Cost-Sensitive Label Propagation for Semi-Supervised Face Recognition

Jianwu Wan and Yi Wang, *Member, IEEE*

***Abstract*— In real-world applications, different kinds of learning and prediction errors are likely to incur different costs for the same system. Moreover, in practice, the cost label information is often available only for a few training samples. In a semi-supervised setting, label propagation is critical to infer the cost information for unlabeled training data. The existing methods typically conduct label propagation independently ahead of supervised cost-sensitive learning. The precomputed label information is kept fixed, which may become suboptimal in the subsequent learning process and hence degrade the overall system performance. In this paper, we develop a unified cost-sensitive framework for semi-supervised face recognition that can jointly optimize the inferred label information and the classifier in an iterative manner. Our experiments on face benchmark datasets demonstrate that in comparison with the state-of-the-art methods for label propagation and cost-sensitive learning, the proposed approach can significantly improve the overall system performance, especially in terms of classification errors associated with high costs.**

***Index Terms*— Cost-sensitive learning, semi-supervised learning, label propagation, face recognition.**

I. Introduction

COST-SENSITIVE learning takes into account the fact that in real-world applications different classification errors incur different penalties [1]. For face recognition systems, possible mistakes in making a classification decision include 1) false rejection that incorrectly identifies a gallery person as an imposter, 2) false acceptance that incorrectly identifies an imposter as a gallery person, and 3) false identification that incorrectly identifies a gallery person as another gallery person. In a door-locker scenario, for example, a false acceptance error that admits an imposter could result in a security breach far more severe than the cost of inconvenience that blocks a legitimate user due to a false rejection error. On the contrary, in a watch-list lookup scenario, any false rejection error, i.e., a criminal suspect on the watch list passes screening by mistake, could become a serious threat to public security.

In traditional face recognition systems, the three kinds of classification errors are assumed to have an equal cost in minimizing the overall loss of misclassification. However, as pointed out in [1], the de facto unequal costs are likely to affect the optimal decision threshold of a face learner. Lowering the total or average recognition error, as attained in conventional machine learning techniques, cannot solve the problem as long as the different types of errors exist and contribute unequally to the misclassification loss. Therefore, cost-sensitive learning methods were proposed to incorporate different misclassification costs into the loss function for different applications [2]–[15].

Most of existing cost-sensitive learning methods are supervised, which require labels that define the misclassification cost of each sample for training. Often in practice, however, only a few training samples can acquire the cost label information while the rest of the training data are *unlabeled*. This led to the development of cost-sensitive semi-supervised techniques to make use of both labeled and unlabeled data effectively. In particular, *label propagation* is critical in semi-supervised learning as it infers the label information for unlabeled training data.

State-of-the-art methods [4], [6], [8] conduct label propagation in advance, then followed by supervised cost-sensitive learning based on the given and inferred label information. Such a two-step pipeline approach has two limitations. Firstly, the label propagation step therein is cost *in*sensitive and therefore may result in wrong label estimates for the unlabeled data. Secondly, as the label propagation step is conducted independently of the subsequent step of cost-sensitive learning, the precomputed label information is kept fixed and may become suboptimal as the cost-sensitive model evolves. As a result, such cost-insensitive and unadaptive label propagation may impair classifier learning and degrade the overall system performance.

We address these issues by updating label propagation with the cost-sensitive model to minimize the overall misclassification loss. As a result, we develop a unified cost-sensitive framework in an iterative manner for extracting high-level face semantic features, conducting label propagation and learning the classifier simultaneously. In this paper, we consider the door-locker system for a cost-sensitive scenario to develop the

Manuscript received April 18, 2018; revised September 27, 2018; accepted November 19, 2018. Date of publication December 6, 2018; date of current version March 20, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61502058, Grant 61876038, Grant 61403324, and Grant 61572085, in part by the Natural Science Foundation of Educational Committee of Jiangsu Province under Grant 15KJB520002, and in part by the Dongguan University of Technology under Grant KCYKYQD2017003. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Karthik Nandakumar. *(Corresponding author: Yi Wang.)*

J. Wan is with the School of Information Science and Engineering, Changzhou University, Changzhou 213164, China (e-mail: jianwuwan@gmail.com).

Y. Wang is with the School of Computer Science and Network Security, Dongguan University of Technology, Dongguan 523808, China (e-mail: wangyi@dgut.edu.cn).

Digital Object Identifier 10.1109/TIFS.2018.2885252

cost-sensitive semi-supervised framework. It is worth noting that the proposed methodology can be conveniently extended to other applications of face recognition systems.

The main contributions of this paper are:

- We design cost-sensitive latent semantic regression to embed face images into the label space of training data. This enables a cost-sensitive process of label propagation that updates the estimated labels with the learned classifier information in an iterative manner.
- We introduce cost-sensitive regularization for guiding the label propagation process. Our experimental results show that imposing regularization on the labeled data improves the accuracy of label propagation and thus the system performance in terms of the total cost due to misclassification.
- We propose to jointly optimize learning of latent semantic features, cost-sensitive label propagation and the classifier all in a unified framework. We devise an algorithm for solving the unified framework by iteration. We show that the algorithm is iteratively descent and its computational complexity is linear with the size of the training dataset in each iteration of the training process.

The rest of this paper is organized as follows. Section II reviews the related work. Section III describes the problem formulation. Section IV provides details of the unified cost-sensitive framework. Experimental results are reported in Section V. Section VI draws the conclusion.

II. Related Work

Cost-sensitive classifiers were proposed to make an optimal decision by minimizing some loss function that incorporates different misclassification costs. The most representative work includes multi-class cost-sensitive kernel logistic regression (McKLR) and multi-class cost-sensitive k-nearest neighbor (McKNN) methods [2], cost-sensitive Laplacian support vector machine [15], cost-sensitive support vector machine [16], multi-category support vector machine [17], cost-sensitive decision trees with example-dependent misclassification costs [18]. In particular, sparse cost-sensitive classifier [9] and cost-sensitive sparse representation based classification (CS_SRC) [10] were proposed for the door-locker system based on face recognition. Recent work on cost-sensitive algorithms can also be found in [19]–[23]. These cost-sensitive classifiers all involve parts, such as dimensionality reduction or feature selection, that are cost *in*sensitive.

Lu *et al.* [3]–[5] considered that useful cost-sensitive information may be lost in the dimensionality reduction phase. To address the problem, they embedded misclassification costs into the cost-insensitive dimensionality reduction phase and proposed cost-sensitive principal component analysis (CSPCA), cost-sensitive linear discriminant analysis (CSLDA) and cost-sensitive locality preserving projections (CSLPP), respectively. Under similar motivation, cost-sensitive Laplacian score (CSLS) [12] and discriminative cost-sensitive Laplacian score (DCSLS) [7] were also proposed to embed cost information into the feature selection phase. Such cost-sensitive methods are mostly based on supervised learning. Using supervised information can lead to better performance, but in practice it is difficult to obtain labels for all the training data.

TABLE I
KEY NOTATIONS

Notation	Description
C	Cost matrix of size $c \times c$
X	Training data matrix of size $D \times N$
X_{L}	Labeled training data matrix of size $D \times N_l$
X_{U}	Unlabeled training data matrix of size $D \times (N - N_l)$
Y_{L}	Given label matrix of size $c \times N_l$ for X_{L}
F	Inferred cost-sensitive label matrix of size $c \times N$
W	Projection matrix of size $d \times c$ for cost-sensitive classifier
B	Learned latent semantic space of size $D \times d$
S	Latent semantic representations of size $d \times N$ for X

Semi-supervised learning makes use of both labeled and unlabeled data. For example, the two self-training methods proposed in [24] and [25] iteratively add the pseudo-class labels of unlabeled data to increase the labeled training data set. Two assumptions widely used in semi-supervised learning are the cluster assumption [26] and the manifold assumption. The former assumes that near neighbors have similar labels. The latter assumes that the high-dimensional data are distributed in a low-dimensional manifold structure. To obtain better performance, the objective function usually obeys the graph preserving criterion. The representative work includes the kNN graph [27]–[29] and L_1 graph [30]–[33] based approaches.

As far as we know, cost-sensitive semi-supervised discriminant analysis (CS^{3}DA) [4] is the first work proposed for cost-sensitive semi-supervised learning. CS^{3}DA first uses the sparsity learning technique [34] to infer the soft label of unlabeled data and then learns the projection by incorporating misclassification costs into linear discriminant analysis. Considering that the sparse representation is computationally inefficient and the sparsity assumption may not hold in applications like face recognition [35], Wan *et al.* [6] proposed a method named PCSDA, which adopts the L_2 norm approach to learn the label information of unlabeled data. Furthermore, to obtain more effective label information for unlabeled data, a soft L_2 norm approach was proposed in cost-sensitive semi-supervised canonical correlation analysis (CS^{3}CCA) [8]. As discussed in Section I, such a two-step pipeline approach has limitations that can affect the overall system performance.

III. Problem Formulation

Table I lists key notations used in this paper. In the door-locker system based on face recognition [2]–[10], suppose that there are in total c classes including $c-1$ classes of gallery subjects and the class c of imposters. Let C denote the cost matrix of size $c \times c$, where C_{ij} is the cost of misclassifying class i as class j and $C_{ij} = 0$ for $i = j$ by definition.

Let $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N] \in \mathbb{R}^{D \times N}$ contain N face image vectors each from the D dimensional space. The training set X is then divided into two parts, i.e., $X = [X_{\mathrm{L}}, X_{\mathrm{U}}]$

where $X_{\mathrm{L}} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N_l}]$ is the labeled subset and $X_{\mathrm{U}} = [\mathbf{x}_{N_l+1}, \mathbf{x}_{N_l+2}, \ldots, \mathbf{x}_N]$ is the unlabeled subset. Let Y_{L} be the given label matrix for X_{L}. For the labeled training data, the label information $y_{ji} = 1$ if sample $\mathbf{x}_i$ belongs to class j, where $j = 1, 2, \ldots, c$, otherwise $y_{ji} = 0$.

For the unlabeled training data, existing cost-sensitive semi-supervised methods such as [4], [6], and [8] have label information estimated through some label propagation function in the form of $L(X_{\mathrm{L}}, X_{\mathrm{U}}, Y_{\mathrm{L}})$ that is typically cost *in*sensitive.

In this paper, we propose a unified cost-sensitive framework for conducting label propagation and classifier learning simultaneously. Let $F = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_N]$ denote the inferred cost-sensitive label matrix, where $\mathbf{f}_i$ is a one-hot vector, i.e., only one of its c elements is one and all the others are zero. In our approach, each label vector $\mathbf{f}_i$ for $i = 1, 2, \ldots, N$ is estimated in a cost-sensitive way by regressing the current classification results. The joint optimization problem can be solved by minimizing some misclassification loss function in the general form of

$$\min_{W,F} \text{loss}\{\phi(X, W), F, C\} \tag{1}$$

where ϕ classifies the input training data X with the projection matrix W. Then, the classification results can be used to evaluate the label matrix F with the cost matrix C. The cost-sensitive label information is used in turn to update the classifier ϕ with respect to W. This process is iterated until the overall misclassification loss is minimized. In this way, both label propagation and classifier learning are embedded in a cost-sensitive framework.

To deal with face feature variations, we further propose to conduct cost-sensitive semi-supervised learning in some latent semantic space of face images. The last two key notations in Table I specify the robust high-level features used in our approach. In particular, $B \in \mathbb{R}^{D\times d}$ spans the learned latent semantic space and $S \in \mathbb{R}^{d\times N}$ accommodates the d-dimensional latent semantic representations of X.

IV. The Unified Cost-Sensitive Framework

In this section, we elaborate our unified cost-sensitive framework for semi-supervised face recognition. Section IV-A proposes cost-sensitive latent semantic regression for label propagation and learning of the classifier. Section IV-B introduces cost-sensitive regularization to guide the label propagation process. Section IV-C presents design of the misclassification loss function for cost-sensitive learning in the latent semantic space. Section IV-D describes the iterative algorithm for solving the unified framework. Section IV-E explains the procedure for inference.

A. Cost-Sensitive Learning in the Latent Semantic Space

Considering facial expressions, lighting and poses of face images taken at different times, it is necessary to extract robust feature representations for cost-sensitive face recognition. To address this issue, we adopt matrix factorization to extract high-level features that can reflect the inherent structure between data [34], [36]–[39]. The latent semantic space B and the high-level features S can be jointly learned from

$$L_1(B, S) = \|X - BS\|_F^2 \tag{2}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. We do not include any sparsity constraint in (2) for matrix factorization because face recognition is not commonly considered as a compressive sensing problem [6], [8], [35].

We then use a linear predictive classifier to project S into the label space, i.e., $\phi(X, W) = W^T S(X)$ where $S(X)$ is the latent semantic features learned from (2) with input X, and cast least square minimization for the loss function. Note that it is possible to consider other classifiers for ϕ and optimization rules. In our context, linear regression makes an update simpler in every iteration and yet can achieve effective results for the unified framework. Thus, we introduce *cost-sensitive latent semantic regression* as

$$L_2(W, S, F) = \sum_{i=1}^{N} h(i)\|W^T\mathbf{s}_i - \mathbf{f}_i\|_2^2 \tag{3}$$

where $\mathbf{s}_i$ denotes the latent semantic representation of sample $\mathbf{x}_i$ and $h(i)$, known as the importance function [3], [6]–[8], depicts the importance of sample $\mathbf{x}_i$ in the training process.

In supervised learning scenarios [3], [40], the importance function is often defined as the total cost of misclassifying sample $\mathbf{x}_i$ whose true class label is denoted by $l(\mathbf{x}_i)$. In our context of semi-supervised learning, sample $\mathbf{x}_i$ can be either labeled or unlabeled. Accordingly, the importance of sample $\mathbf{x}_i$ is evaluated as

$$h(i) = \begin{cases} \sum_{j=1}^{c} C_{l(\mathbf{x}_i)j}, & \text{if } i \le N_l \\ \tau, & \text{if } i > N_l \end{cases} \tag{4}$$

where the hyper-parameter τ is set for unlabeled training data and its value is found empirically to stress the importance of unlabeled data in cost-sensitive learning.

Proposition 1: Assume that $\mathbf{x}_i \in X$ for $i = 1, 2, \ldots, N$ are conditionally independent of each other given their label classes $l(\mathbf{x}_i) = 1, 2, \ldots, c$ whose densities are multivariate Gaussian's with a common covariance matrix. Given the label matrix $F = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_N]$, minimizing the least squares criterion in the form of $\min_W \|W^T S - F\|_2^2$ results in a solution $\hat{W} = [\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \ldots, \hat{\mathbf{w}}_c]$ that projects the latent semantic feature $\mathbf{s}_i$ of each sample $\mathbf{x}_i$ into the label space with regressed terms proportional to the posteriori class probabilities, i.e., $\hat{\mathbf{w}}_k^T\mathbf{s}_i \propto p(l(\mathbf{x}_i) = k|\mathbf{x}_i)$ for $k = 1, 2, \ldots, c$, and

$$\|\hat{W}^T\mathbf{s}_i - \mathbf{f}_i\|_2^2 \propto \sum_{j:j\le c, j\ne l(\mathbf{x}_i)} p(j|\mathbf{x}_i)^2 + [1 - p(l(\mathbf{x}_i)|\mathbf{x}_i)]^2. \tag{5}$$

Proof: Let $\mathbf{g}_k^T$ be a row vector in F containing one-hot vectors for label class $k = 1, 2, \ldots, c$ such that $g_{ki} = 1$ if $l(\mathbf{x}_i) = k$ and $g_{ki} = 0$ otherwise for all $i = 1, 2, \ldots, N$ in the training dataset. The least squares solution can also be obtained by solving

$$\min_{\mathbf{w}_k} \|S^T\mathbf{w}_k - \mathbf{g}_k\|_2^2 \tag{6}$$

for each label classifier individually [41].

Note that the problem expressed in (6) is two-class regression with class k and a *null* class that contains all samples that do *not* belong to class k, i.e., $l(\mathbf{x}_i) \neq k$. Suppose that the mean for the two classes are $\mathbf{m}_k$ and $\mathbf{m}_0$, respectively. Since all label classes have the same covariance matrix Σ, the least squares solution of two-class regression satisfies the following relationship [41]:

$$\hat{\mathbf{w}}_k \propto \Sigma^{-1}(\mathbf{m}_k - \mathbf{m}_0). \tag{7}$$

On the other hand, we may use a Gaussian Naive Bayes (GNB) classifier to estimate the posteriori class probability $p(l(\mathbf{x}_i) = k|\mathbf{x}_i)$, or equivalently $p(g_{ki} = 1|\mathbf{s}_i)$, for the above two-class problem with Gaussian densities. According to the Bayes rule, we have

$$p(g_{ki} = 1|\mathbf{s}_i) = \frac{1}{1 + \exp(a_k - [\Sigma^{-1}(\mathbf{m}_k - \mathbf{m}_0)]^T \mathbf{s}_i)} \tag{8}$$

with

$$a_k = \frac{1}{2}(\mathbf{m}_k + \mathbf{m}_0)^T \Sigma^{-1}(\mathbf{m}_k - \mathbf{m}_0) + \ln \frac{n_0}{n_k} \tag{9}$$

where n_k and n_0 are sizes of class k and the null class, respectively. It can be seen that (8) is linear in the latent semantic feature $\mathbf{s}_i$. Therefore, we have

$$[\Sigma^{-1}(\mathbf{m}_k - \mathbf{m}_0)]^T \mathbf{s}_i \propto p(l(\mathbf{x}_i) = k|\mathbf{x}_i). \tag{10}$$

Combining (7) and (10) yields

$$\hat{\mathbf{w}}_k^T \mathbf{s}_i \propto p(l(\mathbf{x}_i) = k|\mathbf{x}_i) \tag{11}$$

for $k = 1, 2, \ldots, c$.

Let $\mathbf{p}_i = [p(1|\mathbf{x}_i), p(2|\mathbf{x}_i), \ldots, p(c|\mathbf{x}_i)]^T$. Note from (11) that $\|\hat{W}^T \mathbf{s}_i - \mathbf{f}_i\|_2^2 \propto \|\mathbf{p}_i - \mathbf{f}_i\|_2^2$ and the residual-sum-of-squares (RSS) between $\mathbf{p}_i$ and $\mathbf{f}_i$ can be expressed as

$$\begin{aligned}
&\|\mathbf{p}_i - \mathbf{f}_i\|_2^2 \\
&= (\mathbf{p}_i - \mathbf{f}_i)^T (\mathbf{p}_i - \mathbf{f}_i) \\
&= \mathbf{p}_i^T \mathbf{p}_i - \mathbf{p}_i^T \mathbf{f}_i - \mathbf{f}_i^T \mathbf{p}_i + \mathbf{f}_i^T \mathbf{f}_i \\
&= \sum_{j:j\leq c, j\neq l(\mathbf{x}_i)} p(j|\mathbf{x}_i)^2 + p(l(\mathbf{x}_i)|\mathbf{x}_i)^2 - 2\ p(l(\mathbf{x}_i)|\mathbf{x}_i) + 1 \\
&= \sum_{j:j\leq c, j\neq l(\mathbf{x}_i)} p(j|\mathbf{x}_i)^2 + [1 - p(l(\mathbf{x}_i)|\mathbf{x}_i)]^2.
\end{aligned} \tag{12}$$

■

It can be seen that (12) contains two parts. The first part corresponds to Type I errors of misclassifying $\mathbf{x}_i$ as another class $j \neq l(\mathbf{x}_i)$ and the second part indicates Type II errors of misclassifying $\mathbf{x}_i$ as the null class. Thus, $\|\hat{W}^T \mathbf{s}_i - \mathbf{f}_i\|_2^2$ is proportional to both Type I and Type II errors. A small value of RSS indicates that $\mathbf{x}_i$ is likely to be regressed to the right class $l(\mathbf{x}_i)$, i.e., $p(l(\mathbf{x}_i)|\mathbf{x}_i)$ is close to 1 and $p(j|\mathbf{x}_i)$ is close to 0 for $j \leq c$ and $j \neq l(\mathbf{x}_i)$.

Proposition 1 states that the projection matrix $\hat{W}$ learned in the latent semantic space can minimize classification errors. In addition, we multiply the error probabilities with the associated costs as in (3). In this way, minimizing (3) minimizes the overall misclassification loss of all training samples, including both labeled and unlabeled ones, with respect to their individual importance in terms of the misclassification cost.

B. Cost-Sensitive Regularization for Label Propagation

In this section, we further improve the proposed scheme by introducing cost-sensitive regularization for label propagation. The main idea is to guide the semi-supervised learning process with supervised label information Y_{L}. Intuitively, we should keep estimated labels in F close to the corresponding ones in Y_{L} with the label propagation loss minimized.

Proposition 2: Let $\mathbf{y}_i \in Y_{\mathrm{L}}$ be the label vector of $\mathbf{x}_i \in X_{\mathrm{L}}$ with label class $l(\mathbf{x}_i)$ and $\mathbf{q}_i = [\sqrt{C_{l(\mathbf{x}_i)1}}, \ldots, \sqrt{C_{l(\mathbf{x}_i)c}}]^T$ contains the cost of classifying $\mathbf{x}_i$ as class j for $j = 1, 2, \ldots, c$. Following Proposition 1, the estimate $\hat{\mathbf{f}}_i = \hat{W}^T \mathbf{s}_i$ for $i = 1, 2, \ldots, N_l$ in the label space satisfies

$$\|(\hat{\mathbf{f}}_i - \mathbf{y}_i) \odot \mathbf{q}_i\|_2^2 \propto \sum_{j:j\leq c, j\neq l(\mathbf{x}_i)} p(j|\mathbf{x}_i)^2\ C_{l(\mathbf{x}_i)j} \tag{13}$$

where the symbol $\odot$ denotes element-wise multiplication.

Proof: The true label vector $\mathbf{y}_i$ has $y_{l(\mathbf{x}_i)i} = 1$ and $y_{ji} = 0$ for $j \neq l(\mathbf{x}_i)$. Noting $C_{l(\mathbf{x}_i)l(\mathbf{x}_i)} = 0$ and due to (11), we have

$$\begin{aligned}
\|(\hat{\mathbf{f}}_i - \mathbf{y}_i) \odot \mathbf{q}_i\|_2^2 &= \sum_{j=1}^{c} \left[(\hat{\mathbf{w}}_j^T \mathbf{s}_i - y_{ji})\sqrt{C_{l(\mathbf{x}_i)j}}\right]^2 \\
&= \sum_{j:j\leq c, j\neq l(\mathbf{x}_i)} \left(\hat{\mathbf{w}}_j^T \mathbf{s}_i\right)^2 C_{l(\mathbf{x}_i)j} \\
&\propto \sum_{j:j\leq c, j\neq l(\mathbf{x}_i)} p(j|\mathbf{x}_i)^2\ C_{l(\mathbf{x}_i)j}\ .
\end{aligned} \tag{14}$$

■

According to Proposition 2, we can reduce the label propagation loss on the labeled training data by adding the regularization term $\|(\mathbf{f}_i - \mathbf{y}_i) \odot \mathbf{q}_i\|_2^2$ in optimization together with cost-sensitive latent semantic regression. Consider that the labeled and unlabeled data have a similar structure if they come from the same class. The posterior class probabilities $p(j|\mathbf{x}_i)$, $j = 1, 2, \ldots, c$, learned from the labeled samples can also be applied to the unlabeled ones, thus minimizing the label propagation loss for the unlabeled data. The regularization term for label propagation is defined as

$$L_3(F) = \sum_{i=1}^{N} \|(\mathbf{f}_i - \mathbf{y}_i) \odot \mathbf{q}_i\|_2^2 \tag{15}$$

where the sample size is extended to N for incorporating $L_3(F)$ into the overall objective function. In particular, we define $\mathbf{y}_i = \mathbf{0}$ and $\mathbf{q}_i = \mathbf{0}$ for $i > N_l$ in (15), because label estimates of the unlabeled samples should not have influence on regularization. Note that the regularization term in (15) is cost sensitive due to the vector $\mathbf{q}_i$ as well as the fact that $\mathbf{f}_i$ is estimated by cost-sensitive latent semantic regression in (3).

C. Misclassification Loss Function

In our design of the misclassification loss function, we include the term $L_1(B, S)$ in (2) that performs matrix factorization for learning latent semantic representations, the term $L_2(W, S, F)$ in (3) that performs cost-sensitive label regression in the latent semantic space for updating the classifier and the label matrix, and the regularization term $L_3(F)$ in (15) for

Algorithm 1 Iterative Algorithm for Optimization

Input: Training set X, given label matrix Y_{L}, cost matrix C, hyper-parameters μ, γ, λ and τ.

Output: Latent semantic space B, projection matrix W.

1: Initialize $B^{[k]}$, $S^{[k]}$, $W^{[k]}$ and $F^{[k]}$ randomly with $k = 0$.
2: **repeat**
3: Update $B^{[k+1]}$ using (19) with $S = S^{[k]}$.
4: Update $\mathbf{s}_i^{[k+1]}$ for $i = 1, 2, \ldots, N$ using (21) with $B = B^{[k+1]}$, $W = W^{[k]}$ and $F = F^{[k]}$.
5: Update $W^{[k+1]}$ using (23) with $S = S^{[k+1]}$ and $F = F^{[k]}$.
6: Update $\mathbf{f}_i^{[k+1]}$ for $i = 1, 2, \ldots, N$ that minimizes (24) with $S = S^{[k+1]}$ and $W = W^{[k+1]}$.
7: Set $k = k + 1$.
8: **until** termination criterion reached.

supervising label propagation in a cost-sensitive way. In addition, to resist overfitting, we introduce another regularization term

$$R(B, S, W) = \|B\|_F^2 + \|S\|_F^2 + \|W\|_F^2. \tag{16}$$

The resulting loss function is of the form

$$\mathrm{loss}(B, S, W, F) = L_1(B, S) + \mu L_2(W, S, F) + \gamma L_3(F) + \lambda R(B, S, W) \tag{17}$$

where the hyper-parameters μ, γ and λ are used to trade off the corresponding terms. In particular, μ and λ affect updates of the learned classifier in the latent semantic space while μ and γ control the label propagation process. Note that it is possible to use separate parameters for controlling the three regularization terms in $R(B, S, W)$ though at the expense of complicating parameter selection. Section V-D.1 discusses parameter selection in more details.

D. Iterative Algorithm for Optimization

Here we present the iterative algorithm to learn the latent semantic space B and the projection matrix W for the unified framework with the aim of minimizing the misclassification loss function (17). Specifically, as shown in Algorithm 1, the algorithm updates one matrix variable at a time by fixing all the other variables in every step of an iteration. The updating steps of the algorithm in the kth iteration, $k = 0, 1, \ldots$, are described as follows:

1) *Update the latent semantic space $B^{[k+1]}$.* We note that, by fixing S, W and F, minimizing (17) with respect to B is equivalent to

$$\min_B \|X - BS\|_F^2 + \lambda \|B\|_F^2 \tag{18}$$

which is a quadratic minimization problem [42] and has a unique solution of the form

$$B = XS^T(SS^T + \lambda I)^{-1} \tag{19}$$

where $I \in \mathbb{R}^{d \times d}$ is the identity matrix. Thus, we update $B^{[k+1]}$ using (19) with $S = S^{[k]}$.

2) *Update the latent semantic representations $S^{[k+1]}$.* We update $\mathbf{s}_i^{[k+1]}$, $i = 1, 2, \ldots, N$, independently with each other. In this way, by fixing B, W and F, minimizing (17) with respect to $\mathbf{s}_i$ is equivalent to

$$\min_{\mathbf{s}_i} \|\mathbf{x}_i - B\mathbf{s}_i\|_2^2 + \mu h(i) \|W^T\mathbf{s}_i - \mathbf{f}_i\|_2^2 + \lambda \|\mathbf{s}_i\|_2^2 \tag{20}$$

which is a quadratic minimization problem [42] and has a unique solution of the form

$$\mathbf{s}_i = (B^T B + \mu h(i) W W^T + \lambda I)^{-1}(B^T \mathbf{x}_i + \mu h(i) W \mathbf{f}_i). \tag{21}$$

Thus, for $i = 1, 2, \ldots, N$, we update $\mathbf{s}_i^{[k+1]}$ using (21) with $B = B^{[k+1]}$, $W = W^{[k]}$ and $F = F^{[k]}$.

3) *Update the projection matrix $W^{[k+1]}$.* We note that, by fixing B, S and F, minimizing (17) with respect to W is equivalent to

$$\min_W \mu \sum_{i=1}^{N} h(i) \|W^T\mathbf{s}_i - \mathbf{f}_i\|_2^2 + \lambda \|W\|_F^2 \tag{22}$$

which is a quadratic minimization problem [42] and has a unique solution of the form

$$W = (\mu S H S^T + \lambda I)^{-1}(\mu S H F^T) \tag{23}$$

where $H = \mathrm{diag}(h(1), \ldots, h(N))$. Thus, we update $W^{[k+1]}$ using (23) with $S = S^{[k+1]}$ and $F = F^{[k]}$.

4) *Update the cost-sensitive label matrix $F^{[k+1]}$.* We update $\mathbf{f}_i^{[k+1]}$, $i = 1, 2, \ldots, N$, independently with each other. In this way, by fixing B, S and W, minimizing (17) with respect to $\mathbf{f}_i$ is equivalent to minimizing

$$\mu h(i) \|W^T\mathbf{s}_i - \mathbf{f}_i\|_2^2 + \gamma \|(\mathbf{f}_i - \mathbf{y}_i) \odot \mathbf{q}_i\|_2^2 \ . \tag{24}$$

Recall by definition that $\mathbf{f}_i$ is a one-hot vector. Thus, for $i = 1, 2, \ldots, N$, we update $\mathbf{f}_i^{[k+1]}$ by enumerating all the c possible solutions in this context and finding the one that minimizes (24) with $S = S^{[k+1]}$ and $W = W^{[k+1]}$.

The above steps are repeated until the algorithm reaches the termination criterion, i.e., $\mathrm{loss}(B^{[k]}, S^{[k]}, W^{[k]}, F^{[k]}) - \mathrm{loss}(B^{[k+1]}, S^{[k+1]}, W^{[k+1]}, F^{[k+1]}) \leq \epsilon$, where ϵ is an arbitrarily small value.

Remark 1: It can be observed that the proposed algorithm is *iteratively descent*. This is because each of the quadratic minimization problems derived in Steps 1) to 3) is convex and hence guarantees that the objective function is strictly decreasing. The minimization problems involved in Step 4, though not necessarily convex, ensure that the objective function is at least non-increasing. Accordingly, for all k, we have

$$\begin{aligned} &\mathrm{loss}(B^{[k]}, S^{[k]}, W^{[k]}, F^{[k]}) \\ &\quad > \mathrm{loss}(B^{[k+1]}, S^{[k]}, W^{[k]}, F^{[k]}) \\ &\quad > \mathrm{loss}(B^{[k+1]}, S^{[k+1]}, W^{[k]}, F^{[k]}) \\ &\quad > \mathrm{loss}(B^{[k+1]}, S^{[k+1]}, W^{[k+1]}, F^{[k]}) \\ &\quad \geq \mathrm{loss}(B^{[k+1]}, S^{[k+1]}, W^{[k+1]}, F^{[k+1]}) \ . \end{aligned} \tag{25}$$

Remark 2: In the proposed algorithm, the time complexity for updating $B^{[k+1]}$ is $O(N(Dd + d^2) + Dd + d^3)$.

The time complexity for updating each $\mathbf{s}_i^{[k+1]}$, $i = 1, 2, \ldots, N$, is $O(d(D + c) + d^2(D + c + 1) + d^3)$. As H in (23) is a diagonal matrix, the time complexity for updating $W^{[k+1]}$ is $O(N(2d + d^2 + dc) + d^2c + d^3)$. For updating each $\mathbf{f}_i^{[k+1]}$, $i = 1, 2, \ldots, N$, we enumerate its c possible solutions and the time complexity is $O(3c^2 + c)$. Therefore, the computational complexity of the proposed algorithm is linear with the size N of the training dataset in each iteration of the training process.

E. Inference for Face Recognition

At the inference stage, given a test sample $\mathbf{x}$, we extract its high-level feature representations $\mathbf{s_x}$ in the latent semantic space B learned from the training process. The high-level feature $\mathbf{s_x}$ is then projected into the c-dimensional label space using the learned classifier W via $W^T\mathbf{s_x}$. The class label k is assigned to $\mathbf{x}$ if the kth element of $W^T\mathbf{s_x}$ has the maximum value which corresponds to the minimum misclassification loss as shown in Proposition 1.

V. Experimental Results

In this section, we evaluate the performance of our proposed approach for semi-supervised face recognition in door-locker systems. The experiments are conducted on five public face datasets. The following describes main specifications of these benchmark datasets.

- Extended Yale B [43]: It contains 2,414 front-view face images of 38 individuals under different illuminations. The images are cropped to 32×32 pixels.
- AR [44]: We use a subset of the AR face dataset pre-processed by Martinez and Kak [45]. This subset contains 1,400 face images corresponding to 100 individuals, where each individual has 14 images taken with different expressions and illumination conditions. The resolution of face images is resized to 66×48 pixels.
- PIE [46]: It contains 41,368 images of 68 individuals with 13 different poses and 4 different expressions, under 43 different illumination conditions. In this paper, we choose five near frontal poses (C05, C07, C09, C29) with all different expressions and illuminations. This results in 170 images for each individual and the image size is 32×32 pixels.
- LFW [47]: This dataset is more challenging for studying unconstrained face recognition, which contains 13,233 face images of 5,749 subjects crawled from the web. We use the pre-aligned version LFW-a [48]. Similar to [49], we crop each face image to 50×50 pixels and gather the subjects that have no less than 30 samples per subject from LFW-a to use in our experiments.
- CASIA-WebFace [50]: It is a large-scale face dataset that contains 494,414 face images of 10,575 subjects collected from the Internet in a semi-automatic way. Similar to [51], we use MTCNN [52] to detect the faces and find the five face landmarks. Then, all the face images are converted to gray-scale images and normalized to 144×144 pixels via the landmarks. Inspired by [53], we further rotate the two eye points to be horizontal to overcome the pose variations. We gather the subjects that have no less than 200 samples per subject from CASIA-WebFace to be used in our experiments.

TABLE II
Experimental Settings

	c	G_{L}	G_{U}	I_{L}	I_{U}	$N_{\mathrm{G}}^{\mathrm{te}}$	$N_{\mathrm{I}}^{\mathrm{te}}$
Extended Yale B	31	3	7	24	56	1,602	432
AR	81	3	7	60	140	320	80
PIE	61	3	27	24	216	8,394	1,120
LFW-a	29	3	17	18	102	1,467	223
CASIA-WebFace	151	3	17	375	2,125	44,056	36,119

In all cases, face images are preprocessed with PCA for dimension reduction to speed up matrix factorization of the high-dimensional features.

For each of the five datasets, we randomly select $c - 1$ subjects as the gallery person and the rest as imposters. The training data is formed by randomly selecting $N_{\mathrm{G}}^{\mathrm{tr}}$ images for each gallery person containing G_{L} labeled and G_{U} unlabeled samples, and $N_{\mathrm{I}}^{\mathrm{tr}}$ images for the imposters containing I_{L} labeled and I_{U} unlabeled images. The remaining $N_{\mathrm{G}}^{\mathrm{te}}$ gallery images and $N_{\mathrm{I}}^{\mathrm{te}}$ imposter images are used for testing. The total number of training images and test images is $N_{\mathrm{tr}} = (c-1)N_{\mathrm{G}}^{\mathrm{tr}} + N_{\mathrm{I}}^{\mathrm{tr}}$ and $N_{\mathrm{te}} = N_{\mathrm{G}}^{\mathrm{te}} + N_{\mathrm{I}}^{\mathrm{te}}$, respectively. Table II specifies the settings for each of the five datasets used in our experiments.

We perform three-fold cross validation on the training data for choosing the values of the hyper-parameters μ, λ, γ and τ. Section V-D.1 discusses parameter selection in more details. As a result, we choose $\mu = 0.007$, $\gamma = 1$, $\lambda = 0.01$ and $\tau = 0.2$ for all datasets.

For ease of discussion, in our experiments, we assume

$$C_{ij} = \begin{cases} 0, & \text{if } i = j \\ C_{\mathrm{GG}}, & \text{if } i < c,\ j < c \text{ and } i \neq j \\ C_{\mathrm{GI}}, & \text{if } i < c \text{ and } j = c \\ C_{\mathrm{IG}}, & \text{if } i = c \text{ and } j < c \end{cases} \tag{26}$$

where C_{GG}, C_{GI} and C_{IG} are the cost of false identification, false rejection and false acceptance, respectively. In door-locker scenarios, it often considers $C_{\mathrm{IG}} > C_{\mathrm{GI}} > C_{\mathrm{GG}}$. That is, misrecognizing an imposter as a gallery person is more serious than misrecognizing a gallery person as an imposter, whereas the latter is more serious than misrecognizing a gallery person as another gallery person. In Section V-D.2, we discuss how the cost ratios $C_{\mathrm{IG}} : C_{\mathrm{GI}} : C_{\mathrm{GG}}$ may influence the system performance. If not specified elsewhere, we set $C_{\mathrm{IG}} : C_{\mathrm{GI}} : C_{\mathrm{GG}} = 20 : 2 : 1$ as in [2]–[10].

Given the cost matrix C in the form of (26) and based on the definition of the importance function $h(i)$ in (4), we have

$$h(i) = \begin{cases} (c-2)C_{\mathrm{GG}} + C_{\mathrm{GI}}, & \text{if } i \leq N_l \text{ and } l(\mathbf{x}_i) < c \\ (c-1)C_{\mathrm{IG}}, & \text{if } i \leq N_l \text{ and } l(\mathbf{x}_i) = c \\ \tau, & \text{if } i > N_l \end{cases} \tag{27}$$

for labeled gallery subjects, labeled imposters and unlabeled training data, respectively. Note that, in semi-supervised

TABLE III
CONTRIBUTION OF EACH TERM IN THE LOSS FUNCTION WITH REGARD TO *Total Cost* (*Accuracy*)

	Extended Yale B	AR	PIE	LFW-a	CASIA-WebFace
Remove the $L_1(B,S)$ term from (17)	3,134 (72.4%)	89 (96.1%)	15,796 (53.0%)	2,874 (89.4%)	192,440 (76.0%)
Remove the $L_2(W,S,F)$ term from (17)	3,637 (67.9%)	944 (55.5%)	22,830 (30.6%)	3,730 (45.5%)	466,150 (43.9%)
Remove the $L_3(F)$ term from (17)	3,051 (81.8%)	64 (99.3%)	12,635 (97.3%)	2,870 (90.6%)	162,375 (92.1%)
Remove the $R(B,S,W)$ term from (17)	2,727 (78.0%)	63 (97.9%)	8,544 (80.8%)	2,653 (76.5%)	120,085 (82.0%)
The proposed approach with all terms in (17)	1,879 (94.0%)	49 (99.4%)	8,311 (98.2%)	2,552 (98.0%)	83,383 (99.7%)

learning, it is not uncommon to assume that the unlabeled data are less important than their labeled counterparts in the training process. Thus, we have $\tau < (c-2)C_{\mathrm{GG}} + C_{\mathrm{GI}} < (c-1)C_{\mathrm{IG}}$ for $C_{\mathrm{GG}} < C_{\mathrm{GI}} < C_{\mathrm{IG}}$. For convenience, we normalize $h(i)$ by $(c-2)C_{\mathrm{GG}} + C_{\mathrm{GI}}$ and the resulting importance function is of the simplified form

$$h(i) = \begin{cases} 1, & \text{if } i \le N_l \text{ and } l(\mathbf{x}_i) < c \\ \dfrac{(c-1)C_{\mathrm{IG}}}{(c-2)C_{\mathrm{GG}} + C_{\mathrm{GI}}}, & \text{if } i \le N_l \text{ and } l(\mathbf{x}_i) = c \\ \tau \in [0,1], & \text{if } i > N_l \ . \end{cases} \quad (28)$$

For performance evaluation, we adopt five widely-used metrics in cost-sensitive learning [2]–[5], [9], [10], i.e., *total cost*, Err_{IG} (error rate of false acceptance), Err_{GI} (error rate of false rejection), Err_{GG} (error rate of false identification), and Err (total error rate). We also introduce the metric *accuracy* to evaluate the effectiveness of cost-sensitive approaches in imposter detection. In our context, the metrics are defined as

$$\begin{cases} total\ cost = C_{\mathrm{IG}} \times N_{\mathrm{fa}} + C_{\mathrm{GI}} \times N_{\mathrm{fr}} + C_{\mathrm{GG}} \times N_{\mathrm{fi}} \\ Err_{\mathrm{IG}} = N_{\mathrm{fa}} / N_{\mathrm{I}}^{\mathrm{te}} \times 100\% \\ Err_{\mathrm{GI}} = N_{\mathrm{fr}} / N_{\mathrm{G}}^{\mathrm{te}} \times 100\% \\ Err_{\mathrm{GG}} = N_{\mathrm{fi}} / N_{\mathrm{G}}^{\mathrm{te}} \times 100\% \\ Err = (N_{\mathrm{fa}} + N_{\mathrm{fr}} + N_{\mathrm{fi}}) / N_{\mathrm{te}} \times 100\% \\ accuracy = 1 - Err_{\mathrm{IG}} \end{cases} \quad (29)$$

where N_{fa}, N_{fr} and N_{fi} denote the number of false acceptance, false rejection and false identification, respectively.

A. Contribution of Each Term in the Loss Function

Recall that the loss function (17) of our proposed unified framework contains four terms, i.e., $L_1(B,S)$, $L_2(W,S,F)$, $L_3(F)$, and $R(B,S,W)$. Here we evaluate the contribution of each component by removing it from (17). The corresponding results in terms of *total cost* and *accuracy* are reported in Table III for each of the five benchmark datasets.

It is clear that the most critical term is $L_2(W,S,F)$ which performs cost-sensitive label regression in the latent semantic space. According to Proposition 1, the projection matrix W learned from $L_2(W,S,F)$ helps to minimize classification errors. In fact, by removing the $L_2(W,S,F)$ term from (17), the proposed approach is reduced to vanilla matrix factorization on PCA features. Without cost-sensitive label propagation, the related terms involving W and F are no longer included in the reduced loss function for regularization. The results have *total cost* increased by more than 17 times on the AR dataset and *accuracy* reduced by almost 70% on the PIE dataset.

The next significant term is $L_1(B,S)$ for extracting high-level feature representations in the latent semantic space. Removing the $L_1(B,S)$ term from (17) results in *total cost* increased by 130% on the CASIA-WebFace dataset and *accuracy* reduced by 46% on the PIE dataset. This demonstrates the effectiveness of learning latent semantic representations on top of raw image features to deal with face variations.

The remaining two regularization terms also help to improve the performance of the proposed approach. In particular, $L_3(F)$ is for supervising label propagation. According to Proposition 2, the $L_3(F)$ term can reduce the label propagation loss and thus improve the overall system performance. This is supported by the results in Table III where we see that, by removing the $L_3(F)$ term from (17), *total cost* is almost doubled on the CASIA-WebFace dataset. The other regularization term $R(B,S,W)$ is to resist overfitting. The results in Table III show that removing $R(B,S,W)$ from (17) can reduce *accuracy* by 22% on the LFW-a dataset.

B. Comparison With Other Cost-Sensitive Learning Methods

We demonstrate the effectiveness of the proposed unified cost-sensitive framework by comparing it with the following *nine* cost-sensitive learning methods:

- *Cost-sensitive classifiers* including McKLR [2], McKNN [2] and CS_SRC [10].
- *Cost-sensitive feature selection* including CSLS [12] and DCSLS [7].
- *Supervised cost-sensitive dimensionality reduction* such as CSLDA [3].
- *Semi-supervised cost-sensitive dimensionality reduction* including CS^{3}DA [4], PCSDA [6] and CS^{3}CCA [8].

In particular, McKLR, McKNN, CS_SRC, CSLS, DCSLS and CSLDA are supervised and trained on *labeled* samples only, while the remaining methods including the proposed one are semi-supervised and trained on *all* samples. The cost-sensitive methods of feature selection and dimensionality reduction are evaluated with a kNN classifier with $k = 3$.

The results are presented in Table IV for comparison on *total cost* and in Table V for comparison on *accuracy*. We observe that semi-supervised methods outperform supervised methods in most cases. This suggests that unlabeled data can provide useful additional information for training. Among the ten comparing methods, the proposed unified framework performs the best over all five benchmark datasets.

For a more rigorous study of the experimental results, we also conduct a statistical analysis of the difference between the proposed method and each of the other

TABLE IV

COMPARISON OF COST-SENSITIVE LEARNING METHODS ON *Total Cost*

	Supervised Methods						Semi-Supervised Methods			
	McKLR	McKNN	CS_SRC	CSLS	DCSLS	CSLDA	CS^{3}DA	PCSDA	CS^{3}CCA	Proposed
Extended Yale B	2,650	4,242	5,538	6,191	5,142	2,161	3,202	1,954	2,149	1,879
AR	324	685	553	898	875	156	475	55	155	49
PIE	12,139	18,971	20,091	22,709	21,271	9,118	14,458	9,102	10,393	8,311
LFW-a	2,819	3,153	4,467	4,123	3,847	2,813	2,896	2,849	3,020	2,552
CASIA-WebFace	103,034	105,403	316,206	399,044	383,260	234,775	203,468	91,482	188,722	83,383

TABLE V

COMPARISON OF COST-SENSITIVE LEARNING METHODS ON *Accuracy*

	Supervised Methods						Semi-Supervised Methods			
	McKLR	McKNN	CS_SRC	CSLS	DCSLS	CSLDA	CS^{3}DA	PCSDA	CS^{3}CCA	Proposed
Extended Yale B	89.8%	81.6%	57.4%	44.9%	57.3%	88.8%	88.5%	91.7%	92.1%	94.0%
AR	98.3%	86.3%	77.4%	58.0%	59.0%	98.4%	98.2%	99.3%	99.3%	99.4%
PIE	83.8%	70.7%	46.7%	31.0%	36.0%	90.2%	98.0%	90.2%	93.3%	98.2%
LFW-a	84.6%	87.1%	32.5%	40.2%	49.9%	76.8%	97.0%	82.5%	81.4%	98.0%
CASIA-WebFace	97.7%	97.4%	68.1%	54.1%	56.4%	78.3%	82.9%	99.5%	87.6%	99.7%

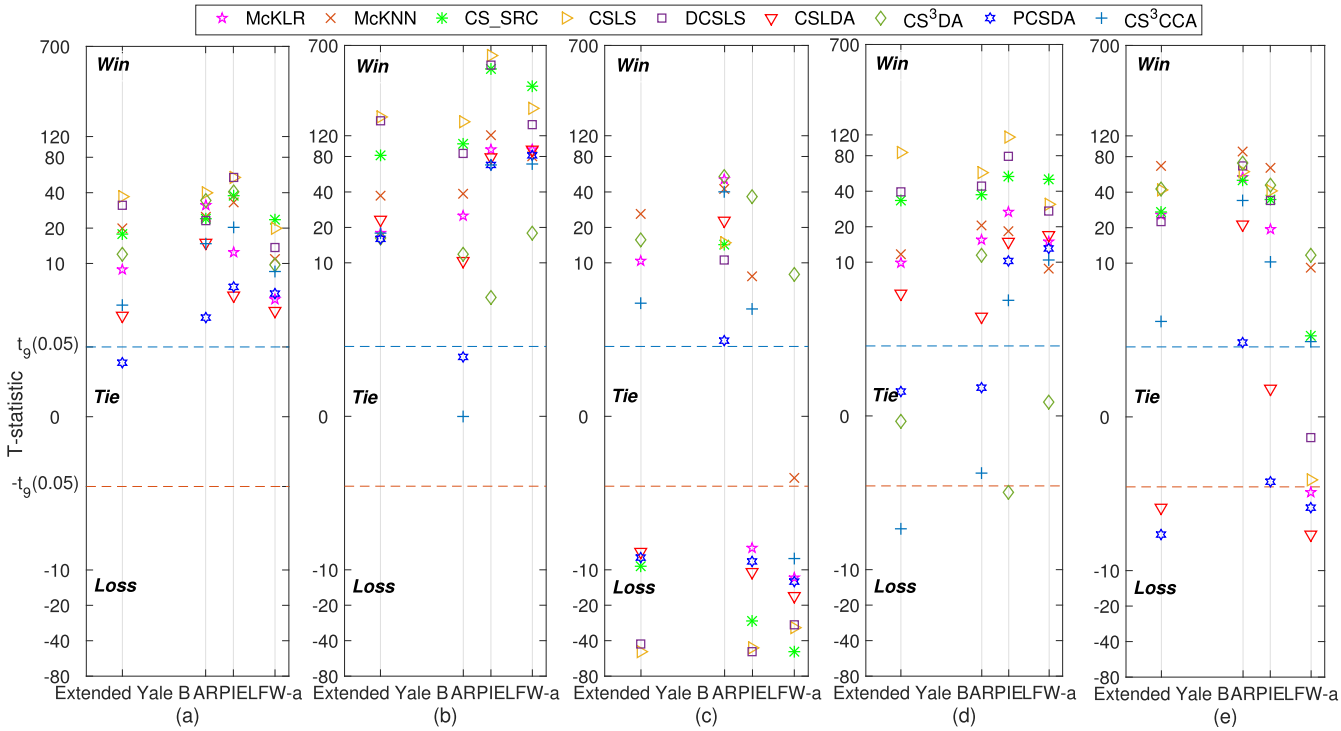


Fig. 1. Paired t-tests comparing the proposed method with the other methods on: (a) *total cost*. (b) Err_{IG}. (c) Err_{GI}. (d) Err_{GG}. (e) *Err*.

comparing methods. Specifically, let $E_{\mathrm{p}}^{(i)}$ and $E_{\mathrm{o}}^{(i)}$ denote the result of the proposed approach and that of the other comparing method, respectively, in the ith independent run of the experiment, and let $Z_i = E_{\mathrm{o}}^{(i)} - E_{\mathrm{p}}^{(i)}$, $i = 1, 2, \ldots, n$. Let $\mu_{E_{\mathrm{o}}}$ and $\mu_{E_{\mathrm{p}}}$ denote the expected result of the proposed method and that of the other comparing method. We are interested in testing if the difference between $\mu_{E_{\mathrm{o}}}$ and $\mu_{E_{\mathrm{p}}}$ is statistically significant with the null hypothesis $H_0 : \mu_{E_{\mathrm{o}}} = \mu_{E_{\mathrm{p}}}$ and the alternatives $H_1 : \mu_{E_{\mathrm{o}}} > \mu_{E_{\mathrm{p}}}$; $H_2 : \mu_{E_{\mathrm{o}}} < \mu_{E_{\mathrm{p}}}$. The test statistic that we use to make an inference decision is the T-statistic defined as [54]:

$$T = \frac{\bar{Z} - \mu_Z}{s_Z / \sqrt{n}} \tag{30}$$

where $\bar{Z}$ and s_Z are the sample mean and standard deviation of $\{Z_i\}$, and μ_Z is the hypothesized difference between $\mu_{E_{\mathrm{o}}}$ and $\mu_{E_{\mathrm{p}}}$. Under the null hypothesis, i.e., H_0, we have $\mu_Z = 0$ and the test statistic follows a t-distribution with $n - 1$ degrees of freedom for paired samples [54]. This allows us to determine the rejection region on T for a test at a significance level α. Accordingly, we decide that the proposed approach yields:

Win if $T > t_{n-1}(\alpha)$ by rejecting H_0 for H_1 at level α;
Loss if $T < -t_{n-1}(\alpha)$ by rejecting H_0 for H_2 at level α;
Tie otherwise, i.e., $|T| \leq t_{n-1}(\alpha)$.

Figure 1 plots the paired t-test results with $\alpha = 0.05$ and $n = 10$. Note that, for better visualization, the scale of all

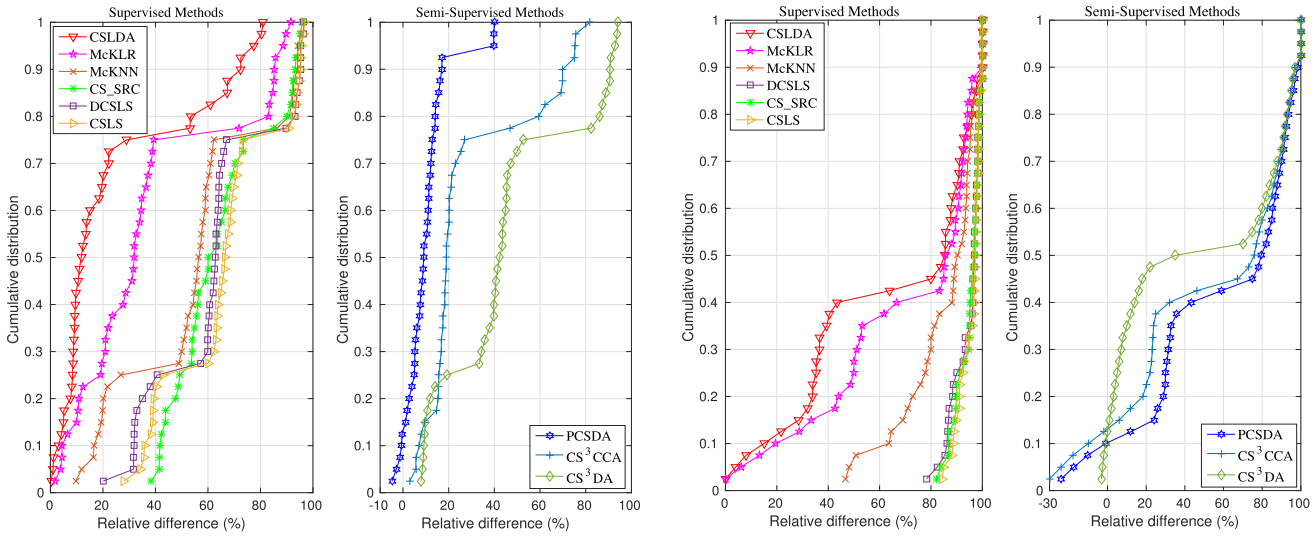


Fig. 2. Cumulative distributions of relative difference on: (a) *total cost*. (b) Err_{IG}.

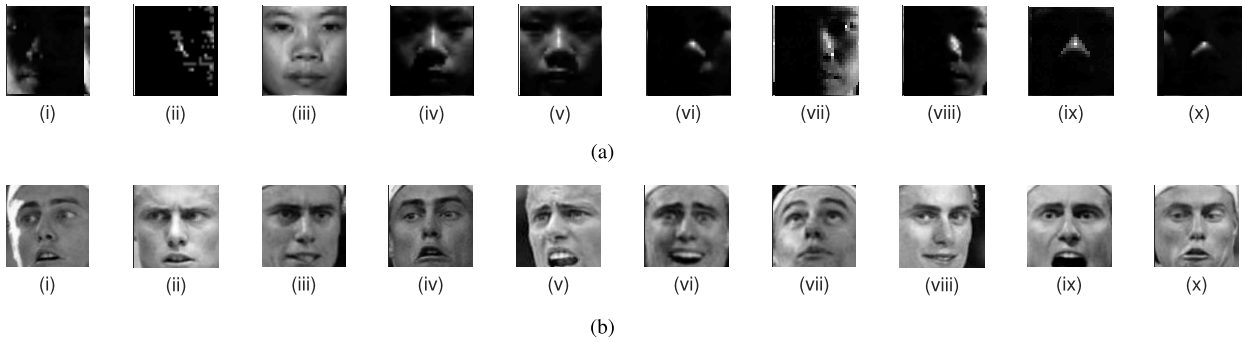


Fig. 3. Success and failure cases in imposter detection. (a) Example in the Extended Yale B dataset. (b) Example in the LFW-a dataset.

vertical axes is changed by taking the inverse hyperbolic sine function of the T values. We observe that the proposed method wins or ties in the majority of cases, especially on *total cost* and Err_{IG}. The loss is mainly on Err_{GI}, which is less critical in door-locker systems. For Err_{GG} and Err, the proposed method is comparable to the best performing method.

In Fig. 2, we further demonstrate the gain on *total cost* and Err_{IG} achievable by the proposed method. We present the results in the form of cumulative distributions of *relative difference*. Specifically, given $E_{\mathrm{p}}^{(i)}$ and $E_{\mathrm{o}}^{(i)}$ of the i-th run of the experiment on a dataset, their relative difference is defined as $Z_i/E_{\mathrm{o}}^{(i)}$. As expected, the proposed method, being a semi-supervised one, outperforms the supervised methods. It can be seen that, in nearly 75% of the cases, the gain in *total cost* achieved by the proposed method is more than 8%, 19%, 26%, 40%, 42% and 49% over CSLDA, McKLR, McKNN, DCSLS, CSLS and CS_SRC, respectively, while the gain in Err_{IG} is at least 35%. The median reduction is up to 66% in terms of *total cost* and up to 97% in terms of Err_{IG}. Comparing with the three semi-supervised methods, we observe that the proposed method outperforms CS^{3}CCA and CS^{3}DA in all cases in terms of *total cost*. Although in 12% of the cases the proposed method yields a larger *total cost* of up to 5% than that of PCSDA, the gain is up to 40% in 88% of the cases. In 12% of the cases, the proposed method also yields a larger Err_{IG} of up to 30% and 24% than that of CS^{3}CCA and PCSDA, respectively. However, the gain is up to 100% in 88% of the cases with a median reduction of nearly 80%.

In Fig. 3, we present success and failure cases of the proposed method in imposter detection. In particular, we consider two examples, one taken from the Extended Yale B dataset and the other taken from the LFW-a dataset. For both examples, images (i)-(iii) are training samples labeled as imposters, while images (iv)-(x) are test samples of the same subject. Among the test samples, (iv)-(v) in Fig. 3(a) and (iv) in Fig. 3(b) are falsely accepted by the proposed method but PCSDA and CS^{3}CCA. In contrast, (vii)-(x) in both examples are correctly identified by the proposed method but are falsely accepted by PCSDA and CS^{3}CCA. Observing these success and failure cases, it seems that PCSDA and CS^{3}CCA are better in identifying face images visually closer to the training samples, owing to their use of the Fisher criterion

TABLE VI
COMPARING LABEL PROPAGATION METHODS BY COST-SENSITIVE CORRELATION COEFFICIENT Δ BENCHMARKED AGAINST THE UPPER BOUND Δ^*

	Δ^*	kNN (k=3)	MF	Soft SR	L_2	Soft L_2	LPCR	L_1 graph	MASC	SODA	Proposed
Extended Yale B	4.86	2.00	2.14	1.78	3.49	2.95	3.15	2.93	2.51	2.79	4.30
AR	4.75	2.56	2.58	1.23	4.03	4.01	4.12	3.32	2.85	3.23	4.61
PIE	3.19	0.87	0.93	0.56	1.87	1.84	1.70	1.30	1.26	1.25	2.27
LFW-a	4.23	1.55	1.63	0.83	2.05	2.01	1.42	2.40	2.44	2.48	3.41

directly established at the pixel level. In contrast, since the proposed method is conducted in the latent semantic space which learns high-level feature representations, it is more robust in identifying face images with semantic variations of illumination, pose and facial expression. However, none of the three methods is able to identify (vi) in Fig. 3(a) and (v)-(vi) in Fig. 3(b) that deviate too much at both the pixel level and the semantic level.

C. Comparison With Other Label Propagation Methods

Next, we compare the proposed cost-sensitive label propagation method with the following *nine* state-of-the-art label propagation methods:

- kNN classifies an unlabeled sample by checking the labels of its k nearest neighbors in the supervised training set (k is set to 3).
- MF extracts high-level features of unlabeled data by matrix factorization first, and then propagates their labels with the kNN classifier (k is set to 3).
- Soft SR [4] obtains the soft label information of unlabeled data according to its sparse representation.
- L_2 [6] estimates the label information of unlabeled data by an L_2 norm approach.
- Soft L_2 [8] estimates the soft label information of unlabeled data by a soft L_2 norm approach.
- LPCR [55] uses the unlabeled training samples as the dictionary to reconstruct labeled training data. The label information of unlabeled data is then linearly propagated with reconstruction coefficients.
- L_1 graph [33] infers the label information of unlabeled data with the L_1 graph preserving criterion.
- MASC [27] predicts unlabeled data by using the smoothness criterion on the manifold.
- SODA [28] propagates the label information from labeled data to unlabeled data according to the distribution of labeled and unlabeled data.

To evaluate the accuracy of label inference, we define a measure of cost-sensitive correlation coefficient as

$$\Delta = \frac{1}{(N - N_l)} \sum_{i=N_l+1}^{N} \tilde{h}(i) \langle \mathbf{y}_i, \mathbf{f}_i \rangle \tag{31}$$

where $\mathbf{y}_i$ and $\mathbf{f}_i$ denote the true and estimated label vectors of the unlabeled sample $\mathbf{x}_i$, $\langle , \rangle$ is the dot product, and

$$\tilde{h}(i) = \begin{cases} 1, & \text{if } l(\mathbf{x}_i) < c \\ \dfrac{(c-1)C_{\text{IG}}}{(c-2)C_{\text{GG}} + C_{\text{GI}}}, & \text{if } l(\mathbf{x}_i) = c \end{cases} \tag{32}$$

indicates the importance of sample $\mathbf{x}_i$ in the learning process. Intuitively, a larger value of Δ indicates a higher expectation of cost-sensitive correlation between $\mathbf{f}_i$ and $\mathbf{y}_i$. In the ideal case where $\mathbf{f}_i$ is identical to $\mathbf{y}_i$, (31) reduces to

$$\Delta^* = \frac{1}{(N - N_l)} \sum_{i=N_l+1}^{N} \tilde{h}(i) \tag{33}$$

which represents the *upper bound* that can be achieved by a label propagation method. Table VI reports the Δ values of the various label propagation methods benchmarked against the upper bound Δ^*. In all cases, it is clear that the Δ value of the proposed method is the closest to Δ^*. The results confirm that the proposed method infers labels more accurately than the comparing methods.

D. Influential Factors

In this section, we study the influence of various factors in the proposed framework, including the choice of hyper-parameters, cost ratios, the number of gallery subjects, and the number of labeled training samples per class.

1) Choice of Hyper-Parameters: We choose the values of hyper-parameters μ, λ, γ in (17) and τ in (28) empirically from three-fold cross validation performed on the training dataset. As shown in Section IV-D, μ and λ affect updates of the latent semantic representations S and the projection matrix W in the training process. Figure 4 shows the effect of μ and λ on *total cost* with the value of μ chosen from the set $\{0.001, 0.005, 0.01, 0.015, 0.02, 0.025, 0.03\}$ and the value of λ chosen from the set $\{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}\}$. It can be seen that the minimum of *total cost* is obtained when $\mu \in [0.005, 0.015]$ and $\lambda \in [10^{-3}, 10^{-1}]$.

The parameter γ controls the regularization term for label propagation, which may be tuned with respect to μ in updating the cost-sensitive label matrix F. Figure 5 shows the effect of γ on *total cost* where we fix $\mu = 0.007$ and vary γ from $10^{-2}\mu$ to $10^{4}\mu$. It is clear that regularization improves the performance of label propagation. We also observe that there tends to be a performance degradation when $\gamma < 10^{0}\mu$ or when $\gamma > 10^{3}\mu$. Thus, a proper value of γ may be chosen from the range $[10^{0}\mu, 10^{3}\mu]$.

The parameter τ sets the importance of unlabeled data in cost-sensitive learning. In the extreme case where $\tau = 0$, we do not consider any cost for misclassifying an unlabeled sample while training. On the other hand, when $\tau = 1$, the unlabeled data and their estimated labels are treated equally as the labeled ones. Figure 6 shows the effect

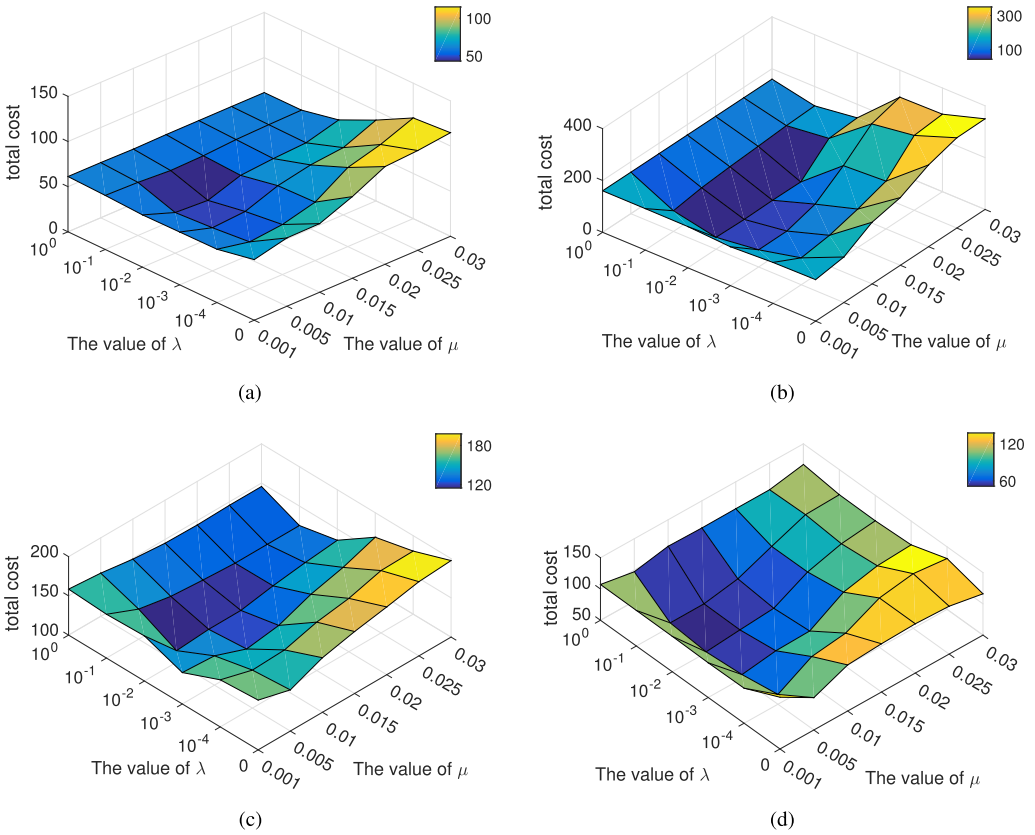


Fig. 4. Influence of the parameters μ and λ on *total cost*. (a) Extended Yale B. (b) AR. (c) PIE. (d) LFW-a.

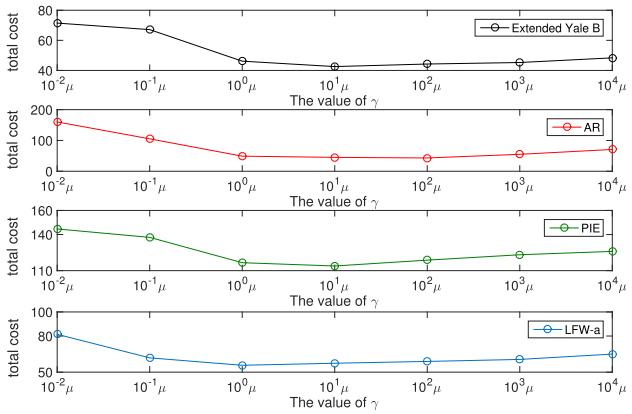


Fig. 5. Influence of the parameter γ on *total cost*.

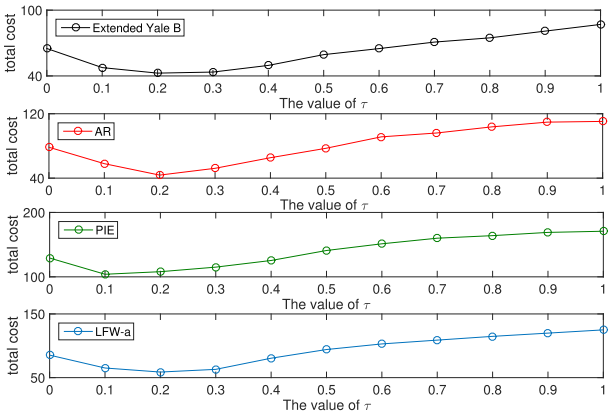


Fig. 6. Influence of the parameter τ on *total cost*.

of τ on *total cost* with τ varied from 0 to 1 at a step of 0.1. An appropriate value of τ may be selected from the range $[0.1, 0.3]$.

2) Cost Ratios: It is clear in (28) that normalizing both the numerator and the denominator by C_{GG} has no effect on the importance function $h(i)$. It can also be seen from (13), (15) and (17) that normalizing the objective function of $L_3(F)$ by C_{GG} and adjusting the parameter γ accordingly do not change the optimization result. Thus, in our design, the learning results depend on the ratios $C_{\mathrm{IG}} : C_{\mathrm{GI}} : C_{\mathrm{GG}}$ rather than the absolute values of the cost items.

Figure 7 shows the effect of cost ratios on *total cost* by fixing $C_{\mathrm{GI}} : C_{\mathrm{GG}} = 2 : 1$ and varying $C_{\mathrm{IG}} : C_{\mathrm{GI}}$ from 2.5 to 15 at a step of 2.5. We see that, when $C_{\mathrm{IG}} : C_{\mathrm{GI}}$ increases, CSLS, DCSLS and CS_SRC increase *total cost* at a significantly higher rate than the other methods. In all cases, the proposed approach is among the best performing methods.

Figure 8 plots the ROC curves based on Err_{IG} and Err_{GI} as a result of the varying cost ratios configured in Fig. 7. In most of the cases, we observe a clear trade-off between false acceptance and false rejection. That is, when $C_{\mathrm{IG}} : C_{\mathrm{GI}}$ increases, Err_{IG} decreases but Err_{GI} increases. Note that

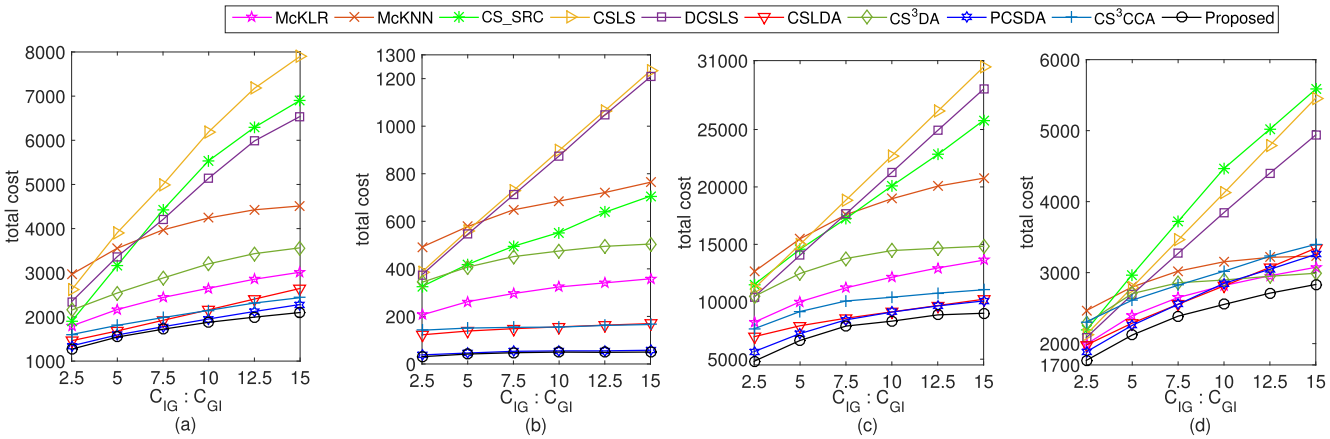


Fig. 7. Influence of the ratio $C_{\mathrm{IG}} : C_{\mathrm{GI}}$ on *total cost*. (a) Extended Yale B. (b) AR. (c) PIE. (d) LFW-a.

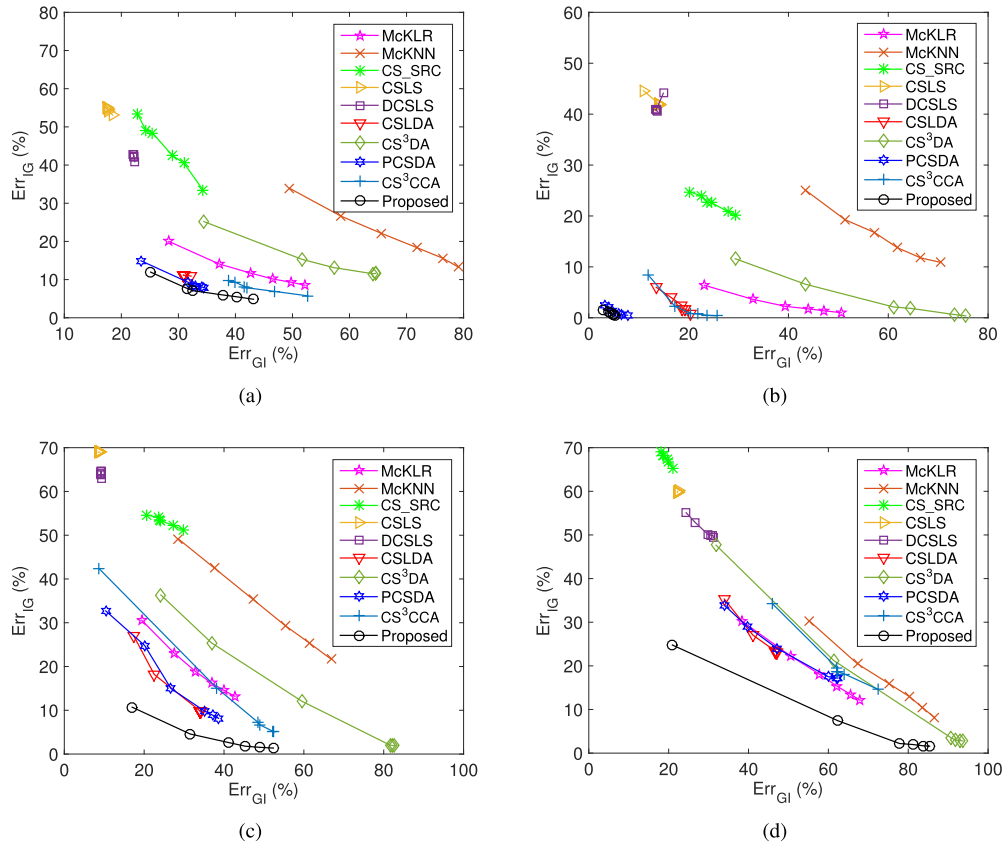


Fig. 8. ROC curves based on Err_{IG} vs. Err_{GI} influenced by the ratio $C_{\mathrm{IG}} : C_{\mathrm{GI}}$. (a) Extended Yale B. (b) AR. (c) PIE. (d) LFW-a.

some curves such as those of CSLS and DCSLS are short and nearly vertical in all cases, which indicates that *variation* of the cost ratio does not have much effect on the error rates of such cost-sensitive methods. On the other hand, the proposed approach not only keeps Err_{IG} to the minimum in cases where $C_{\mathrm{IG}} : C_{\mathrm{GI}}$ is small, but also reduces the high-cost error even further in response to the change of system requirements as expected. This observation supports our analysis in Proposition 1 where the cost-sensitive latent semantic regression error is shown proportional to the probability of misclassifying a training sample including both Type I and Type II errors. This relationship enables the unified framework to minimize the overall misclassification loss of all training samples, including both labeled and unlabeled ones, with respect to their importance played in cost-sensitive learning.

3) Number of Gallery Subjects: We investigate how the system performs by changing the training data distribution between gallery and imposter classes. In doing so, we vary the number of gallery persons from small to large until gallery classes become a majority in the training datasets. Specifically, we vary the number of gallery persons for training from 16 to 36 at a step of 4 for Extended Yale B, from 40 to 90 at a step of 10 for AR, from 38 to 63 at a step of 5 for PIE, and from 12 to 32 at a step of 4 for LFW-a. The results of *total cost* and Err_{IG} are presented in Fig. 9 and Fig. 10, respectively, where we observe that the proposed approach outperforms the nine comparing methods in all cases.

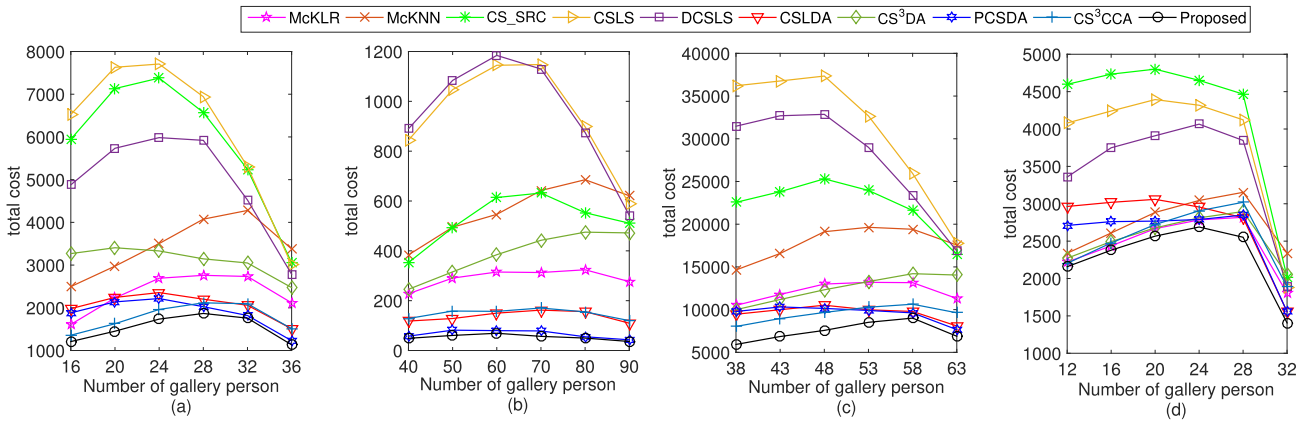


Fig. 9. Influence of the number of gallery subjects on *total cost*. (a) Extended Yale B. (b) AR. (c) PIE. (d) LFW-a.

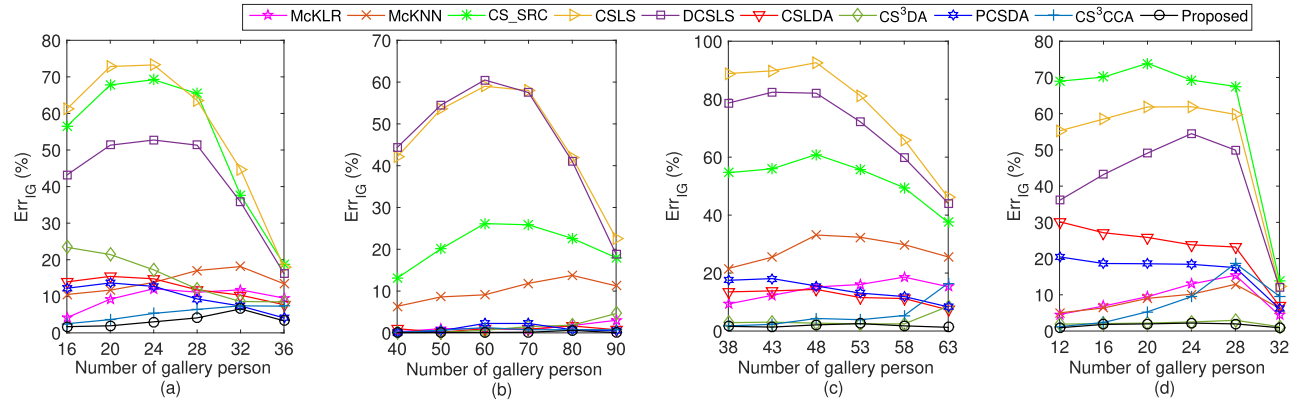


Fig. 10. Influence of the number of gallery subjects on Err_{IG}. (a) Extended Yale B. (b) AR. (c) PIE. (d) LFW-a.

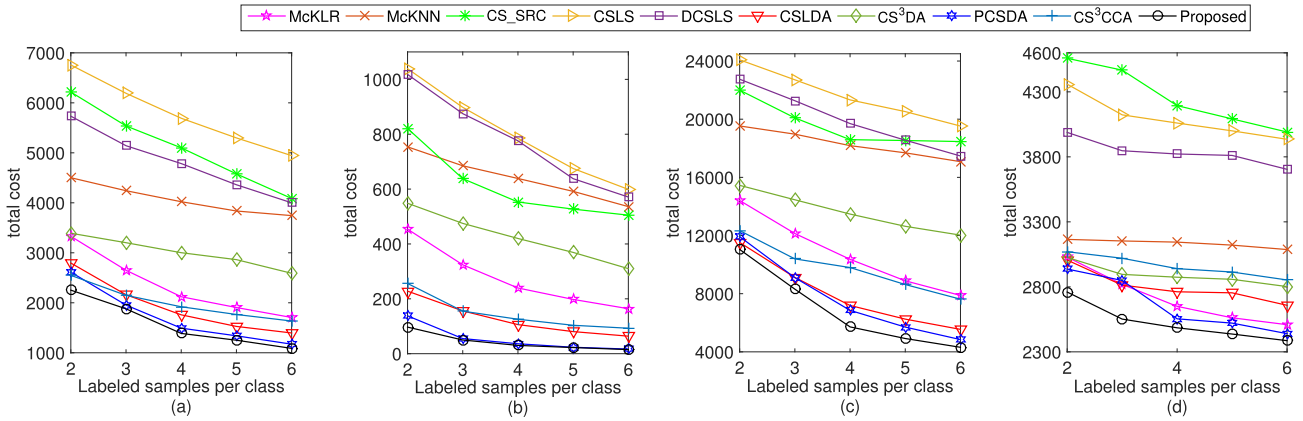


Fig. 11. Influence of the number of labeled training samples per class on *total cost*. (a) Extended Yale B. (b) AR. (c) PIE. (d) LFW-a.

We notice also in both Fig. 9 and Fig. 10 that the results do not change monotonically with an increasing number of gallery persons. This effect can be explained as follows. In cases where there are significantly more imposters than gallery subjects in the training dataset, the resulting system is more likely to misclassify a gallery subject as an imposter. This leads to a larger Err_{GI} and a smaller Err_{IG}. Thus, *total cost* is reduced as Err_{IG} is associated with a higher cost in door-locker systems. On the other hand, Err_{IG} may also become smaller when there are less instances of imposters causing false acceptance. Compared with the other cost-sensitive learning methods, the proposed approach demonstrates less variation in Err_{IG} when the data distribution between gallery and imposter classes changes.

4) Number of Labeled Training Samples per Class: In Fig. 11, we vary the number of labeled training samples per class from 2 to 6 and evaluate its influence on *total cost*. As expected, the performance is improved by including more labeled data for cost-sensitive learning. It can be seen that, compared with the other methods, the proposed approach may require less labeled data in order to reduce *total cost*.

E. Computational Cost

All methods considered in this paper are implemented on a machine with 3.5 GHz CPU and 32 GB RAM under the same experimental settings. Table VII reports the average training and test time. In general, the supervised methods such

TABLE VII
COMPARISON OF COMPUTATIONAL COST (IN SECONDS)

		Supervised Methods						Semi-Supervised Methods			
		McKLR	McKNN	CS_SRC	CSLS	DCSLS	CSLDA	CS^{3}DA	PCSDA	CS^{3}CCA	Proposed
Extended Yale B	Training	1.540	0.004	15	0.112	0.247	0.220	2.210	0.530	1.360	4.192
	Test	0.002	0.360	102	0.205	0.077	0.013	0.014	0.014	0.030	0.008
AR	Training	25	0.024	130	1.133	1.598	1.420	25	6.480	22	14
	Test	0.001	0.340	15	0.393	0.302	0.011	0.008	0.011	0.021	0.002
PIE	Training	8.761	0.012	58	0.317	0.388	0.761	20	7.050	18	26
	Test	0.021	6.248	381	2.212	0.987	0.112	0.145	0.131	0.341	0.037
LFW-a	Training	1.114	0.004	11	0.309	0.316	0.125	5.713	0.687	2.312	7.550
	Test	0.002	0.261	65	0.545	0.113	0.012	0.009	0.014	0.035	0.007
CASIA-WebFace	Training	656	0.131	6,585	84	108	4.691	2,312	413	1,578	141
	Test	0.670	360	14,775	577	58	8.470	7.480	3.652	15	0.348

as McKNN, CSLS, DCSLS and CSLDA are more efficient than the semi-supervised methods (including the proposed approach) at the training stage. However, it is worth noting that training is usually done offline. Thus, the test time is more of a concern in practice. The results in Table VII show that the proposed approach is comparable to the best performing methods in terms of the test time.

Note also that the high-level features of test data can be solved simultaneously by matrix factorization in our approach. Thus, the time complexity can become linear with the number of classes at the inference stage. In this regard, the proposed approach can conduct a test more efficiently than those using kNN classifiers, such as CSLS, DCSLS, CSLDA, CS^{3}DA, PCSDA and CS^{3}CCA, whose time complexity depends on the size of the labeled training data set that is usually much larger than the number of classes in a dataset.

VI. CONCLUSION

In this paper, we proposed to incorporate label propagation and classifier learning in a unified cost-sensitive framework for semi-supervised face recognition in the application scenario of door-locker systems. In particular, we showed that cost-sensitive learning in the latent semantic space is able to minimize classification errors of all training samples, including both labeled and unlabeled ones, with respect to their misclassification costs. We also showed both analytically and experimentally that the cost-sensitive regularization term introduced in the proposed approach is able to reduce the label propagation loss and thus improve the classifier performance. The proposed approach jointly updates high-level semantic features, label propagation and the semi-supervised classifier in the unified framework which is optimized by an iterative descent algorithm. As a result, the proposed approach significantly improves the system performance in comparison with state-of-the-art cost-sensitive learning methods. In future work, we shall consider using explicit structures of unlabeled training data for more robust learning, e.g., by adding constraints in the misclassification loss function that force nearest neighbors to have similar labels. We shall also consider extending label propagation on large datasets that may contain wrong labels affecting both training and test processes.

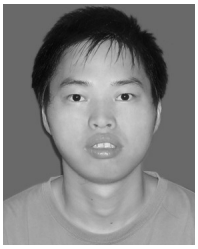
ACKNOWLEDGMENT

The authors wish to thank the anonymous reviewers and the editor for their valuable comments and suggestions that contributed to the improved quality of this paper.

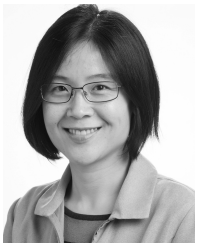
REFERENCES

[1] C. Elkan, "The foundations of cost-sensitive learning," in *Proc. Int. Joint Conf. Artif. Intell.*, Seattle, WA, USA, Aug. 2001, pp. 973–978.
[2] Y. Zhang and Z. H. Zhou, "Cost-sensitive face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 10, pp. 1758–1769, Oct. 2010.
[3] J. Lu and Y.-P. Tan, "Cost-sensitive subspace learning for face recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 2661–2666.
[4] J. Lu, X. Zhou, Y. P. Tan, Y. Shang, and J. Zhou, "Cost-sensitive semi-supervised discriminant analysis for face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 3, pp. 944–953, Jun. 2012.
[5] J. Lu and Y.-P. Tan, "Cost-sensitive subspace analysis and extensions for face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 3, pp. 510–519, Mar. 2013.
[6] J. Wan, M. Yang, Y. Gao, and Y. Chen, "Pairwise costs in semisupervised discriminant analysis for face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 10, pp. 1569–1580, Oct. 2014.
[7] J. Wan, M. Yang, and Y. Chen, "Discriminative cost sensitive Laplacian score for face recognition," *Neurocomputing*, vol. 152, pp. 333–344, Mar. 2015.
[8] J. Wan, H. Wang, and M. Yang, "Cost sensitive semi-supervised canonical correlation analysis for multi-view dimensionality reduction," *Neural Process. Lett.*, vol. 45, no. 2, pp. 411–430, Apr. 2017.
[9] J. Man, X. Jing, D. Zhang, and C. Lan, "Sparse cost-sensitive classifier with application to face recognition," in *Proc. 18th IEEE Int. Conf. Image Process.*, Brussels, Belgium, Sep. 2011, pp. 1773–1776.
[10] G. Zhang, H. Sun, Z. Ji, Y.-H. Yuan, and Q. Sun, "Cost-sensitive dictionary learning for face recognition," *Pattern Recognit.*, vol. 60, pp. 613–629, Dec. 2016.
[11] Y. Li, J. T.-Y. Kwok, and Z.-H. Zhou, "Cost-sensitive semi-supervised support vector machine," in *Proc. AAAI Nat. Conf. Artif. Intell.*, Atlanta, GA, USA, Jul. 2010, pp. 500–505.
[12] L. Miao, M. Liu, and D. Zhang, "Cost-sensitive feature selection with application in software defect prediction," in *Proc. 21st Int. Conf. Pattern Recognit.*, Tsukuba, Japan, Nov. 2012, pp. 967–970.
[13] M. Liu, L. Miao, and D. Zhang, "Two-stage cost-sensitive learning for software defect prediction," *IEEE Trans. Rel.*, vol. 63, no. 2, pp. 676–686, Jun. 2014.
[14] F. Wu, X.-Y. Jing, X. Dong, J. Cao, B. Xu, and S. Ying, "Cost-sensitive local collaborative representation for software defect prediction," in *Proc. Int. Conf. Softw. Anal., Test. Evol.*, Kunming, China, Dec. 2016, pp. 102–107.
[15] J.-S. Wu and Z.-H. Zhou, "Sequence-based prediction of microRNA-binding residues in proteins using cost-sensitive Laplacian support vector machines," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 10, no. 3, pp. 752–759, May 2013.

[16] U. Brefeld, P. Geibel, and F. Wysotzki, "Support vector machines with example dependent costs," in *Proc. 14th Eur. Conf. Mach. Learn.*, Cavtat, Croatia, Sep. 2003, pp. 23–34.
[17] Y. Lee, Y. Lin, and G. Wahba, "Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data," *J. Amer. Statist. Assoc.*, vol. 99, no. 465, pp. 67–81, 2004.
[18] O. M. Aodha and G. J. Brostow, "Revisiting example dependent cost-sensitive learning with decision trees," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 193–200.
[19] J. L. Wang, P. Zhao, and S. C. H. Hoi, "Cost-sensitive online classification," in *Proc. IEEE 12th Int. Conf. Data Mining*, Brussles, Belgium, Dec. 2012, pp. 1140–1145.
[20] H.-Y. Lo, S.-D. Lin, and H.-M. Wang, "Generalized k-Labelsets ensemble for multi-label and cost-sensitive classification," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 7, pp. 1679–1691, Jul. 2014.
[21] H.-Y. Lo, J.-C. Wang, H.-M. Wang, and S.-D. Lin, "Cost-sensitive multi-label learning for audio tag annotation and retrieval," *IEEE Trans. Multimedia*, vol. 13, no. 3, pp. 518–529, Jun. 2011.
[22] G. Zhang, H. Sun, G. Xia, and Q. Sun, "Multiple kernel sparse representation-based orthogonal discriminative projection and its cost-sensitive extension," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4271–4285, Sep. 2016.
[23] B. Gu, V. S. Sheng, K. Y. Tay, W. Romano, and S. Li, "Cross validation through two-dimensional solution surface for cost-sensitive SVM," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1103–1121, Jun. 2017.
[24] F. Roli and G. L. Marcialis, "Semi-supervised PCA-based face recognition using self-training," in *Proc. Joint IAPR Int. Workshops Stat. Techn. Pattern Recognit. Structural Syntactic Pattern Recognit.*, Hong Kong, Aug. 2006, pp. 560–568.
[25] X. Zhao, N. Evans, and J.-L. Dugelay, "Semi-supervised face recognition with LDA self-training," in *Proc. 18th IEEE Int. Conf. Image Process.*, Brussels, Belgium, Sep. 2011, pp. 3041–3044.
[26] P. Pham, T. Tuytelaars, and M.-F. Moens, "Naming people in news videos with label propagation," *IEEE MultiMedia*, vol. 18, no. 3, pp. 44–55, Mar. 2011.
[27] E. Kokiopoulou and P. Frossard, "Video face recognition with graph-based semi-supervised learning," in *Proc. IEEE Int. Conf. Multimedia Expo*, New York, NY, USA, Aug. 2009, pp. 1564–1565.
[28] F. Nie, S. Xiang, Y. Jia, and C. Zhang, "Semi-supervised orthogonal discriminant analysis via label propagation," *Pattern Recognit.*, vol. 42, no. 11, pp. 2615–2627, 2009.
[29] F. Nie, S. M. Xiang, Y. Liu, and C. Zhang, "A general graph-based semi-supervised learning with novel class discovery," *Neural Comput. Appl.*, vol. 19, no. 4, pp. 549–555, 2010.
[30] F. P. Nie, H. Wang, H. Huang, and C. Ding, "Unsupervised and semi-supervised learning via ℓ_1-norm graph," in *Proc. IEEE Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 2268–2273.
[31] F. Nie, D. Xu, I. W. Tsang, and C. Zhang, "Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction," *IEEE Trans. Image Process.*, vol. 19, no. 7, pp. 1921–1932, Jul. 2010.
[32] V. Kumar, A. M. Namboodiri, and C. V. Jawahar, "Face recognition in videos by label propagation," in *Proc. IEEE 22nd Int. Conf. Pattern Recognit.*, Stockholm, Sweden, Aug. 2014, pp. 303–308.
[33] S. Yan and H. Wang, "Semi-supervised learning by sparse representation," in *Proc. SIAM Int. Conf. Data Mining*, Sparks, NV, USA, May 2009, pp. 792–801.
[34] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.
[35] Q. Shi, A. Eriksson, A. van den Hengel, and C. Shen, "Is face recognition really a compressive sensing problem?" in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, Jun. 2011, pp. 553–560.
[36] G. Ding, Y. Guo, and J. Zhou, "Collective matrix factorization hashing for multimodal data," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 2083–2090.
[37] J. Tang, K. Wang, and L. Shao, "Supervised matrix factorization hashing for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3157–3166, Jul. 2016.
[38] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Amer. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.
[39] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, Denver, CO, USA, Nov. 2000, pp. 556–562.
[40] K. M. Ting, "An instance-weighting method to induce cost-sensitive trees," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 3, pp. 659–665, May 2002.
[41] T. Hastie, P. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer-Verlag, 2001.
[42] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[43] A. S. Georghiades, P. N. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.
[44] A. M. Martinez and R. Benavente, "The AR face database," Comput. Vis. Center, Bellatera, Spain, CVC Tech. Rep. 24, Jun. 1998.
[45] A. M. Martínez and A. C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 228–233, Feb. 2001.
[46] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cogn. Neurosci.*, vol. 31, no. 1, pp. 71–86, Jan. 1991.
[47] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, and G. Hua, "Labeled faces in the wild: A survey," in *Advances in Face Detection and Facial Image Analysis*, M. Kawulok, M. Celebi, and B. Smolka, Eds. Cham, Switzerland: Springer, 2016, pp. 189–248.
[48] L. Wolf, T. Hassner, and Y. Taigman, "Effective unconstrained face recognition by combining multiple descriptors and learned background statistics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 1978–1990, Oct. 2011.
[49] X. Fontaine, R. Achanta, and S. Süsstrunk, "Face recognition in real-world images," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, New Orleans, LA, USA, Mar. 2017, pp. 1482–1486.
[50] D. Yi, Z. Lei, S. C. Liao, and S. Z. Li. (Nov. 2014). "Learning face representation from scratch." [Online]. Available: https://arxiv.org/abs/1411.7923
[51] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 6738–6746.
[52] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[53] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2884–2896, Nov. 2018.
[54] J. A. Rice, *Mathematical Statistics and Data Analysis*, 3rd ed. Duxbury, CA, USA: Duxbury, 2006.
[55] G. Zhang, H. Sun, Z. Ji, and Q. Sun, "Label propagation based on collaborative representation for face recognition," *Neurocomputing*, vol. 171, pp. 1193–1204, Jan. 2016.



Jianwu Wan received the B.S., M.S., and Ph.D. degrees from the School of Computer Science and Technology, Nanjing Normal University, in 2008, 2010, and 2013, respectively. In 2013, he joined the School of Information Science and Engineering, Changzhou University. From 2015 to 2016, he was a Post-Doctoral Research Fellow with the Department of Computer Science, Hong Kong Baptist University. His research interests include machine learning and pattern recognition.



Yi Wang (S'05–M'09) received the Ph.D. degree in computer science from RMIT University, Melbourne, Australia, in 2009. She was a Research Associate with the School of Mathematics and Statistics, The University of New South Wales, Sydney, Australia, from 2009 to 2012, and a Research Assistant Professor with the Department of Computer Science, Hong Kong Baptist University, from 2012 to 2016. In 2017, she joined the School of Computer Science and Network Security, Dongguan University of Technology, China. Her research interests include biometrics, pattern recognition, and privacy-aware computing.