtained from weighting S_b :

$$F_c = \sum_{i=1}^{C-1} \sum_{j=i+1}^C p_i p_j w(i, j) (\mu_i - \mu_j)(\mu_i - \mu_j)^T, \quad (35)$$

where

$$w(i, j) = \frac{1}{n} \sum_{k=1}^n \gamma(\mu_i(k), \mu_j(k)), \quad (36)$$

where n is the vector dimensionality, $\mu_i(k)$, $\mu_j(k)$ are the k th feature in the mean vector of the i th and j th class, respectively, and $\gamma(\mu_i(k), \mu_j(k))$ is the correlation coefficient of the k th feature in the mean vector of the i th and j th class, defined as follows:

$$\gamma(\mu_i(k), \mu_j(k)) = \frac{\min_j \min_k |\mu_i(k) - \mu_j(k)| + \rho \max_j \max_k |\mu_i(k) - \mu_j(k)|}{|\mu_i(k) - \mu_j(k)| + \rho \max_j \max_k |\mu_i(k) - \mu_j(k)|}. \quad (37)$$

Similarly, the objective of the factor analysis criterion is to remove the style factor as much as possible. Thus, the contribution to solve the eigenvalue from the sample pair near each other in the same class should be reduced, since it is possible that this pair of samples has little interference from style factor; the contribution to solve the eigenvalue from the sample pair far away from each other in the same class should be increased, since it is certain that one sample of this pair contains style factor, and the optimization function of the criterion should emphasize this kind of pairs. F_s is obtained from weighting S_w :

$$F_s = \sum_{i=1}^C p_i \sum_{j=1}^{N_i-1} \sum_{k=j+1}^{N_i} w(j, k) (x_j^{(i)} - x_k^{(i)}) (x_j^{(i)} - x_k^{(i)})^T, \quad (38)$$

where the weighting function $w(j, k)$ is defined as follows:

$$w(j, k) = \exp\left(-\frac{\|x_j - x_k\|^2}{2t^2}\right), \quad (39)$$

where t is an empirical constant. The distance within class is more, the weight in S_w^{-1} is heavier, and vice versa.

$$w_{jk}^s = \begin{cases} \exp\left(-\frac{\|x_j - x_k\|^2}{2t}\right) & x_j \text{ and } x_k \text{ belong to the same class and } x_j \text{ is one of the } k\text{-nearest neighbors of } x_k \\ 0 & \text{other.} \end{cases} \quad (42)$$

The content factor of FA-LPP is

$$F_c = \sum_{i=1}^{C-1} \sum_{j=i+1}^C w(i, j) (\mu_i - \mu_j) (\mu_i - \mu_j)^T. \quad (43)$$

μ_i, μ_j are the mean values of the k -nearest neighbors of the i th and j th classes, respectively, and $w(i, j)$ is computed by (36). Given the definition of F_c, F_s , FA-LPP can be defined as follows:

$$J_F(W) = \arg \max_W \frac{\text{trace}(W^T F_c W)}{\text{trace}(W^T F_s W)}. \quad (44)$$

4. Experiments

4.1. FA-LDA. To evaluate the proposed FA-LDA algorithm, we compare it with the LDA algorithm on artificial data and real-world face dataset. During the testing phases, the nearest neighbor (NN) rule was used in classification. Note that [24] proposed a probabilistic graphical model framework to mimic the process of generating a face image. However, they focus on addressing the obstacles of small sample set, occlusion, and illumination variations in one sample face identification. So we do not compare our algorithm with [24].

Experiments on Artificial Data. The artificial data includes 3 subsets with 3, 4, and 5 classes, respectively, and the prior

Given the weighted F_c, F_s , FA-LDA is defined as follows:

$$J_F(W) = \arg \max_W \frac{\text{trace}(W^T F_c W)}{\text{trace}(W^T F_s W)}. \quad (40)$$

3.2. FA-LPP. The style variance in the same locality manifold is minimized after projecting in LPP algorithm; namely, the locality manifold is preserved in low-dimensional space after projecting. We improve the LPP algorithm based on factor analysis criterion so that the style variance in locality manifold of the same class is minimized and the content variance in locality manifold of different classes is maximized after projecting. Thus, the style factor of FA-LPP is

$$F_s = \sum_{i=1}^C \sum_{j=1}^{N_i-1} \sum_{k=j+1}^{N_i} w_{jk}^s \|x_j^{(i)} - x_k^{(i)}\|^2, \quad (41)$$

where weight matrix w_{jk}^s is

probability of each class is identical. Each class contains 60 samples with 9 dimensions. There is one class easy to discriminate in each subset. Thus, its within class covariance matrix plays a dominant role compared to other classes. Meanwhile, assume that the other classes obey the normal distribution and have identical within class covariance matrix. In experiments, 10 samples are selected as train set and the other 50 samples are equally divided into 10 groups as test set. The experiments of each class are conducted 10 times and each experiment is repeated 10 times. The classification rate on average and variance of the two algorithms are listed in Table 1. It can be seen that the classification performance of LDA algorithm is not good on this artificial data, which indicates the limitation of LDA. FA-LDA is superior to LDA in classification property. Since LDA algorithm simply uses the average of the within class scatter matrix as unified covariance matrix, it leads to an overemphasis on the scatter matrix of the class easy to discriminate. However, for FA-LDA algorithm the weights of this easily separated class are small and more attention is on the confusable classes.

Experiments on Real-World Data. The ORL database contains 400 images of 40 individuals which is used for real-world experiment. The images were captured at different times and with different variations including expression and facial details, which can be seen from Figure 2. In order to evaluate the superiority of FA-LDA, we compare it with the LDA and

TABLE 1: Classification accuracies of LDA and FA-LDA on the artificial database.

3 classes		4 classes		5 classes	
FA-LDA	LDA	FA-LDA	LDA	FA-LDA	LDA
83.5 ± 7.45	62 ± 8.34	75.5 ± 5.6	38.54 ± 10.21	78.26 ± 5.4	45.26 ± 7.85
81 ± 9.01	56.7 ± 7.45	73 ± 6.68	45.85 ± 7.37	75.66 ± 6.28	46.15 ± 8.21
83.5 ± 9.07	59.5 ± 10.53	73.5 ± 5.62	46.33 ± 5.36	75.5 ± 7.25	39.54 ± 9.25
80.5 ± 8.67	57 ± 9.57	76.67 ± 5.6	42.64 ± 7.6	79.45 ± 5.5	48.26 ± 6.51
81 ± 9.01	53.5 ± 12.45	78.56 ± 5.21	43.85 ± 7.49	76.5 ± 7.50	44.58 ± 9.21
81.5 ± 6.26	61.5 ± 9.57	78.85 ± 5.34	40.28 ± 6.35	70.57 ± 9.14	39.83 ± 5.68
84.5 ± 6.58	60 ± 6.7	80.21 ± 8.63	42.84 ± 5.49	75.64 ± 5.60	46.25 ± 7.35
82.5 ± 8.08	59.5 ± 11.56	78.26 ± 5.23	41.45 ± 8.16	76.42 ± 5.60	42.35 ± 5.36
81 ± 8.45	64.5 ± 6.87	75.95 ± 6.72	43.58 ± 6.53	77.38 ± 6.58	46.25 ± 8.24

FIGURE 2: The sample facial images of ORL database.

FIGURE 3: Best recognition rates of four algorithms on the testing sets of ORL.

current weighted algorithms, in which WLDA1 [17] is based on Euclidean distance and WLDA2 [19] is based on confusion matrix. For training, we randomly selected different numbers (three, four, five, six, seven, eight, and nine) of images per individual and used the rest of the images for testing. Such a trial was independently performed 10 times, and then the average recognition results were calculated. The results are shown in Figure 3. The classification property of LDA is the worst, which illustrated the validity the weighting scheme based on factor analysis criterion. FA-LDA algorithm achieves the best performance for it not only weights S_b to

reduce the influence of the far away class, but also weights S_w to improve the anti-interference ability of discriminant criterion.

4.2. FA-LPP. To evaluate the proposed FA-LPP algorithm, we compare it with LPP and MFA/DLA on ORL and Yale database. The Yale database is built by computational vision and control center at Yale University. Everyone has 11 images varying in expressions (blinking, glad, sad, or surprised), illumination conditions (frontal, left, or right), and facial details (glasses or no glasses), which can be seen from Figure 4.

For the ORL and Yale database, the image set is partitioned into the different gallery and probe sets where G_m/P_n indicates that m images per person are randomly selected for training and the remaining n images are used for testing. For the LPP, the important parameters include k (the number of Neighbor Measurements). For the MFA and DLA algorithm, the important parameters include k_1 (the number of Neighbor Measurements of the Same Class), k_2 (the number of Neighbor Measurements of Different Classes). For the proposed FA-LPP, the important parameters include k (the number of Neighbor Measurements of the Same Class).

Each experiment is repeated 10 times and the average classification rate on ORL and Yale database is listed in Tables 2 and 3, which show the superiority of our algorithm. The LPP algorithm only utilizes the nearest neighborhood for partition, while MFA, DLA, and FA-LPP algorithms consider not only the nearest neighborhood but also the class information. Thus, the classification accuracies of MFA, DLA, and FA-LPP algorithms are higher than LPP algorithm. Since the weighting scheme is also taken into account for FA-LPP, the classification accuracy of FA-LPP is higher than MFA and

FIGURE 4: The sample facial images of Yale database.

FIGURE 5: Sample images of FG-Net aging database (the same person with face images of different ages).

TABLE 2: Best recognition rates of four algorithms on the testing sets of ORL.

	G_3/P_7	G_4/P_6	G_5/P_5
LPP	85.0% (60)	89.6% (141)	96.0% (158)
MFA	89.3% (39, 2, 2)	91.4% (46, 2, 1)	96.0% (42, 4, 2)
DLA	94% (42, 2, 2)	92.5% (53, 2, 1)	95.4% (40, 4, 2)
FA-LPP	94.5% (39, 2)	92% (45, 2)	96.6% (40, 4)

For LPP, the numbers in the parentheses are the selected subspace dimensions. For FA-LPP, the first numbers in the parentheses are the selected subspace dimensions and the second numbers are the parameters. For MFA and DLA, the first numbers in the parentheses are the selected subspace dimensions and the second and the third numbers are the parameters k_1 and k_2 , respectively.

TABLE 3: Best recognition rates of four algorithms on the testing sets of Yale.

	G_3/P_8	G_5/P_6	G_7/P_4
LPP	63.5% (13)	72.3% (14)	80.5% (14)
MFA	61.0% (13, 2, 2)	74.5% (14, 2, 1)	79.7% (21, 4, 2)
DLA	66.8% (15, 2, 2)	78.9% (23, 2, 1)	81.2% (13, 4, 2)
FA-LPP	73.2% (14, 2)	79.5% (16, 2)	85.4% (15, 4)

For LPP, the numbers in the parentheses are the selected subspace dimensions. For FA-LPP, the first numbers in the parentheses are the selected subspace dimensions and the second numbers are the parameters. For MFA and DLA, the first numbers in the parentheses are the selected subspace dimensions and the second and the third numbers are the parameters k_1 and k_2 , respectively.

DLA. What is more, there is only one parameter of nearest neighborhood to design for FA-LPP algorithm.

4.3. Comprehensive Experiments. In order to comprehensively analyze the performance of FA-LDA and FA-LPP, we conducted facial age group classification experiment and compare it with LDA, WLDA1, WLDA2, LPP, and DLA. In this experiment, the images for age group classification are collected from multiple data sources like 1002 facial images from FG-NET database, 500 images from Google database, and 600 images from the scanned photographs leading to a total of 2102 sample facial images. A few of them are shown in Figure 5. For experimental purposes the total 2102 images are split into two training and test sets. Training set consists of 700 images from FG-Net aging database, 300 images

FIGURE 6: Comparisons of the performance of different approaches on facial age database.

from Google, and 400 images collected from scanned images leading to a total of 1400 images. Testing set consists of 302 images from FG-Net database, 200 images from Google, and 200 images from scanned photographs, totaling 702 images.

For the four-class classification, faces in the training dataset will be assigned to age groups according to their labeled age. As every face has an age value labeled, it can provide necessary data for building the factors analysis framework.

In this experiment, the Gabor feature was used first in original facial image, and then we used the FA-LDA, FA-LPP, LDA, WLDA1, WLDA2, LPP, and DLA algorithms in feature extraction and dimensionality reduction. The comparative results in Figure 6 show that our methods (FA-LDA, FA-LPP) outperformed others. From this result, the classification accuracies of DLA and our FA-LDA and FA-LPP are higher than LDA, WLDA1, WLDA2, and LPP algorithms. The reason is that LDA and LPP are used for linear dimensional reduction without considering weight design problems, respectively, and WLDA1 and WLDA2 are not considered partition

design. Furthermore, since LPP is based on manifold learning which is suitable for nonlinear problem such as face age group classification, therefore, the LPP and FA-LPP are better than LDA and FA-LDA, respectively.

Recently, deep learning is a new area of machine learning research which has got hot attention [25, 26]. Our factors analysis framework can be embedded in deep learning based object recognition system as feature dimension reduction part. Therefore, our factors analysis framework can be seamlessly integrated with the deep learning method.

5. Conclusions

In this paper, we aim to provide insights into the relationship among the state-of-the-art dimensionality reduction algorithms as well as to facilitate the design of new algorithms. A general framework known as factor analysis, along with its linearization, kernelization, and tensorization, has been proposed to provide a unified perspective for the understanding and comparison of many popular dimensionality reduction algorithms. Moreover, the factor analysis framework can be used as a general platform to develop new algorithms for dimensionality reduction. As shown in this paper, we have proposed two novel dimensionality reduction algorithms called FA-LDA and FA-LPP by designing the objective function of factors separation that characterize the weight matrix and the partition and by optimizing their corresponding criteria based on the factor analysis framework. These new algorithms are shown to effectively overcome the data distribution assumption of the traditional LDA and LPP algorithm. Thus, FA-LDA and FA-LPP are more general algorithms for discriminant analysis.

A byproduct of this paper is a series of kernelization and tensorization versions of the factors analysis. One of our future works is to systematically compare all possible extensions of the algorithms mentioned in this paper.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This work is supported partially by China Postdoctoral Science Foundation (no. 2015M582355), the Doctor Scientific Research Start Project from Hubei University of Science and Technology (no. BK1418), the National Natural Science Foundation of China (NSFC) (no. 61271256), Provincial Natural Science Foundation of Hubei Province (no. 2015cfc778), and National Natural Science Foundation of China (NSFC) (no. 61373108).

References

- [1] Q. Zhang, L. Zhang, Y. Yang, Y. Tian, and L. Weng, "Local patch discriminative metric learning for hyperspectral image feature extraction," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 3, pp. 612–616, 2014.
- [2] Y. Fu, S. Yan, and T. S. Huang, "Classification and feature extraction by simplexization," *IEEE Transactions on Information Forensics & Security*, vol. 3, no. 1, pp. 91–100, 2008.
- [3] A. M. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228–233, 2001.
- [4] J. Ye, R. Janardan, C. H. Park, and H. Park, "An optimization criterion for generalized discriminant analysis on undersampled problems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 8, pp. 982–994, 2004.
- [5] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using Laplacianfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328–340, 2005.
- [6] Y. Su, Y. Fu, X. Gao, and Q. Tian, "Discriminant learning through multiple principal angles for visual recognition," *IEEE Transactions on Image Processing*, vol. 21, no. 3, pp. 1381–1390, 2012.
- [7] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.
- [8] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
- [9] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in Neural Information Processing Systems 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds., pp. 585–591, 2001.
- [10] K. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Schölkopf, "An introduction to kernel-based learning algorithms," *IEEE Transactions on Neural Networks*, vol. 12, no. 2, pp. 181–201, 2001.
- [11] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H.-J. Zhang, "Discriminant analysis with tensor representation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, pp. 526–532, IEEE, June 2005.
- [12] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 1, pp. 131–137, 2004.
- [13] J. Ye, "Generalized low rank approximations of matrices," in *Proceedings of the 21st International Conference on Machine Learning (ICML '04)*, p. 112, Banff, Canada, July 2004.
- [14] J. Ye, R. Janardan, and Q. Li, "Two-dimensional linear discriminant analysis," *Advances in Neural Information Processing Systems*, vol. 17, pp. 1569–1576, 2005.
- [15] V. Popa, J. Nurminen, and M. Gabbouj, "A study of bilinear models in voice conversion," *Journal of Signal and Information Processing*, vol. 2, no. 2, pp. 125–139, 2011.
- [16] H. Tan, B. Cheng, J. Feng, G. Feng, and Y. Zhang, "Tensor recovery via multi-linear augmented lagrange multiplier method," in *Proceedings of the 6th International Conference on Image and Graphics (ICIG '11)*, pp. 141–146, Hefei, China, August 2011.
- [17] Y. Li, Y. Gao, and H. Erdogan, "Weighted pairwise scatter to improve linear discriminant analysis," in *Proceedings of the 6th International Conference on Spoken Language Processing*, vol. 4, pp. 608–611, Beijing, China, October 2000.
- [18] M. Loog, R. P. W. Duin, and R. Haeb-Umbach, "Multiclass linear dimension reduction by weighted pairwise Fisher criteria," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 7, pp. 762–766, 2001.

- [19] H.-S. Lee and B. Chen, "Linear discriminant feature extraction using weighted classification confusion information," in *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH '08)*, pp. 2254–2257, Brisbane, Australia, September 2008.
- [20] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: a general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40–51, 2007.
- [21] T. Zhang, D. Tao, X. Li, and J. Yang, "Patch alignment for dimensionality reduction," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1299–1313, 2009.
- [22] J. Yang, S. Yan, and T. S. Huang, "Ubiquitously supervised subspace learning," *IEEE Transactions on Image Processing*, vol. 18, no. 2, pp. 241–249, 2009.
- [23] R. Lotlikar and R. Kothari, "Fractional-step dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 6, pp. 623–627, 2000.
- [24] X. Zhao, X. Li, Z. Wu, Y. Fu, and Y. Liu, "Multiple subcategories parts-based representation for one sample face identification," *IEEE Transactions on Information Forensics & Security*, vol. 8, no. 10, pp. 1654–1664, 2013.
- [25] Y. Jia, E. Shelhamer, J. Donahue et al., "Caffe: convolutional architecture for fast feature embedding," <http://arxiv.org/abs/1408.5093>.
- [26] S. Yang, P. Luo, C. Loy, K. W. Shum, and X. Tang, "Deep representation learning with target coding," in *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI '15)*, Austin Tex, USA, January 2015.