Cost-Sensitive Semi-Supervised Discriminant Analysis for Face Recognition

Jiwen Lu, *Member, IEEE*, Xiuzhuang Zhou, *Member, IEEE*, Yap-Peng Tan, *Senior Member, IEEE*, Yuanyuan Shang, *Member, IEEE*, and Jie Zhou, *Senior Member, IEEE*

***Abstract*—This paper presents a cost-sensitive semi-supervised discriminant analysis method for face recognition. While a number of semi-supervised dimensionality reduction algorithms have been proposed in the literature and successfully applied to face recognition in recent years, most of them aim to seek low-dimensional feature representations to achieve low classification errors and assume the same loss from all misclassifications in the feature representation/extraction phase. In many real-world face recognition applications, however, this assumption may not hold as different misclassifications could lead to different losses. For example, it may cause inconvenience to a gallery person who is misrecognized as an impostor and not allowed to enter the room by a face recognition-based door locker, but it could result in a serious loss or damage if an impostor is misrecognized as a gallery person and allowed to enter the room. Motivated by this concern, we propose in this paper a new method to learn a discriminative feature subspace by making use of both labeled and unlabeled samples and exploring different cost information of all the training samples simultaneously. Experimental results are presented to demonstrate the efficacy of the proposed method.**

***Index Terms*—Cost sensitive, discriminant analysis, face recognition, semi-supervised.**

I. Introduction

In face recognition applications, one is often confronted with high-dimensional data, and it is necessary to transform these high-dimensional face data into a low-dimensional feature space because high-dimensional data usually deteriorate the performances of classifiers and lead to high computational complexity. Over the past two decades, a large number of dimensionality reduction algorithms have been proposed and successfully applied to face recognition [1], [2], [6], [9]–[11], [16], [18], [27], [29], [35].

Linear discriminant analysis (LDA) [1] is a popular discriminative dimensionality reduction method and has been widely used in face recognition [1], [16]. The aim of LDA is to seek a set of discriminative projections to maximize the ratio of between-class variance to within-class variance. In many real world applications, it is usually difficult, expensive and time consuming to collect sufficient labeled data because laborious human labeling effort is required, which makes the estimated projection vectors of LDA inaccurate and correspondingly degrades the final recognition performance. To address this problem, some semi-supervised learning algorithms which utilize both labeled and unlabeled data to improve the face recognition performance have been proposed in recent years [4], [40], among which the graph-based approach is one of the most active and effective areas [40]. The basic idea of graph-based semi-supervised learning is to model the whole data (both labeled and unlabeled) as a graph and then propagate label information from the labeled data to the unlabeled data through the graph constructed by both labeled and unlabeled samples.

More recently, semi-supervised learning techniques have also been incorporated into dimensionality reduction, and many such algorithms have been proposed [3], [21], [25], [26], [30], [34], [36], [37]. For example, Cai *et al.* [3] proposed a semi-supervised discriminant analysis (SDA) method to extract discriminative features and preserve geometrical information of both the labeled and unlabeled samples for dimensionality reduction; Zhang and Yeung [37] applied a robust path-based similarity measure to better exploit the graph of SDA, Xu and Yan [34] presented an adaptive regularization method to better characterize the interplay between the labeled and unlabeled data in SDA, and Wang *et al.* [30] introduced a semi-parametric approach to better estimate the projection directions of SDA, to improve the discriminative power for semi-supervised dimensionality reduction methods. Moreover, Nie *et al.* [21] imposed an orthogonal constrain and formulated a semi-supervised orthogonal discriminant analysis (SODA) method for dimensionality reduction, Song *et al.* [25] proposed a semi-supervised submanifold discriminant analysis algorithm to discover the nonlinear relationships of samples in semi-supervised dimensionality reduction, and Wang *et al.* [32] presented a semi-supervised dimensionality reduction with pairwise constraints (SDAPC) method which utilizes pairwise constraints of samples to perform semi-supervised dimensionality reduction. Despite different assumptions of these methods, most of them can

Manuscript received November 02, 2011; revised January 18, 2012; accepted January 26, 2012. Date of current version May 08, 2012. This work was supported in part by the National Natural Science Foundation of China under Grant 11178017 and the Beijing Nova Program under Grant 2008B57. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Fabio Scotti.

J. Lu is with the Advanced Digital Sciences Center, Singapore 138632, Singapore (e-mail: jiwen.lu@adsc.com.sg).

X. Zhou and Y. Shang are with the Beijing Engineering Research Center of High Reliable Embedded System, and the College of Information Engineering, Capital Normal University, Beijing 100048, China (e-mail: zxz@xeehoo.com; syy@bao.ac.cn).

Y.-P. Tan is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore (e-mail: eyptan@ntu.edu.sg).

J. Zhou is with the Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: jzhou@tsinghua.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2012.2188389

be unified into a general semi-supervised dimensionality reduction framework with different constraints [26].

All existing semi-supervised dimensionality reduction methods only seek low recognition errors and implicitly assume that the losses of all misclassifications are the same in the feature extraction phase for a practical recognition system. However, this assumption may not be suitable for many real-world face recognition tasks because different misclassifications could lead to different losses. For example, in a face recognition-based door locker system, it may cause inconvenience to a gallery person who is misrecognized as an impostor and not allowed to enter the room, but may result in a serious loss or damage if an impostor is misrecognized as a gallery person and allowed to enter the room. Another example is the face sketch recognition-based criminal search system. It could cause some trouble to an innocent citizen who is misrecognized as a criminal suspect by the criminal search system, but may result in a large insecurity if an criminal suspect is misrecognized as innocent citizen and allowed to go free. From these two examples, we conclude that face recognition is a cost-sensitive pattern classification problem.

Cost-sensitive learning is one important topic in the data mining and machine learning community [13], [20], [22], [28], [38]. In such settings, cost information is introduced to measure the importance of different samples in different classes, and different costs reflect different amounts of losses. The aim of cost-sensitive learning is to minimize the total cost rather than the total error. Generally, there are two kinds of misclassification cost. The first is class-dependent, where the costs of misclassifying any example in class A to class B are the same. The second is example-dependent, where the costs of classifying examples in class A to class B are different. In this paper, we focus on the former one because face recognition is generally a class-dependent cost-sensitive problem.

There have been several cost-sensitive learning algorithms proposed in the literature, such as cost-sensitive boosting [20], [28], cost-sensitive support vector machine [13], [38], cost-sensitive semi-supervised learning [13], and cost-sensitive neural networks [22]. While these cost-sensitive learning methods can attain better performance than conventional cost-blind methods, most of them are designed for classification [13], [20], [22], [28], [38] rather than dimensionality reduction. While improved recognition performance (smaller recognition loss) has been obtained, these methods only use cost-sensitive techniques in the classification phase and not in the feature extraction (dimensionality reduction) phase. In other words, the samples with different costs have not been utilized in the feature extraction phase, and hence some useful cost-sensitive information could be lost in this phase.

In this paper, we propose a cost-sensitive semi-supervised discriminant analysis $(\mathrm{CS^3DA})$ method to learn a discriminative feature space by making use of both labeled and unlabeled samples in a cost-sensitive setting. Specifically, we first reconstruct each unlabeled sample from the labeled data and use the reconstruction coefficients to infer a soft label for each unlabeled sample. Then, we learn a low-dimensional feature subspace by utilizing both label and unlabeled data, such that the low overall loss is achieved when recognition is performed in the feature space derived. Experimental results on cost-sensitive face recognition are presented to demonstrate the efficacy of the proposed method.

The remainder of the paper is organized as follows. Section II details our proposed cost-sensitive semi-supervised discriminant analysis method. The experimental results and discussions are presented in Section III. We conclude the paper in Section IV.

TABLE I

COST MATRIX OF GENERIC PATTERN CLASSIFICATION SYSTEM. HERE C_{ij} DENOTES COST OF MISCLASSIFYING SAMPLES OF iTH CLASS AS THE jTH CLASS. DIAGONAL ELEMENTS IN COST MATRIX ARE ZERO BECAUSE THERE IS NO LOSS FOR CORRECT CLASSIFICATION

	l_1	$\cdots$	l_j	$\cdots$	l_c
l_1	0	$\cdots$	C_{1j}	$\cdots$	C_{1c}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
l_i	C_{i1}	$\cdots$	C_{ij}	$\cdots$	C_{i1}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
l_c	C_{c1}	$\cdots$	C_{cj}	$\cdots$	0

II. COST-SENSITIVE SEMI-SUPERVISED DISCRIMINANT ANALYSIS $(\mathrm{CS^3DA})$

Different from existing semi-supervised dimensionality reduction methods which implicitly assume that the losses of all misclassifications are the same in the feature extraction (dimensionality reduction) phase of a recognition system, we explicitly incorporate the cost-sensitive information of both the labeled and unlabeled samples for feature extraction.

A. Soft Label Estimation

Let $X = [x_1, x_2, \ldots, x_N, x_{N+1}, \ldots, x_M]$ be a training set, where $x_i|_{i=1}^{N}$ and $x_i|_{i=N+1}^{M}$ are the labeled and unlabeled samples, respectively, the class label of $x_i|_{i=1}^{N}$ is known and assumed to be $l_i \in \{1, 2, \ldots, c\}$, where c is the number of classes. We can construct a cost matrix for the labeled data $X_L = [x_1, x_2, \ldots, x_N]$ as shown in Table I, where C_{ij} denotes the cost of misclassifying the sample of the ith class as the jth class. The diagonal elements in the cost matrix are zero because there is no loss for a correct classification. Generally, it is easy for users to specify which kind of error leads to a higher cost and which leads to a lower cost. Therefore, the cost matrix is assumed to be specified by users in this work and we focus on how to extract discriminative features that benefits from the cost information of both labeled and unlabeled samples for recognition tasks.

In order to perform cost-sensitive discriminant analysis on the whole data for feature extraction, we need to first infer the cost information of the unlabeled samples such that both labeled and unlabeled data can be used. A natural solution is to predict the cost information of unlabeled data from labeled data. While many graph-based semi-supervised learning algorithms have been proposed in recent years, the construction of the graph is the heart of these methods because most existing graph construction methods are sensitive to noises. While there are a number of label reconstruction methods proposed in the literature [12], [31], graph-based is one of the most effective and popular methods for label reconstruction, and the construction of the graph is the heart of these methods because most existing graph construction methods such as KNN and ϵ-ball are sensitive to noises. To address this issue, we propose exploiting the sparse property of samples for label reconstruction.

Inspired by recent sparse coding methods [5], [33] which represent each point as a linear combination of a dictionary and such representations are usually sparse and robust to noises, we here apply all the labeled data $X_L = [x_1, x_2, \ldots, x_N]$ to construct a dictionary and each unlabeled data can be linearly reconstructed by this dictionary X_L. Given one unlabeled sample x_i $(i = N+1, \ldots, M)$, we reconstruct it by solving the following optimization problem:

$$w_i = \arg\min \|w_i\|_1 \quad \text{subject to} \quad \|X_L w_i - x_i\|_2 \le \varepsilon \tag{1}$$

where ε is the reconstruction error bound which is set to 0.01 in our experiments, $w_i \in R^N$ is the representation coefficient vector for the sample x_i.

Let $w_i = [w_{i1}, w_{i2}, \ldots, w_{iN}]$, where w_{ij} is the representation coefficient of x_i from the labeled sample x_j, $u_{ij} = \|w_{ij}\| / \sum_j \|w_{ij}\|$, $1 \le j \le N$. It is easy to see that each unlabeled sample x_i is decomposed into a linear combination of all labeled data x_j associated with a reconstruction coefficient u_{ij}. Obviously, the larger u_{ij} is, the greater the contribution of the sample x_j to the reconstruction of x_i; especially, if $u_{ij} = 1$, it means that x_i is completely reconstructed by x_i, which indicates that x_i is the same as x_j. Therefore, we can now predict the label information of x_i from x_j, $N+1 \le i \le M$.

Since x_i is linearly reconstructed from x_j with a weight u_{ij}, the label information l_{x_i} of x_j can also be linearly propagated to x_i with the same reconstruction coefficients. However, we cannot assign a fixed label value to l_{x_i} because the k nearest labeled neighbors of x_i are usually from different classes as they are heavily dependent on the value of k. Alternatively, we apply a soft label $f_{x_i} = [f_{i1}, f_{i2}, \ldots, f_{ic}]$ to characterize the class information of x_i, where f_{iq} measures the probability of x_i belonging to the qth class and can be calculated as follows:

$$f_{iq} = \sum_{j=1}^{N} u_{ij} \delta(l_j - q) \tag{2}$$

where $\delta(n)$ is a discrete unit impulse function, and defined as

$$\delta(n) = \begin{cases} 1, & n = 0 \\ 0, & n \ne 0. \end{cases} \tag{3}$$

There are several advantages of such operation including robustness to the data noise and an adaptive neighborhood size. Most existing label propagation methods construct the graph by the k-nearest-neighbor (KNN) method based on the pairwise Euclidean distance, which is very sensitive to data noise and the graph structure is easy to change when unfavorable noises come in, such as someone wearing sunglasses that heavily changes the local structure of face samples. Moreover, there is one important parameter k to be tuned in KNN and it is very challenging to select an optimal value of k. These shortcomings can be overcome in our adopted label estimation method, such that better cost information of the unlabeled samples can be exploited for cost-sensitive semi-supervised dimensionality reduction.

B. Objective Function

After the label prediction, we obtain the label information for each unlabeled data and represent it as a label vector. For the labeled data, the label information is fixed and is a number from 1 to c. Therefore, we also represent each labeled data x_j as a label vector $f_j = [f_{j1}, f_{j2}, \ldots, f_{jc}]$, where $f_{jq} = 1$ if x_j belongs to the qth class and zero otherwise. Now, we can obtain a label matrix F for the whole data as follows:

$$F = [f_1; f_2; \cdots; f_M] \tag{4}$$

where the rth row of F is a label vector f_r for x_r, where $1 \le r \le M$.

Different from LDA and SDA, we apply the following tow criteria to calculate the cost-sensitive within-class and between-class variances of CS^{3}DA for discriminative feature extraction.

1) The larger the cost of the qth class, the more importance the class is, and a larger weight should be assigned to this class to calculate the within-class variance of CS^{3}DA.
2) The larger the cost of misclassifying the samples from the qth class as the pth class, $q \ne p$, the larger weight should be applied to calculate the between-class variance of CS^{3}DA.

To characterize the cost of each class, we first define an importance function $h(q)$ to depict the importance of the samples in the qth class, where $1 \le q \le c$. Obviously, there is a number of potential strategies to define $h(q)$ and it is generally believed that $h(q)$ should be a monotone function of cost $C_{q,p}$. In this paper, we use a power function to achieve this goal as follows:

$$h(q) = \sum_{q=1}^{c} C_{q,p}^{a} \tag{5}$$

where a is a tuning parameter to balance the contribution of different classes to computing the cost of the qth class, and $a \ge 1$.

To characterize the different costs of misclassifying the samples from the qth class as the pth class, we define another penalty function $g(q,p)$ to calculate these costs. Obviously, $g_{q,p}$ can be easily found from Table I if the cost matrix C is prespecified, calculated as

$$g(q,p) = C_{q,p}. \tag{6}$$

Now, we calculate the cost-sensitive within-class and between-class scatter matrices of CS^{3}DA as follows:

$$S_b^{cs} = \sum_{q=1}^{c}\sum_{p=1}^{c} g(q,p)(\tilde{m}_q - \tilde{m}_p)(\tilde{m}_q - \tilde{m}_p)^T \tag{7}$$

$$S_w^{cs} = \sum_{q=1}^{c}\sum_{j=1}^{M} h(q) F_{j,q} (x_j - \tilde{m}_q)(x_j - \tilde{m}_q)^T \tag{8}$$

where $\tilde{m}_q$ and $\tilde{m}_q$ are the estimated means of the qth and pth classes, respectively, which can be calculated as follows:

$$\tilde{m}_q = \frac{1}{N_q}\sum_{j=1}^{M} F_{jq} x_j, \quad \tilde{m}_p = \frac{1}{N_p}\sum_{j=1}^{M} F_{jp} x_j \tag{9}$$

where $N_q = \sum_{j=1}^{M} F_{jq}$ and $N_p = \sum_{j=1}^{M} F_{jp}$.

Having obtained S_b^{cs} and S_w^{cs}, we can obtain the objective function of our cost-sensitive semi-supervised discriminant analysis

$$\arg\max_w \frac{w^T S_b^{cs} w}{w^T S_w^{cs} w}. \tag{10}$$

The projection vector w that maximizes the objective function can be easily obtained by selecting the maximal eigenvalue solution to the following generalized eigenvalue equation:

$$S_b^{cs} w = \lambda S_w^{cs} w. \tag{11}$$

When the number of data points is smaller than the dimensionality of each sample, S_w^{cs} in (11) may be singular and hence the eigen-decomposition problem will be unstable. To avoid this problem, we apply the idea of Tikhonov regularization as in regularized discriminant analysis [8]. Therefore, the generalized eigenvalue problem in (11) becomes

$$S_b^{cs} w = \lambda (S_w^{cs} + \mu I) w \tag{12}$$

where $\mu > 0$ and I is an identity matrix.

C. Algorithm

The $\mathrm{CS^3DA}$ algorithm can be summarized as follows:

1) **Reconstruction Weight Calculation**: Compute the reconstruction weight for each unlabeled sample by using (1).
2) **Label Prediction**: Obtain a soft label vector for each unlabeled sample by using (2).
3) **Label Matrix Construction**: Construct a label matrix F for the whole data including both labeled and unlabeled samples by using (4).
4) **Cost-Sensitive Scatter Calculation**: Calculate the cost-sensitive within-class and between-class scatter matrices of $\mathrm{CS^3DA}$ by using (5)–(9).
5) **Eigen-Problem**: Solve the generalized eigenvalue equation in (12).
6) **$\mathbf{CS^3DA}$ Embedding**: Let $\{w_1, w_2, \ldots, w_k\}$ be the eigenvectors corresponding to the k largest eigenvalues ordered such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$, and $W = [w_1, w_2, \ldots, w_k] \in R^{d\times k}$ is the projection matrix of $\mathrm{CS^3DA}$. The samples can be embedded into a k dimensional subspace by

$$x \rightarrow y = W^T x. \tag{13}$$

D. Theoretic Analysis

We now briefly analyze how our proposed cost-sensitive scatter matrices affect the classification error. Since it is generally too complex to use the Bayes error directly as a criterion to derive the feature subspace for classification, one resorts to the Fisher criteria that is suboptimal but that is easier to optimize. However, as proved by [15], such criteria aims to seek the linear transformation that maximizes the mean-squared distance between the classes in the lower-dimensional space, which is clearly different from minimizing the classification error. To address this shortcoming, Loog *et al.* [15] introduced a weighted pairwise Fisher criteria to seek a feature subspace which is more closely related to the classification error, described as follows:

$$\max J(w) = \sum_{i=1}^{c} \sum_{j=i+1}^{c} p_i p_j f(\triangle_{ij}) \frac{tr(w^T S_{ij} w)}{tr(w^T S_w w)} \tag{14}$$

where p_i and p_j are *priori* probabilities of the ith and jth classes, $\triangle_{ij}$ is the Mahanalobis distance between the ith and jth classes, calculated as

$$\triangle_{ij} = \sqrt{(m_i - m_j)^T S_w^{-1} (m_i - m_j)} \tag{15}$$

S_w is the within-class scatter, S_{ij} is the between-class scatter between the ith and jth classes, S_w is the within-class scatter, $f(\cdot)$ is a weighting function which depends on the Mahanalobis distance $\triangle_{ij}$ between the ith and jth classes, defined as

$$S_{ij} = (m_i - m_j)^T (m_i - m_j). \tag{16}$$

Having compared (7)–(14), we can observe the objective of our proposed $\mathrm{CS^3DA}$ is consistent with that of the weighted pairwise Fisher criteria in terms of the classification error, and the functions of the penalty functions $g(q,p)$ and $h(q)$ are similar to that of the weighting function $f(\cdot)$ in the weighted pairwise Fisher criteria. Moreover, according to the definitions of cost-sensitive within-class and between-class scatter matrices of $\mathrm{CS^3DA}$ [(7) and (8)], we can see that the samples in the classes which have a larger importance play a more important role in seeking the feature subspace because these classes contributes more to calculate the between-class Mahanalobis distance. In other words, the classes are more difficult to be misclassified and hence a smaller total cost can be obtained.

E. Discussion

In this section, we discuss the relationship between our proposed $\mathrm{CS^3DA}$ method and some related works: the existing discriminant analysis algorithms [1], [3], [11], [21], [27], [30], [34], [35], [37], cost-sensitive learning techniques [13], [20], [22], [28], [38], and propagation/label estimation methods.

1) Relationship With Existing Discriminant Analysis Algorithms: Most existing discriminant analysis algorithms, supervised [1], [11], [27], [35] or semi-supervised [3], [21], [30], [34], [37], aim to seek a low recognition error rate and implicitly assume that the losses of all misclassifications are the same. Hence, they can be considered as cost-blind discriminant analysis methods. More recently, Lu and Tan [17] proposed a cost-sensitive linear discriminant analysis (CSLDA) method for face recognition. However, their method is supervised and the label information of all training samples needs to be known in advance. As we discussed in Section I, it is usually difficult, expensive and time-consuming to collect sufficient labeled data because laborious human labeling effort is required, which will limit the convenience of supervised methods in practical applications. Moreover, when there is only one labeled sample per class, these supervised methods will fail to work as the intraclass variance cannot be estimated. However, our proposed $\mathrm{CS^3DA}$ method performs feature extraction and recognition to attain low losses rather than low recognition errors and extracts cost-sensitive discriminant features from both labeled and unlabeled data simultaneously.

2) Relationship With Existing Cost-Sensitive Learning Techniques: Cost-sensitive learning is one important and hot research topic in the pattern recognition and machine learning community, and several cost-sensitive learning methods have been proposed in the literature [13], [20], [22], [28], [38], such as cost-sensitive boosting [20], [28], cost-sensitive support vector machine [13], cost-sensitive k-nearest neighbor classifier [38], and cost-sensitive neural networks [22]. While improved recognition performance (smaller recognition loss) can be obtained, these methods apply the cost-sensitive learning techniques only in the classification phase and not in the feature extraction phase since they make use of the conventional cost-blind features for recognition. Different from these methods, we focus on extracting cost-sensitive discriminative features in this work from both labeled and unlabeled samples for recognition tasks.

3) Relationship With Existing Label Propagation/Estimation Methods: Label propagation/estimation has been widely used in semi-supervised learning [21], [31], [40], multiclass learning [7], and multilabel learning [12], [14]. As we mentioned in Section I, the graph-based method is one of the most active and effective approaches for label propagation. Different from most previous label propagation methods which reconstruct each data linearly from its neighborhood, we apply the sparse coding technique to estimate the labels of the unlabeled training samples from the labeled training samples. There are several advantages of such an operation including robustness to the data noise and an adaptive neighborhood size. Most existing label propagation methods construct the graph by the k-nearest-neighbor (KNN) method based on the pairwise Euclidean distance, which is very sensitive to data noise and the graph structure is easy to change when unfavorable noises come in, such as someone wearing sunglasses that heavily change the local structure of face samples. Moreover, there is one important parameter k to be tuned in KNN and it is very challenging to select an optimal value. These shortcomings can be overcome in our adopted label estimation method, such that better cost information of the unlabeled samples can be exploited for cost-sensitive semi-supervised dimensionality reduction.

III. EXPERIMENTS

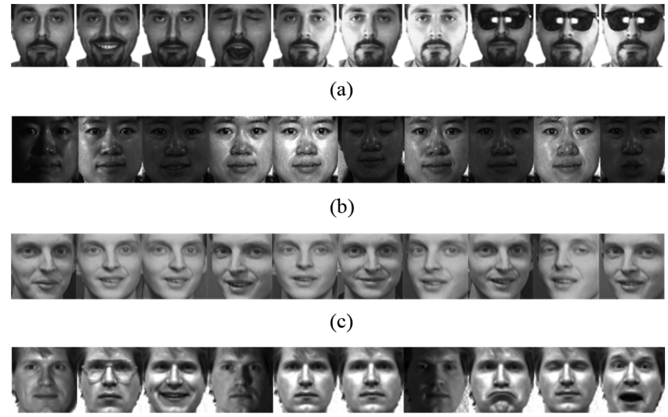
In this section, we report experimental results on four widely used face databases to evaluate the performance of our proposed $\mathrm{CS}^3\mathrm{DA}$ for face recognition.

A. Data Sets

We made use of the AR [19], CMU PIE [24], ORL [23] and Yale [1] face databases in our experiments.

The AR database contains over 4000 face images of 126 subjects (70 men and 56 women), including frontal facial images with different facial expressions, lighting conditions and occlusions. There are 26 images for each person taken in two sessions, each having 13 images.

The PIE database comprises 68 subjects with 41 368 face images of different poses, illuminations and expressions. In our experiments, we selected the frontal pose with varying expressions and illuminations to construct a subset of the PIE database. It contains 3060 frontal face images of different expressions and illuminations from 68 subjects with 45 images from each subject.



(a)

(b)

(c)

(d)

Fig. 1. Sample images for one subject in the (a) AR, (b) PIE, (c) ORL, and (d) Yale databases, respectively.

The ORL face database, contains a total of 400 images of 40 subjects with 10 grayscale face images for each. The images show all frontal and slight tilt/rotation of the face up to 20. For some subjects, the images were taken at different times, varying lighting, facial expressions (open or closed eyes, smiling or not smiling), and facial details (glasses or no glasses).

The Yale face database is constructed at the Yale center for computational vision and control. It contains 165 grayscale images of 15 individuals. The images demonstrate variations in lighting condition (left-light, center-light, right-light), facial expression (normal, happy, sad, sleepy, surprised, and wink), and with/without glasses.

In our experiments, the images from all of the four datasets were manually aligned, cropped, and resized to 32×32 pixels according to the eyes' positions. Fig. 1 shows some sample images of one subject from each dataset.

B. Experimental Settings

While there are a number of application scenarios for face recognition, in this study, we only take face recognition-based access control as an example to illustrate the effectiveness of our proposed method. It is noted that our method can be easily extended to other face recognition applications by designing a different cost matrix according to different application requirements and preferences.

For a face recognition-based access control system, there are usually three types of misclassifications on recognizing a test face sample:

1) *False rejection:* misrecognizing a gallery person as an impostor;
2) *False acceptance:* misrecognizing an impostor as a gallery person;
3) *False identification:* misrecognizing a gallery person as another gallery person.

While all possible forms of loss are allowed in a face recognition-based access control system, for convenience of our discussion, we assume that accepting any impostor will cause

TABLE II
COST MATRIX OF FACE RECOGNITION-BASED ACCESS CONTROL SYSTEM

	G_1	$\cdots$	G_c	I
G_1	0	$\cdots$	C_{GG}	C_{GI}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
G_c	C_{GG}	$\cdots$	0	C_{GI}
I	C_{IG}	$\cdots$	C_{IG}	0

the same loss, and misrecognizing a gallery person as another gallery person or an impostor will cause different amounts of loss. Without any *a priori* information or preference, this assumption is reasonable for most practical face recognition-based access control systems. Let C_{GI}, C_{IG} and C_{GG} be the costs incurred by a false rejection, false acceptance and false identification, respectively, then the cost matrix for this face recognition system will be simplified into Table II, where C_{ij} denotes the cost of misrecognizing a face image of the ith class as the jth class. Without loss of generality, we assume that $C_{GI} = C_{GI}/C_{GG}$, $C_{IG} = C_{IG}/C_{GG}$ and $C_{GG} = 1$ as this will not change the final results.

For each database, we randomly selected M subjects as the gallery persons and the remaining subjects as the imposters. In the training phase, we randomly selected N_G^{tr} images for each gallery person, where G_L samples are labeled and the other G_U samples are unlabeled. For the impostors in the training set, we randomly selected N_I^{tr} images, where N_L samples are labeled and the other N_U are unlabeled. In the testing phase, there are N_G^{te} images for each gallery person, as well as N_I^{te} imposter images. We repeated this selection ten times for each database and computed the average recognition accuracy as the final recognition result. The parameters M, N_G^{tr}, G_L, G_U, N_I^{tr}, I_L, I_U, N_G^{te}, N_I^{te}, C_{IG} and C_{GI} and C_{GG} are specified in Table III.

C. Results and Analysis

1) Comparisons of Different Discriminant Analysis Methods: We compared the proposed $\mathrm{CS^3DA}$ with the following five other discriminant analysis methods.

1) LDA [1]: Since LDA is a supervised method, only labeled samples in the training set can be used for feature extraction. We applied a Tikhonov regularization term μI rather than performing PCA to avoid the well-known small sample size (SSS) problem in LDA.
2) PLDA [39]: We followed the parameter setting in [39] to make use of perturbation covariance matrices to calculate the between-class and within-class scatters for discriminant analysis.
3) SDA [3]: There is a parameter α to be tuned in SDA and it was empirically set to one in our experiments. The Tikhonov regularization was also adopted to avoid the SSS problem.
4) SDAPC [32]: SDAPC uses the pairwise constraints for semi-supervised dimensionality reduction. We obtain the pairwise constraints from the labeled samples in the training set. If two samples are from the same class, there is a must-link constraint and otherwise there is a cannot-link constraint.
5) CSLDA [17]: Different from the above four discriminant analysis algorithms, CSLDA is a cost-sensitive dimensionality reduction method. Since it is supervised, only the labeled samples in the training set are employed for feature extraction. The Tikhonov regularization was also used to avoid the SSS problem.

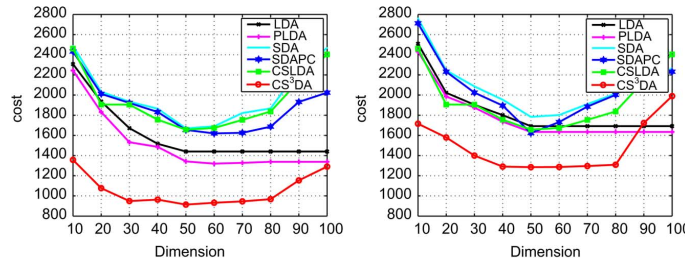


Fig. 2. Total cost versus different feature dimensions. Results on the (a) unlabeled data of the AR database and (b) testing data of the AR database, respectively.

For each database, the recognition performance of different methods were evaluated on the unlabeled data and the testing data separately. We applied the conventional nearest neighbor (NN) rule in the Euclidean space to perform recognition. We compared the total cost (cost), total error rate (err), error rate of false rejection (err_{GI}), and error rate of false acceptance (err_{IG}) of the methods. Tables IV–VII report the results obtained on the AR, CMU PIE, ORL and Yale databases, respectively. Fig. 2 shows the total cost of different methods versus different feature dimensions on the AR database. We can see our proposed method consistently achieves smaller total cost than other compared methods versus different feature dimensions.

2) Comparisons of Different Cost-Sensitive Learning Methods: We compared our cost-sensitive semi-supervised dimensionality reduction method with the state-of-the-art cost-sensitive learning method: cost-sensitive semi-supervised support vector machine $(\mathrm{CS^4VM})$ [13] for cost-sensitive face recognition experiments. We conducted three sets of face recognition experiments according to the following settings.

1) $\mathrm{CS^3DA}$ + NN: We performed feature extraction using our proposed $\mathrm{CS^3DA}$ method and performed recognition using the conventional nearest neighborhood classifier with the Euclidean metric.
2) $\mathrm{SDA+CS^4VM}$: We performed feature extraction using the SDA method and performed recognition using the state-of-the-art $\mathrm{CS^4VM}$ cost-sensitive classifier.
3) $\mathrm{CS^3DA+CS^4VM}$: Both the feature extraction and recognition phases use the cost-sensitive techniques for face recognition.

Fig. 3 shows the total costs of these experiments on different databases. We can see from this figure that our cost-sensitive approach is comparable to existing cost-sensitive classifier, and the total cost can be further reduced when both our cost-sensitive feature extraction method and the existing cost-sensitive classifier was applied simultaneously.

3) Comparisons of Different Label Estimation Methods: We compared our label estimation method with the linear neighborhood propagation (LNP) [31] which estimates the labels of the unlabeled samples by linearly reconstructing it from its

TABLE III
EXPERIMENTAL SETTINGS FOR FACE RECOGNITION EXPERIMENTS

Database	M	N_G^{tr}	G_L	G_U	N_I^{tr}	N_L	N_U	N_G^{te}	N_I^{te}	$C_{IG}:C_{GI}:C_{GG}$
AR	50	13	2	11	130	20	110	650	130	20:2:1
PIE	60	30	2	28	240	16	224	900	120	20:2:1
ORL	30	7	4	3	70	30	40	90	30	20:2:1
Yale	12	8	3	5	24	9	15	36	9	20:2:1

TABLE IV
COMPARISON OF TOTAL COST (COST), TOTAL ERROR RATE (ERR), ERROR RATE OF FALSE REJECTION (err_{GI}), AND ERROR RATE OF FALSE ACCEPTANCE (err_{IG}) OF DIFFERENT DISCRIMINANT ANALYSIS METHODS ON AR DATABASE

Method	Unlabeled data				Testing data			
	cost	err (%)	err_{GI} (%)	err_{IG} (%)	cost	err (%)	err_{GI} (%)	err_{IG} (%)
LDA	1440	57.42	60.00	44.55	1692	75.38	84.46	30.00
PLDA	1320	55.32	60.00	42.55	1634	72.32	81.46	28.62
SDA	1670	81.82	89.82	41.82	1784	85.26	96.92	26.92
SDAPC	1620	80.54	86.52	36.82	1624	82.12	94.92	23.92
CSLDA	1327	54.79	56.00	42.73	1657	74.23	83.08	30.00
CS^{3}DA	914	70.91	85.09	0.00	1285	81.41	97.54	0.77

TABLE V
COMPARISON OF TOTAL COST (COST), TOTAL ERROR RATE (ERR), ERROR RATE OF FALSE REJECTION (err_{GI}), AND ERROR RATE OF FALSE ACCEPTANCE (err_{IG}) OF DIFFERENT DISCRIMINANT ANALYSIS METHODS ON CMU PIE DATABASE

Method	Unlabeled data				Testing data			
	cost	err (%)	err_{GI} (%)	err_{IG} (%)	cost	err (%)	err_{GI} (%)	err_{IG} (%)
LDA	4286	68.96	70.06	60.71	2766	63.92	61.33	90.83
PLDA	3896	63.24	66.84	56.42	2654	62.32	58.96	84.66
SDA	3646	78.10	85.48	22.77	1952	70.00	74.67	35.00
SDAPC	3543	77.54	83.82	21.82	1878	68.22	95.22	34.23
CSLDA	3795	62.34	63.69	52.23	2518	51.47	46.89	85.83
CS^{3}DA	2648	69.59	78.39	3.57	1647	69.31	60.33	30.83

TABLE VI
COMPARISON OF TOTAL COST (COST), TOTAL ERROR RATE (ERR), ERROR RATE OF FALSE REJECTION (err_{GI}), AND ERROR RATE OF FALSE ACCEPTANCE (err_{IG}) OF DIFFERENT DISCRIMINANT ANALYSIS METHODS ON ORL DATABASE

Method	Unlabeled data				Testing data			
	cost	err (%)	err_{GI} (%)	err_{IG} (%)	cost	err (%)	err_{GI} (%)	err_{IG} (%)
LDA	983	36.67	22.78	78.33	976	35.54	20.42	25.55
PLDA	974	35.45	21.34	76.86	952	34.42	20.12	24.86
SDA	954	34.32	20.12	74.32	934	32.22	19.86	23.46
SDAPC	932	33.32	19.32	73.36	930	32.22	19.82	22.42
CSLDA	912	33.15	19.86	72.32	924	32.12	19.68	22.22
CS^{3}DA	864	32.15	32.34	50.20	846	32.01	32.44	12.22

TABLE VII
COMPARISON OF TOTAL COST (COST), TOTAL ERROR RATE (ERR), ERROR RATE OF FALSE REJECTION (err_{GI}), AND ERROR RATE OF FALSE ACCEPTANCE (err_{IG}) OF DIFFERENT DISCRIMINANT ANALYSIS METHODS ON YALE DATABASE

Method	Unlabeled data				Testing data			
	cost	err (%)	err_{GI} (%)	err_{IG} (%)	cost	err (%)	err_{GI} (%)	err_{IG} (%)
LDA	683	25.67	20.12	75.33	732	28.44	22.42	26.86
PLDA	654	25.33	19.88	75.13	714	26.33	21.36	25.68
SDA	662	24.33	19.96	74.86	708	24.56	20.23	24.96
SDAPC	624	23.87	19.82	74.55	694	22.68	20.12	23.46
CSLDA	618	23.50	19.64	74.33	682	22.54	19.32	22.26
CS^{3}DA	594	23.33	32.45	32.33	654	22.32	32.34	10.22

t-nearest neighbors. In our experiments, t was empirically set to be five for all datasets. We performed recognition experiments with the NN classifier. Fig. 4 shows the total cost of our method when using different label estimation methods. We can observe from this figure that our soft label estimation method can attain smaller cost than the existing linear neighborhood label propagation method. The reason is our label estimation method applies the sparse coding technique, which is naturally discriminative in label propagation.

4) Parameter Analysis: Lastly, we investigated the effect of parameter k on the performance of CS^{3}DA. The total cost, total error rate, error rate of false rejection, and error rate of false acceptance versus different values of parameters on the AR and ORL databases are recorded in Tables VIII and IX. We can see our proposed method is insensitive to the parameter k on both the AR and ORL datasets.

D. Discussion

We can make the following four observations from the results listed in Tables IV–IX and Figs. 2–4.

1) CSLDA consistently outperforms LDA in terms of total cost on all experiments, which further implies that explic-

TABLE VIII
TOTAL COST (COST), TOTAL ERROR RATE (ERR), ERROR RATE OF FALSE REJECTION (err_{GI}), AND ERROR RATE OF FALSE ACCEPTANCE (err_{IG}) OF OUR METHOD VERSUS DIFFERENT VALUES OF k ON AR DATABASE

k	Unlabeled data				Testing data			
	cost	err (%)	err_{GI} (%)	err_{IG} (%)	cost	err (%)	err_{GI} (%)	err_{IG} (%)
1	1136	78.32	86.32	0.00	1424	82.66	97.86	0.77
2	1085	75.98	86.24	0.00	1398	82.32	97.82	0.77
3	986	73.12	86.12	0.00	1342	82.02	97.74	0.77
4	936	71.43	85.96	0.00	1296	81.86	97.64	0.77
5	914	70.91	85.09	0.00	1285	81.41	97.54	0.77
6	933	73.32	85.56	0.00	1298	81.12	97.58	0.77
7	933	74.36	85.88	0.00	1324	82.23	97.64	0.77
8	941	76.56	86.24	0.00	1346	82.65	97.76	0.77
9	950	77.42	86.46	0.00	1368	82.97	97.82	0.77
10	968	78.68	86.86	0.00	1388	83.32	97.86	0.77

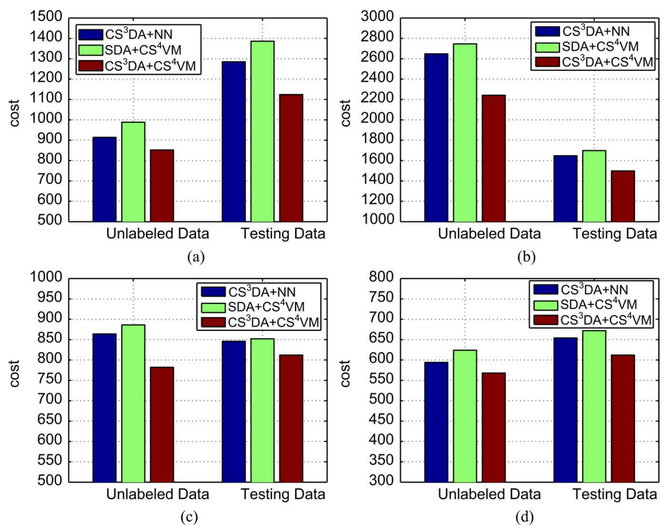


Fig. 3. Total cost on different databases for different cost-sensitive dimensionality reduction methods when the uniform-cost and cost-sensitive classifiers are applied. Results on (a) AR, (b) CMU PIE, (c) ORL, and (d) Yale database, respectively.

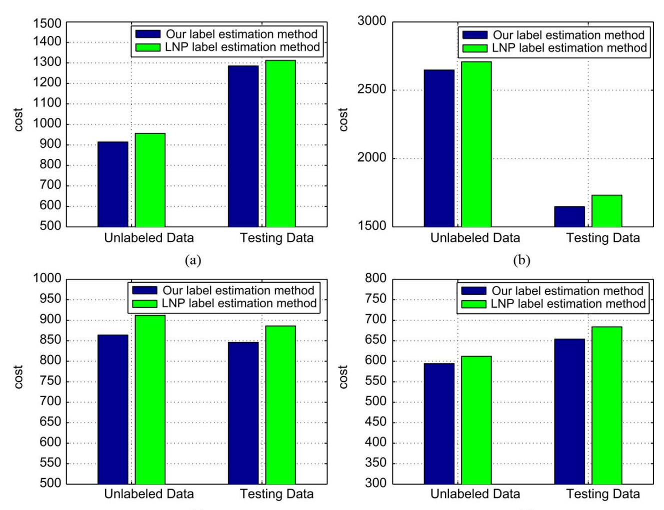


Fig. 4. Total cost on different databases when different label estimation methods are applied. Results on the (a) AR, (b) CMU PIE, (c) ORL and (d) Yale database, respectively.

itly extracting cost-sensitive discriminant features can help improve the recognition performance of the existing uniform-cost discriminant analysis methods.

2) SDA outperforms CSLDA on the CMU PIE database, and CSLDA outperforms SDA on the AR database. The reason is there are some images with occlusion in the AR database which resulting in large intraclass variance such that nearby points may belong to different classes. Under this scenario, the assumption of SDA may not hold and semi-supervised discriminant information preserved in SDA may not be benefit to recognition.
3) CS^{3}DA consistently outperforms the other discriminant analysis methods in all experiments, which implies extracting a discriminative feature space by making use of both labeled and unlabeled data and explicitly considering the cost of each data simultaneous can achieve the best recognition performance for discriminant analysis.
4) Our proposed cost-sensitive feature extraction method can achieve comparable performance with the existing cost-sensitive classifier when only the cost-blind NN classifier is used. Moreover, the performance can be further improved when both the cost-sensitive feature extraction and classification methods were used simultaneously. The reason is that there are usually two phases for a practical face recognition system: feature extraction (dimensionality reduction) and classification, and we can improve the recognition performance of the face recognition system at either the feature extraction or the classification phase. Hence, if we exploit the cost information at these two phases simultaneously, better recognition performance (in terms of the cost) can be obtained. Therefore, cost-sensitive classifiers can further improve the performance of cost-sensitive dimensionality reduction methods for face recognition.

IV. CONCLUSION AND FUTURE WORK

We have proposed in this paper a new dimensionality reduction algorithm called cost-sensitive semi-supervised discriminant analysis (CS^3 DA). It can make use of both labeled and unlabeled data to learn a low-dimensional feature space to achieve dimensionality reduction in a cost-sensitive setting. Each unlabeled sample was firstly linearly reconstructed from labeled data and the reconstruction coefficients were used to estimate a soft label for the unlabeled data. Then, both these label and unlabeled data were used to learn a low-dimensional feature subspace aiming at seeking low overall loss when recognition is performed in the feature space derived. Experimental results

TABLE IX
TOTAL COST (COST), TOTAL ERROR RATE (ERR), ERROR RATE OF FALSE REJECTION (err_{GI}), AND ERROR RATE OF FALSE ACCEPTANCE (err_{IG}) OF OUR METHOD VERSUS DIFFERENT VALUES OF k ON ORL DATABASE

k	Unlabeled data				Testing data			
	cost	err (%)	err_{GI} (%)	err_{IG} (%)	cost	err (%)	err_{GI} (%)	err_{IG} (%)
1	1104	35.24	34.84	52.23	932	33.23	34.86	13.86
2	1042	34.32	33.66	51.86	912	33.12	34.32	13.64
3	986	33.55	32.12	51.32	876	32.98	33.94	12.86
4	884	32.65	32.54	50.64	856	32.51	32.84	12.64
5	864	32.15	32.34	50.20	846	32.01	32.44	12.22
6	876	32.66	32.62	50.86	854	32.86	33.24	12.56
7	886	33.12	32.96	51.32	862	33.02	33.86	12.78
8	912	33.26	33.36	51.98	876	33.32	34.32	12.96
9	932	33.88	33.88	52.32	888	33.46	34.32	13.02
10	944	34.24	34.12	52.86	896	33.88	34.32	13.12

on face recognition have demonstrated the efficacy of the proposed method. How to learn an appropriate cost matrix for a specific cost-sensitive pattern recognition application automatically seems another interesting direction of our future work.

ACKNOWLEDGMENT

The authors would like to express their thanks to the associate editor and anonymous reviewers for their supportive comments and useful suggestions that have helped them to improve the quality of the paper.

REFERENCES

[1] P. N. Belhumeur, J. Hespanha, and D. J. Kriegman, “Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[2] W. Bian and D. Tao, “Biased discriminant Euclidean embedding for content-based image retrieval,” *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 545–554, Feb. 2010.

[3] D. Cai, X. He, and J. Han, “Semi-supervised discriminant analysis,” in *Proc. IEEE Int. Conf. Computer Vision*, 2007, pp. 1–7.

[4] O. Chapelle *et al., Semi-Supervised Learning*. Cambridge, MA: MIT Press, 2006.

[5] B. Chen, J. Yang, S. Yan, Y. Fu, and T. Huang, “Learning with $l1$-graph for image analysis,” *IEEE Trans. Image Process.*, vol. 19, no. 5, pp. 856–866, May 2010.

[6] H. Chen, H. Chang, and T. Liu, “Local discriminant embedding and its variants,” in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, 2005, pp. 846–853.

[7] H. Cheng, Z. Liu, and J. Yang, “Sparsity induced similarity measure for label propagation,” in *Proc. IEEE Int. Conf. Computer Vision*, 2009, pp. 317–324.

[8] J. Friedman, “Regularized discriminant analysis,” *J. Amer. Statistical Assoc.*, vol. 84, no. 405, pp. 165–175, 1989.

[9] Y. Fu, S. Yan, and T. S. Huang, “Classification and feature extraction by simplexization,” *IEEE Trans. Inform. Forensics Security*, vol. 3, no. 1, pp. 91–100, Feb. 2010.

[10] X. He, S. Yan, Y. Hu, P. Niyogi, and H. J. Zhang, “Face recognition using Laplacian faces,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.

[11] H. Hu, “Orthogonal neighborhood preserving discriminant analysis for face recognition,” *Pattern Recognition*, vol. 41, no. 6, pp. 2045–2054, 2008.

[12] F. Kang, R. Jin, and R. Sukthankar, “Correlated label propagation with application to multi-label learning,” in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, 2006, vol. 2, pp. 1719–1726.

[13] Y. F. Li, J. Kwok, and Z. Zhou, “Cost-sensitive semi-supervised support vector machine,” in *Proc. Nat. Conf. Artificial Intell.*, 2010, pp. 500–505.

[14] Y. Liu, R. Jin, and L. Yang, “Semi-supervised multi-label learning by constrained non-negative matrix factorization,” in *Proc. Nat. Conf. Artifical Intelligence*, 2006, vol. 21, pp. 421–421.

[15] M. Loog, R. Duin, and R. Haeb-Umbach, “Multiclass linear dimension reduction by weighted pairwise fisher criteria,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, no. 7, pp. 762–766, Jul. 2001.

[16] J. Lu and Y.-P. Tan, “A doubly weighted approach for appearance-based subspace learning methods,” *IEEE Trans. Inform. Forensics Security*, vol. 5, no. 1, pp. 71–81, Feb. 2012.

[17] J. Lu and Y.-P. Tan, “Cost-sensitive subspace learning for face recognition,” in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, 2010, pp. 2661–2666.

[18] J. Lu and Y.-P. Tan, “Regularized locality preserving projections and its extensions for face recognition,” *IEEE Trans. Syst., Man, Cybern., Part B*, vol. 40, no. 2, pp. 958–963, Feb. 2010.

[19] A. M. Martinez and R. Benavente, AR Face Database 1998, CVC Tech. Rep..

[20] H. Masnadi-Shirazi and N. Vasconcelos, “Cost-sensitive boosting,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 33, no. 2, pp. 294–309, Feb. 2011.

[21] F. Nie, S. Xiang, Y. Jia, and C. Zhang, “Semi-supervised orthogonal discriminant analysis via label propagation,” *Pattern Recognition*, vol. 42, no. 11, pp. 2615–2627, 2009.

[22] S. Raudys and A. Raudys, “Pariwise costs in multiclass perceptions,” *IEEE Trans. Pattern Analy. Machine Intell.*, vol. 32, no. 7, pp. 1324–1328, Jul. 2010.

[23] F. Samaria and A. Harter, “Parameterisation of a stochastic model for human face identification,” in *Proc. IEEE Workshop Applications Computer Vision*, 1994, pp. 138–142.

[24] T. Sim, S. Baker, and M. Bsat, “The CMU pose, illumination, and expression (PIE) database,” in *Proc. IEEE Int. Conf. Automatic Face Gesture Recognition*, 2002, pp. 46–51.

[25] Y. Song, F. Nie, and C. Zhang, “Semi-supervised sub-manifold discriminant analysis,” *Pattern Recognition Lett.*, vol. 29, no. 13, pp. 1806–1813, 2008.

[26] Y. Song, F. Nie, C. Zhang, and S. Xiang, “A unified framework for semi-supervised dimensionality reduction,” *Pattern Recognition*, vol. 41, no. 9, pp. 2789–2799, 2008.

[27] M. Sugiyama, “Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis,” *J. Machine Learning Res.*, vol. 8, pp. 1027–1061, 2007.

[28] Y. Sun, M. Kamel, A. Wong, and Y. Wang, “Cost-sensitive boosting for classification of imbalanced data,” *Pattern Recognition*, vol. 40, no. 12, pp. 3358–3378, 2007.

[29] M. Turk and A. Pentland, “Eigenfaces for recognition,” *J. Cognitive Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.

[30] F. Wang, X. Wang, and T. Li, “Beyond the graphs: Semi-parametric semi-supervised discriminant analysis,” in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, 2009, pp. 2113–2120.

[31] F. Wang and C. Zhang, “Label propagation through linear neighborhoods,” *IEEE Trans. Knowledge Data Eng.*, vol. 1, pp. 55–67, 2008.

[32] N. Wang, X. Li, Y. Cui, and J. Pan, “Instance-level based discriminative semi-supervised dimensionality reduction with chunklets,” *Int. J. Innovative Computing, Information Control*, vol. 6, no. 8, pp. 3763–3773, 2010.

[33] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[34] D. Xu and S. Yan, “Semi-supervised bilinear subspace learning,” *IEEE Trans. Image Process.*, vol. 18, no. 7, pp. 1671–1676, Jul. 2009.

[35] 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. Machine Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.

[36] S. Zhang and T. Sim, "Discriminant subspace analysis: A Fukunaga-koontz approach," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 29, no. 10, pp. 1732–1745, Oct. 2007.

[37] Y. Zhang and D. Yeung, "Semi-supervised discriminant analysis using robust path-based similarity," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, 2008, pp. 1–8.

[38] Y. Zhang and Z. Zhou, "Cost-sensitive face recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 32, no. 10, pp. 1758–1769, Oct. 2010.

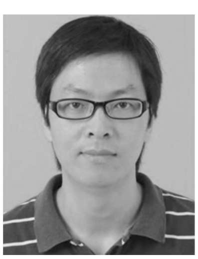
[39] W. Zheng, J. Lai, P. Yuen, and S. Li, "Perturbation lda: Learning the difference between the class empirical mean and its expectation," *Pattern Recognition*, vol. 42, no. 5, pp. 764–779, 2009.

[40] X. Zhu, "Semi-supervised learning literature survey," *Computer Science, Univ. Wisconsin-Madison*, 2006.

Jiwen Lu (S'10–M'11) received the B.Eng. degree in mechanical engineering and the M.Eng. degree in signal and information processing from the Xi'an University of Technology, Xi'an, China, and the Ph.D. degree in electrical engineering from Nanyang Technological University, Singapore, respectively.

He is currently a Research Fellow at the Advanced Digital Sciences Center, Singapore. His research interests include computer vision, pattern recognition, machine learning, multimedia computing and biometrics. He has authored more than 60 scientific papers in peer-reviewed journals and conferences.

Dr. Lu received the First Prize National Scholarship and the National Outstanding Student Recognition awarded by the Ministry of Education of China, in 2002 and 2003, and the Best Student Paper Award from PREMIA in 2012, respectively.



Xiuzhuang Zhou (M'11) received the B.Sci. degree from the Department of Atmosphere Physics, Chengdu University of Information Technology, Chengdu, China, in 1996. He received the M.Eng. and Ph.D. degrees from the School of Computer Science, Beijing Institute of Technology, Beijing, China, in 2005 and 2011, respectively.

He is currently an Assistant Professor in the College of Information Engineering, Capital Normal University, Beijing, China. His research interests include computer vision, pattern recognition, and machine learning. He has authored several scientific papers in peer-reviewed journals and conferences.

Yap-Peng Tan (M'97–SM'04) received the B.S. degree from National Taiwan University, Taipei, Taiwan, in 1993, and the M.A. and Ph.D. degrees from Princeton University, Princeton, NJ, in 1995 and 1997, respectively, all in electrical engineering.

From 1997 to 1999, he was with Intel Corporation, Chandler, AZ, and Sharp Laboratories of America, Camas, WA. In November 1999, he joined the Nanyang Technological University of Singapore, where he is currently Associate Professor and Associate Chair (Academic) of the School of Electrical and Electronic Engineering. His current research interests include image and video processing, content-based multimedia analysis, computer vision, and pattern recognition. He is the principal inventor or co-inventor on 15 U.S. patents in the areas of image and video processing. He was the recipient of an IBM Graduate Fellowship from IBM T. J. Watson Research Center, Yorktown Heights, NY, from 1995 to 1997.

Dr. Tan currently serves as the Secretary of the Visual Signal Processing and Communications Technical Committee of the IEEE Circuits and Systems Society, a member of the Multimedia Signal Processing Technical Committee of the IEEE Signal Processing Society, and a voting member of the ICME Steering Committee. He is an Editorial Board Member of the EURASIP *Journal on Advances in Signal Processing* and EURASIP *Journal on Image and Video Processing*, and an Associate Editor of the *Journal of Signal Processing Systems*. He was also a Guest Editor for special issues of several journals including the IEEE TRANSACTIONS ON MULTIMEDIA. He was the General Co-Chair of the 2010 IEEE International Conference on Multimedia and Expo (ICME 2010).

Yuanyuan Shang (M'12) received the Ph.D. degree from National Astronomical Observatories, Chinese Academy of Sciences, in 2005.

She is currently a Professor and Vice Dean (research) of the College of Information Engineering, Capital Normal University, Beijing, China. Her research interests include digital imaging sensor, image processing, and computer vision. She has authored over more than 40 scientific papers in peer-reviewed journals and conferences.

Jie Zhou (M'01–SM'04) was born in November 1968. He received the B.S. and M.S. degrees both from the Department of Mathematics, Nankai University, Tianjin, China, in 1990 and 1992, respectively. He received the Ph.D. degree from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology (HUST), Wuhan, China, in 1995.

From 1995 to 1997, he served as a Postdoctoral Fellow in Department of Automation, Tsinghua University, Beijing, China. Since 2003, he has been a Full Professor in Department of Automation, Tsinghua University. His research interests include computer vision, pattern recognition and image processing. In recent years, he has authored more than 100 papers in peer-reviewed journals and conferences. Among them, more than 30 papers have been published in top journals and conferences.

Dr. Zhou is an associate editor for *International Journal of Robotics and Automation*, *Acta Automatica* and two other journals.