This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Evolutionary Cost-Sensitive Discriminative Learning With Application to Vision and Olfaction

Lei Zhang, *Member, IEEE*, and David Zhang, *Fellow, IEEE*

***Abstract*— In the design of machine learning models, one often assumes the same loss, which, however, may not hold in cost-sensitive learning scenarios. In a face-recognition-based access control system, misclassifying a stranger as a house owner and allowing entry may result in a more serious financial loss than misclassifying a house owner as a stranger and not allowing entry. That is, different types of recognition mistakes may lead to different losses, and therefore should be treated carefully. It is expected that a cost-sensitive learning mechanism can reduce the total loss when given a cost matrix that quantifies how severe one type of mistake is against another one. However, in many realistic applications, the cost matrix is unknown and unclear to users. Motivated by these concerns, in this paper, we propose an evolutionary cost-sensitive discriminative learning (ECSDL) method, with the following merits: 1) it addresses the definition of cost matrix in cost-sensitive learning without human intervention; 2) an evolutionary backtracking search algorithm is derived for the *NP-hard* cost matrix optimization; and 3) a cost-sensitive discriminative subspace is found, where the between-class separability and within-class compactness are well achieved, such that recognition becomes easier. Experiments in a variety of cost-sensitive vision and olfaction classification tasks demonstrate the efficiency and effectiveness of the proposed ECSDL approach.**

***Index Terms*— Classification, cost-sensitive learning, evolutionary algorithm (EA), subspace learning.**

I. Introduction

SUBSPACE learning is a hotspot in machine learning community for dimension reduction/low-dimensional feature representation. With effective subspace representation, the classification effectiveness and efficiency in pattern recognition tasks can be improved greatly. Several classical subspace learning methods include principal component analysis (PCA) [1], linear discriminant analysis (LDA) [2], locality preservation projection (LPP) [3], and marginal Fisher analysis (MFA) [4]. Their weighted, kernelized, and tensorized variants [5]–[9] have been proposed for compression and recognition in different fields. However, these existing subspace learning methods tend to achieve the lowest error rate by assuming the same loss for any misclassification, which, however, may not hold in many applications. For instance, in the face-recognition recognition-based access control system (FR-ACS) application, different mistakes may lead to different losses. It is natural to imagine that there would be a serious financial loss if the FR-ACS system misclassifies a stranger as the owner of a house and allowed to enter the room. Instead, misclassifying the owner of a house as a stranger and not allowed to enter may be less serious. A number of cost-sensitive recognition tasks happen in our real life, such as face recognition [10], palm recognition [11], odor recognition [12], [13], and object recognition [14]. If we do not treat the different loss problem carefully, the learned subspace would be biased and the feature representation is degraded. Dealing with different losses was first paid attention in [15], in which a cost-sensitive face recognition framework was formulated.

Recently, the cost-sensitive variants of the four classical subspace learning methods, such as CSPCA, CSLDA, CSLPP, and CSMFA, have also been surveyed for face recognition in [16] and [17], in which a cost matrix was predefined in advance under the prototypes of PCA, LDA, LPP, and MFA. These cost-sensitive subspace learning methods were proved to be effective in reducing the misclassification loss by predefining a cost matrix that quantifies how severe one type of mistake is against another one. However, in many realistic cases, the cost matrix is unknown and difficult to be manually defined by users [15], and therefore, the learned low-dimensional representation is biased and leads to poor generalization performance in real-world classification. It is worth noting that the misclassification loss is caused by incorrectly classifying one sample of the ith class into the jth class. Although users may know what type of mistake is more serious than another type, it is still difficult to give a specific cost value of one mistake. Therefore, it is difficult to accurately set the cost matrix via human intervention. The first attempt to address the problem of cost matrix definition can be referred to as [18], in which a cost interval (CI) (e.g., a possible cost range) was introduced instead of a precise cost value. However, it induced a high computational complexity and also the CI should still be manually predefined via human intervention. Therefore, the cost matrix definition problem is an open topic to be resolved in cost-sensitive learning subject. With this motivation, we could imagine that automatic learning of the cost matrix is extremely desired for a number of cost-sensitive recognition tasks. Additionally, for classification-oriented tasks, learning a discriminative subspace for low-dimensional representation with better between-class separability and within-class

Manuscript received October 18, 2016; accepted October 21, 2016. This work was supported in part by the National Natural Science Foundation of China under Grant 61401048 and in part by the Research Fund of Central Universities. The Associate Editor coordinating the review process was Dr. Jochen Lang. *(Corresponding author: Lei Zhang.)*

L. Zhang is with the College of Communication Engineering, Chongqing University, Chongqing 400044, China (e-mail: leizhang@cqu.edu.cn).

D. Zhang is with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong (e-mail: csdzhang@comp.polyu.edu.hk).

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/TIM.2016.2631878

compactness is the essential objective of this paper. That is, this paper targets at constructing a cost-sensitive discriminative subspace learning framework. In summary, the motivations are two-fold: 1) the cost-sensitive nature of subspace learning is very necessary in cost-sensitive classification tasks, and therefore, a discriminative subspace learning method is proposed by incorporating the cost-sensitive concept and 2) the efficient definition of the cost matrix is still an open issue in this community. Therefore, we propose an evolutionary idea for efficiently solving the NP-hard cost matrix optimization. With the two motivations, an evolutionary cost-sensitive discriminative learning (ECSDL) model is proposed in cost-sensitive vision and olfaction application scenarios.

More recently, sparse and low-rank subspace learning methods were also proposed for clustering and representation [36], [37], which tend to reveal the local and global structures implied in the data using sparse and low-rank constraints. In this paper, we would like to learn a discriminative subspace projection, and therefore, these sparse- and low-rank-based subspace clustering methods are not further discussed. When referred to as machine learning, it is necessary to briefly introduce the hottest deep learning techniques, which have manifested a very competitive performance in computer vision and natural language processing. Earlier, convolutional neural network (CNN) has won the highest performance in document and face recognition [38], [39]. Convolutional-based deep nets have also achieved the highest accuracy in face verification on the Faces in the Wild (LFW) data set [40], [41]. Due to the very strong deep feature representation ability, CNNs with convolutional layers, pooling layers, and fully connected layers have also been investigated and shown state-of-the-art performance in vision applications, such as the ImageNet and Pascal VOC [42], [43]. Although these deep learning methods show state-of-the-art performance, they rely on big data and a very high computational demand.

In this paper, inspired by the open issue of cost-sensitive learning [44] and subspace learning, we propose an ECSDL framework for handling real-world cost-sensitive recognition tasks. The concepts of cost-sensitive learning and discriminative subspace learning have been integrated into the ECSDL method. The contributions of this paper are threefold.

1) In ECSDL, a cost-sensitive discriminative subspace learning model is formulated with cost matrix optimization, within-class compactness, and between-class separability.
2) In ECSDL, an efficient evolutionary optimization method based on a backtracking search algorithm is derived for solving the *NP-hard* cost matrix optimization.
3) A number of cost-sensitive recognition tasks in vision (e.g., human attractiveness analysis, face recognition, and face verification) and olfaction (e.g., odor recognition) have been implemented using our ECSDL method.

The remainder of this paper is organized as follows. Section II presents the related work of subspace learning methods. The proposed ECSDL method and optimization algorithm are formulated in Section III. Experiments on multimodal beauty data for human attractiveness prediction are employed in Section IV. Experiments on AR and LFW face data for face recognition and face verification are conducted in Section V. Experiments on E-NOSE data for odor recognition in machine olfaction are presented in Section VI. The computational complexity and time of the proposed methods are discussed in Section VII. The visualization analysis of the cost matrix is shown in Section VIII. A detailed discussion is presented in Section IX. Finally, Section X concludes this paper.

II. Related Work

In this paper, we present four types of classical subspace learning and dimension reduction techniques, such as PCA [1], LDA [2], local preservation projection (LPP) [3], and MFA [4]. Furthermore, their cost-sensitive variants [16] are briefly discussed.

PCA, as a linear technique for unsupervised dimensionality reduction, constructs a low-dimensional representation that describes as much variance in the data as possible. PCA attempts to find a linear mapping $\mathbf{P}$ that maximizes the cost function of $\mathrm{Tr}(\mathbf{P}^T \mathrm{cov}(\mathbf{X})\mathbf{P})$, where $\mathrm{cov}(\mathbf{X})$ denotes the covariance matrix of the data $\mathbf{X}$. It is clear that the eigendecomposition of $\mathrm{cov}(\mathbf{X})$ can be solved, and the eigenvectors with respect to the first d eigenvalues formulate the linear mapping matrix $\mathbf{P}$.

Different from the unsupervised PCA, LDA is a supervised linear technique that attempts to find a mapping $\mathbf{P}$, such that the ratio of the between-class scatter matrix and the within-class scatter matrix is maximized. In this way, the between-class separability and the within-class compactness in the embedded subspace can be achieved. Therefore, the subspace of LDA is recognized to be "discriminative."

LPP, as an unsupervised subspace learning technique, is constructed based on the graph manifold learning theory. The local preservation projection implies that there is a low-dimensional manifold embedding in the high-dimensional data space, and the local affinity data structure can be preserved by constructing a Laplacian matrix $\mathbf{L}$. That is, the local affinity structure of the data in high-dimensional space is preserved in the low-dimensional subspace. Therefore, LPP attempts to find a mapping $\mathbf{P}$ by minimizing $\mathbf{P}^T\mathbf{XLX}^T\mathbf{P}$. However, the label information of the samples is not well exploited and the "discriminative" characteristic is missed.

MFA, as a supervised and linear subspace learning technique, is an effective combination of the "discrimination" in LDA and the "local structure preservation" in LPP. Specifically, two locality graphs with respect to interclass (between-classes) samples and intraclass (within-classes) samples are constructed. As a result, two Laplacian matrices, $\mathbf{L}_{\text{intraclass}}$ and $\mathbf{L}_{\text{interclass}}$, were calculated. Essentially, MFA attempts to find a mapping $\mathbf{P}$ by maximizing the interclass locality graph and minimizing the intraclass locality graph.

Inspired by cost-sensitive learning, Lu and Tan [16], [17] proposed several cost-sensitive variants of the four classical subspace learning methods, such as CSPCA, CSLDA, CSLPP, and CSMFA, for face recognition, in which a cost matrix was predefined in advance under the prototypes of the learning methods. The cost-sensitive subspace learning meth-

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

ods proposed in [16] and [17] aim at imposing a cost matrix as contribution coefficients into the original framework PCA, LDA, LPP, and MFA. However, the cost matrix needs to be manually defined with human intervention based on prior knowledge, which is not suitable in real-world applications.

Zhang and Zhou [15] proposed a cost-sensitive face recognition method (mcKNN) and introduced the cost-sensitive learning concept. Yan [21] proposed a cost-sensitive ordinal regression method (CSOR) for facial attractiveness assessment. Liu and Zhou [18] proposed a cost-interval-based support vector machine (CISVM) for addressing the cost matrix construction using the cost-interval concept. Lee *et al.* [27] proposed a multicategory support vector machine (mcSVM) method for cost-sensitive learning. Zhang and Zhang [44] proposed an evolutionary cost-sensitive method for extreme learning machine. Cost-sensitive learning is closely related to the issues of imbalanced samples and different misclassification losses, thereby attracting people's attention in cost-sensitive tasks. How to automatically optimize the cost matrix instead of manual annotation is necessary.

III. Proposed ECSDL

In this section, the proposed ECSDL framework is formulated and then the optimization is derived.

A. Notations

Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathbb{R}^{D\times N}$ be the feature matrix with c classes, where D is the dimension and N denotes the number of training samples. The label vector with respect to $\mathbf{X}$ is represented $\mathbf{T}$. The cost matrix is represented as $\boldsymbol{\mathcal{N}}$ and the basis of subspace projection is represented as $W_{\boldsymbol{\mathcal{N}}} \in \Re^{D\times d}$. The within-class scatter matrix and the between-class scatter matrix are represented as $\mathbf{S}_{w(\boldsymbol{\mathcal{N}})}$ and $\mathbf{S}_{b(\boldsymbol{\mathcal{N}})}$, respectively. $\mathrm{Tr}(\cdot)$ denotes the trace operator and the subscript ^{T} denotes transpose operator.

B. Formulation of ECSDL

In LDA-based subspace learning, it aims at finding a transformation $W \in \mathbb{R}^{D\times d}$, such that in the projected low-dimensional subspace, the trace of the within-class scatter matrix is minimized for compactness and trace of the between-class scatter matrix is maximized for separability. This goal can be achieved by solving the following problem:

$$\max_{\boldsymbol{W}} \frac{\mathrm{Tr}(\boldsymbol{W}^T\mathbf{S}_b\boldsymbol{W})}{\mathrm{Tr}(\boldsymbol{W}^T\mathbf{S}_w\boldsymbol{W})} \tag{1}$$

where $\mathbf{S}_b$ and $\mathbf{S}_w$ denote the *between-class* and *within-class* scatter matrices, respectively, represented as

$$\mathbf{S}_b = \sum_{l,k=1}^{c} (\mathbf{m}_k - \mathbf{m}_l)(\mathbf{m}_k - \mathbf{m}_l)^T \tag{2}$$

$$\mathbf{S}_w = \sum_{k=1}^{c}\sum_{i=1}^{N_k} 1/N_k(\mathbf{x}_i - \mathbf{m}_k)(\mathbf{x}_i - \mathbf{m}_k)^T \tag{3}$$

where $\mathbf{m}_k$ and $\mathbf{m}_l$ denote the center of class k and the center of class l, respectively.

From (2), we observe that the same scaling coefficient "1" for each class pair is posed in computing the *between-class* scatter matrix. From (3), we see that the coefficient $1/N_k$ depends only on the number of samples of each class in computing the *within-class* scatter matrix. This is obviously not appropriate in cost-sensitive learning tasks, because different losses may be caused in different classes. Therefore, the proposed ECSDL method tends to reformulate the two discriminative matrices using a cost-sensitive strategy.

In this paper, for computing the cost-sensitive *within-class* $\mathbf{S}_w(\boldsymbol{\mathcal{N}})$ and *between-class* scatter matrix $\mathbf{S}_{b(\boldsymbol{\mathcal{N}})}$, the cost matrix $\boldsymbol{\mathcal{N}}$ of ECSDL with c classes is presented as follows:

$$\boldsymbol{\mathcal{N}} = \begin{bmatrix} 0 & \mathcal{N}_{12} & \cdots & \mathcal{N}_{1q} & \cdots & \mathcal{N}_{1c} \\ \mathcal{N}_{21} & 0 & \cdots & \mathcal{N}_{2q} & \cdots & \mathcal{N}_{2c} \\ \vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ \mathcal{N}_{q1} & \mathcal{N}_{q2} & \cdots & 0 & \cdots & \mathcal{N}_{qc} \\ \vdots & \vdots & \cdots & \vdots & \ddots & \vdots \\ \mathcal{N}_{c1} & \mathcal{N}_{c2} & \cdots & \mathcal{N}_{cq} & \cdots & 0 \end{bmatrix}_{c\times c} \tag{4}$$

where $\mathcal{N}_{i,j}$ denotes the misclassification loss of classifying the ith class as the jth class, and the diagonal elements "0" denote the correct classification without loss. To measure the loss of the kth class, an importance function is defined as

$$\wp(k) = \sum_{l=1}^{c} \mathcal{N}_{k,l}, \quad k = 1, \ldots, c. \tag{5}$$

From (5), we can observe that $\wp(k), k = 1, \ldots, c$, describes the misclassification loss of each class. With the definition of $\boldsymbol{\mathcal{N}}$ in (4) and $\wp(k)$ in (5), the cost-sensitive *between-class* scatter matrix and the cost-sensitive *within-class* scatter matrix can be represented as

$$\mathbf{S}_{b(\boldsymbol{\mathcal{N}})} = \sum_{k=1}^{c}\sum_{l=1}^{c} \mathcal{N}_{k,l}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_l)(\boldsymbol{\mu}_k - \boldsymbol{\mu}_l)^T \tag{6}$$

$$\mathbf{S}_{w(\mathbf{N})} = \sum_{k=1}^{c}\sum_{i=1}^{N_k} \wp(k)(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \tag{7}$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\mu}_l$ denote the center of class k and the center of class l, respectively, N_k denotes the number of samples in class k, and $\mathbf{x}_i$ denotes the ith sample vector from class k. From (6) and (7), we can observe that the weighted loss for each class pair is formulated and the derived within-class and between-class scatter matrices becomes "cost sensitive." Note that the definitions of (6) and (7) are inspired by the classical LDA, and the cluster for each class is supposed to be convex.

In ECSDL, we aim at finding a discriminative subspace $W_{\boldsymbol{\mathcal{N}}}$, such that in this subspace, the trace of the within-class scatter matrix is minimized and the trace of the between-class scatter matrix is maximized. To this end, the between-class separability and within-class compactness in the discriminative subspace will be much improved, since the cost matrix is well imposed in $\mathbf{S}_{w(\boldsymbol{\mathcal{N}})}$ (cost-sensitive *within-class* scatter matrix) and $\mathbf{S}_{b(\boldsymbol{\mathcal{N}})}$ (cost-sensitive*between-class* scatter matrix).

Therefore, the two matrices under $W_{\mathcal{N}}$ can be represented as

$$\mathbf{S}'_{w(\mathcal{N})} = \sum_{k=1}^{c}\sum_{i=1}^{N_k} \wp(k)\left(W_{\mathcal{N}}^T\mathbf{x}_i - W_{\mathcal{N}}^T\boldsymbol{\mu}_k\right)\left(W_{\mathcal{N}}^T\mathbf{x}_i - W_{\mathcal{N}}^T\boldsymbol{\mu}_k\right)^T \tag{8}$$

$$\mathbf{S}'_{b(\mathcal{N})} = \sum_{k=1}^{c}\sum_{l=1}^{c} N_{k,l}\left(W_{\mathcal{N}}^T\boldsymbol{\mu}_k - W_{\mathcal{N}}^T\boldsymbol{\mu}_l\right)\left(W_{\mathcal{N}}^T\boldsymbol{\mu}_k - W_{\mathcal{N}}^T\boldsymbol{\mu}_l\right)^T. \tag{9}$$

Obviously, for learning such a discriminative subspace, we would like to achieve the following optimizations:

$$\begin{cases}\min\limits_{W\mathcal{N}} \operatorname{Tr}\left(\mathbf{S}'_{w(\mathcal{N})}\right)\\ \max\limits_{W\mathcal{N}} \operatorname{Tr}\left(\mathbf{S}'_{b(\mathcal{N})}\right).\end{cases} \tag{10}$$

For formulating the optimization problem (10) into a unified minimization or maximization problem, we rewrite the optimization goal of (10) as follows:

$$\min_{W\mathcal{N}} \frac{\operatorname{Tr}\left(\mathbf{S}'_{w(\mathcal{N})}\right)}{\operatorname{Tr}\left(\mathbf{S}'_{b(\mathcal{N})}\right)}. \tag{11}$$

By substituting (8) and (9) into (11), the ECSDL can be derived as follows:

$$\begin{aligned}&\min_{W\mathcal{N}} \frac{\operatorname{Tr}\left(\sum_{k=1}^{c}\sum_{i=1}^{N_k}\wp(k)\left(\boldsymbol{W}_{\mathcal{N}}^T\mathbf{x}_i-\boldsymbol{W}_{\mathcal{N}}^T\boldsymbol{\mu}_k\right)\left(\boldsymbol{W}_{\mathcal{N}}^T\mathbf{x}_i-\boldsymbol{W}_{\mathcal{N}}^T\boldsymbol{\mu}_k\right)^T\right)}{\operatorname{Tr}\left(\sum_{k=1}^{c}\sum_{l=1}^{c}N_{k,l}\left(\boldsymbol{W}_{\mathcal{N}}^T\boldsymbol{\mu}_k-\boldsymbol{W}_{\mathcal{N}}^T\boldsymbol{\mu}_l\right)\left(\boldsymbol{W}_{\mathcal{N}}^T\boldsymbol{\mu}_k-\boldsymbol{W}_{\mathcal{N}}^T\boldsymbol{\mu}_l\right)^T\right)}\\
&= \min_{W\mathcal{N}} \frac{\operatorname{Tr}\left(\boldsymbol{W}_{\mathcal{N}}^T\left(\sum_{k=1}^{c}\sum_{i=1}^{N_k}\wp(k)(\mathbf{x}_i-\boldsymbol{\mu}_k)(\mathbf{x}_i-\boldsymbol{\mu}_k)^T\right)\boldsymbol{W}_{\mathcal{N}}\right)}{\operatorname{Tr}\left(\boldsymbol{W}_{\mathcal{N}}^T\left(\sum_{k=1}^{c}\sum_{l=1}^{c}N_{k,l}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_l)(\boldsymbol{\mu}_k-\boldsymbol{\mu}_l)^T\right)\boldsymbol{W}_{\mathcal{N}}\right)}\\
&= \min_{W\mathcal{N}} \frac{\boldsymbol{W}_{\mathcal{N}}^T\left(\sum_{k=1}^{c}\sum_{i=1}^{N_k}\wp(k)(\mathbf{x}_i-\boldsymbol{\mu}_k)(\mathbf{x}_i-\boldsymbol{\mu}_k)^T\right)\boldsymbol{W}_{\mathcal{N}}}{\boldsymbol{W}_{\mathcal{N}}^T\left(\sum_{k=1}^{c}\sum_{l=1}^{c}N_{k,l}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_l)(\boldsymbol{\mu}_k-\boldsymbol{\mu}_l)^T\right)\boldsymbol{W}_{\mathcal{N}}}\\
&= \min_{W\mathcal{N}} \frac{\boldsymbol{W}_{\mathcal{N}}^T\mathbf{S}_{w(\mathrm{N})}\boldsymbol{W}_{\mathcal{N}}}{\boldsymbol{W}_{\mathcal{N}}^T\mathbf{S}_{b(\mathrm{N})}\boldsymbol{W}_{\mathcal{N}}}\end{aligned} \tag{12}$$

where $\mathbf{S}_{b(\mathcal{N})}$ and $\mathbf{S}_{w(\mathcal{N})}$ are computed by (6) and (7).

According to (12), multiple solutions of $\boldsymbol{W}_{\mathcal{N}}$ may exist due to the ratio minimization. Therefore, for the uniqueness of the solution $\boldsymbol{W}_{\mathcal{N}}$, we impose an equality constraint $\boldsymbol{W}_{\mathcal{N}}^T\mathbf{S}_{b(\mathcal{N})}\boldsymbol{W}_{\mathcal{N}}=\mathbf{I}$, such that the projection basis is normalized in scale. Therefore, the ECSDL model can be rewritten as

$$\begin{aligned}&\min_{W\mathcal{N}} \ \boldsymbol{W}_{\mathcal{N}}^T\mathbf{S}_{w(\mathcal{N})}\boldsymbol{W}_{\mathcal{N}}\\ &\text{s.t. } \boldsymbol{W}_{\mathcal{N}}^T\mathbf{S}_{b(\mathcal{N})}\boldsymbol{W}_{\mathcal{N}} = \mathbf{I}\end{aligned} \tag{13}$$

where $\boldsymbol{W}_{\mathcal{N}}$ denotes the discriminative subspace of ECSDL with respect to a given cost matrix $\mathcal{N}$. From (13), we could find that for better discrimination, $\operatorname{Tr}(\mathbf{S}_{b(\mathcal{N})})$ is maximized and $\operatorname{Tr}(\mathbf{S}_{w(\mathcal{N})})$ is minimized physically. For solving the optimal projection $\boldsymbol{W}_{\mathcal{N}}$ and the optimal cost matrix $\mathcal{N}$, a variable alternating optimization method is used. Under a given cost matrix $\mathcal{N}$, (13) can be easily solved based on eigendecomposition and Lagrange multiplier method. Once the $\boldsymbol{W}_{\mathcal{N}}$ is fixed, the cost matrix $\mathcal{N}$ optimization becomes an *NP*-hard problem, and thus evolutionary search is derived.

Algorithm 1 Solve $\boldsymbol{W}_{\mathcal{N}}$ With a Given Cost Matrix $\mathcal{N}$

Input: Training set $\{\mathbf{x}_i\}_{i=1}^{N}$, the training target matrix $\mathbf{T}$;
Procedure:
1. Initialize the cost matrix $\mathcal{N}$;
2. Compute the importance function $\wp(k), k = 1, \cdots, c$ via Eq.(5);
3. Compute $\mathbf{S}_{b(\mathcal{N})}$ via Eq.(6);
4. Compute $\mathbf{S}_{w(\mathcal{N})}$ via Eq.(7);
5. Perform Eigen-decomposition on the matrix $\mathbf{S}_{b(\mathcal{N})}^{-1}\mathbf{S}_{w(\mathcal{N})}$ in Eq.(16);
6. Get the subspace projection $\boldsymbol{W}_{\mathcal{N}}^* = [\mathbf{w}_1, \cdots, \mathbf{w}_d]$.
Output: $\boldsymbol{W}_{\mathcal{N}}^*$.

C. Optimization

The optimization ECSDL includes two steps: solve $\boldsymbol{W}_{\mathcal{N}}$ under a given $\mathcal{N}$ and solve $\mathcal{N}$ under $\boldsymbol{W}_{\mathcal{N}}$.

First, solving $\boldsymbol{W}_{\mathcal{N}}$ with a given cost matrix $\mathcal{N}$ is as follows.

The Lagrange multiplier function of (13) is written as

$$\operatorname{Lag}(\boldsymbol{W}_{\mathcal{N}}, \lambda) = \boldsymbol{W}_{\mathcal{N}}^T\mathbf{S}_{w(\mathcal{N})}\boldsymbol{W}_{\mathcal{N}} - \lambda\left(\boldsymbol{W}_{\mathcal{N}}^T\mathbf{S}_{b(\mathcal{N})}\boldsymbol{W}_{\mathcal{N}} - \mathbf{I}\right) \tag{14}$$

where λ is the Lagrange multiplier.

Let the partial derivative of $\operatorname{Lag}(\boldsymbol{W}_{\mathcal{N}}, \lambda)$ with respect to $\boldsymbol{W}_{\mathcal{N}}$ be 0, we have the following equality:

$$\mathbf{S}_{w(\mathcal{N})}\boldsymbol{W}_{\mathcal{N}} = \lambda\mathbf{S}_{b(\mathcal{N})}\boldsymbol{W}_{\mathcal{N}}. \tag{15}$$

By multiplying the $\mathbf{S}_{b(\mathcal{N})}^{-1}$ in (15), we have

$$\mathbf{S}_{b(\mathcal{N})}^{-1}\mathbf{S}_{w(\mathcal{N})}\boldsymbol{W}_{\mathcal{N}} = \lambda\boldsymbol{W}_{\mathcal{N}}. \tag{16}$$

From (16), we can observe that $\boldsymbol{W}_{\mathcal{N}}$ is expressed as the eigenvectors of $\mathbf{S}_{b(\mathcal{N})}^{-1}\mathbf{S}_{w(\mathcal{N})}$ with respect to the eigenvalues $\lambda = [\lambda_1, \ldots, \lambda_D]$. Since (13) is a minimization problem, the optimal $\boldsymbol{W}_{\mathcal{N}}^* = [\mathbf{w}_1, \ldots, \mathbf{w}_d] \in R^{D\times d}$ is formulated by the d eigenvectors with respect to the first d minimum eigenvalues, $\lambda_1, \ldots, \lambda_d$.

Specifically, the solving process of $\boldsymbol{W}_{\mathcal{N}}^*$ is summarized in Algorithm 1.

It is clear that $\boldsymbol{W}_{\mathcal{N}}^*$ is closely related to the cost matrix $\mathcal{N}$. Particularly, in LDA, the cost matrix $\mathcal{N}$ is simply recognized as an identity matrix (i.e., the same loss) and the discriminative ability of subspace learning is restricted. Therefore, finding an optimal $\mathcal{N}^*$ that gives rise to the best subspace projection $\boldsymbol{W}_{\mathcal{N}^*}^*$ becomes very important.

Second, solving $\mathcal{N}$ under the computed $\boldsymbol{W}_{\mathcal{N}}^*$ is as follows.

After obtaining $\boldsymbol{W}_{\mathcal{N}}^*$, the ECSDL tends to find the optimal cost matrix $\mathcal{N}^*$ by solving the following subproblem:

$$\begin{aligned}\mathcal{N}^* = \operatorname*{argmin}_{\mathcal{N}} &\sum\nolimits_n \mathcal{L}\{t_n, \mathcal{g}_{\mathrm{NN}}(\hat{\mathbf{x}}_n)\}\\ \text{s.t. } & l_1 \le N_{i,j} \le l_2, \quad \mathcal{N}_{i,i} = 0\\ & i = 1, \ldots, c; \quad j = 1, \ldots, c\end{aligned} \tag{17}$$

where $\mathcal{L}\{\cdot\}$ denotes the misclassification loss function (e.g., recognition error rate), t_n is the label of sample $\mathbf{x}_n$, and $\mathcal{g}_{\mathrm{NN}}(\cdot)$ denotes the nearest neighbor (NN) classifier. Notably, the proposed model serves for classification/recognition problems. Therefore, misclassification loss is instinctively formulated for minimization, as general machine learning models do. $\hat{\mathbf{x}}_n$ denotes the low-dimensional subspace representation of $\mathbf{x}_n$, which is computed as

$$\hat{x}_n = \left(\boldsymbol{W}^*_{\mathcal{N}}\right)^T \mathbf{x}_n. \tag{18}$$

The predicted label $\hat{k}_n$ of a test instance $\mathbf{x}_n$ is obtained by performing the NN classifier as

$$\hat{k}_n = g_{\text{NN}}\left(\left(\boldsymbol{W}^*_{\mathcal{N}}\right)^T \mathbf{x}_n\right). \tag{19}$$

According to (17), the optimal cost matrix $\mathcal{N}^*$ will lead to the minimum misclassification loss. The optimization of $\mathcal{N}$ is presented as follows.

Considering the *NP-hard* problem of $\mathcal{N}$ optimization, the evolutionary algorithm (EA) is employed intuitively with a boundary constraint. EA is a population-based stochastic search strategy and used to search for near-optimal solutions. It tends to evolve a trial individual into a new individual with better fitness, using various genetic or bioinspired operators. It is clear that in EA community, there are a number of population-based EAs, such as swarm-based optimization, simulated annealing, and the genetic algorithm. These algorithms can be integrated in the proposed ECSDL method for evolutionary optimization. Consider that the evolutionary-based algorithm is not the focus of this paper, but a strategy for solving the NP-hard problem of cost matrix $\mathcal{N}$, and therefore, different EAs are not exploited, discussed, and finely compared. For an efficient implementation, in this paper, we derive an efficient EA, i.e., backtracking search optimization algorithm (BSA) structured in [19] and [44], to optimize the cost matrix $\mathcal{N}$. The BSA, as a random searching method, includes three basic genetic operators: selection, mutation, and crossover in a simple structure, and therefore, it is effective, fast, and capable of solving multimodal problems. In summary, it can be briefly described as four stages: 1) *initialization*; 2) *selection-I*; 3) *recombination;* and 4) *selection-II*. Specifically, the four steps are presented as follows.

1) Initialization (Generation and Evaluation of a Population $\mathcal{P}$):

$$\mathcal{P}_{i,j} \sim U\left(l_d^j, l_u^j\right), \quad i = 1, \ldots, N; \;\; j = 1, \ldots, D \tag{20}$$

$$\mathcal{F}_i = \text{ObjFun}_i(\mathcal{P}_i), \quad i = 1, \ldots, N \tag{21}$$

where $\mathcal{P}$ is encoded as the solution of $\mathcal{N}$, N and D denote the population size and problem dimension, respectively, l_d^j and l_u^j denote the low and upper bounds with respect to the jth element, respectively, $U(\cdot)$ denotes uniform distribution, and ObjFun$(\cdot)$ denotes the objective function of (17), i.e., recognition error rate.

2) Selection-I (Update Step for Historical Population $\mathcal{Q}$):

$$\mathcal{Q}_{i,j} \sim U\left(l_d^j, l_u^j\right) \tag{22}$$

$$\text{if} a < b \text{ then } \mathcal{Q} = \mathcal{P}, \quad \forall a, b \sim U(0,1) \tag{23}$$

$$\mathcal{Q}' = \textit{permuting}(\mathcal{Q}) \tag{24}$$

where *permuting*$(\cdot)$ is a random shuffling function. The historical population is for memory characteristics.

3) Recombination (Update Step for Solution Population $\mathcal{P}'_{new}$):

$$\text{Binary mapping matrix } \mathcal{C}_{N\times D}|0-1 \tag{25}$$

$$\mathcal{P}_{\text{new}} = \mathcal{P} + 3r \cdot \mathcal{C} \odot (\mathcal{Q}' - \mathcal{P}) \tag{26}$$

$$\mathcal{P}'_{\text{new}(i,j)} = \begin{cases} \begin{cases} l_d^j, & \text{if rand}^1 < \text{rand}^2 \text{ and } \\ & \mathcal{P}_{\text{new}(i,j)} < l_d^j \\ \text{rand} \times \left(l_u^j - l_d^j\right) + l_d^j, & \text{otherwise} \end{cases} \\ \begin{cases} l_u^j, & \text{if rand}^1 < \text{rand}^2 \text{ and } \\ & \mathcal{P}_{\text{new}(i,j)} > l_u^j \\ \text{rand} \times \left(l_u^j - l_d^j\right) + l_d^j, & \text{otherwise} \end{cases} \end{cases} \tag{27}$$

where $\mathcal{P}'_{\text{new}(i,j)}$ represents the jth element of the ith individual, $\odot$ denotes dot product, $r \sim N(0,1)$, *rand*^{1} and *rand*^{2} $\sim U(0,1)$, and $N(0,1)$ denotes standard normal distribution.

Then, the new population is evaluated by computing

$$\mathcal{F}'_i = \text{ObjFun}_i\left(\mathcal{P}'_{\text{new}}\{i\}\right), \quad i = 1, \ldots, N \tag{28}$$

where $\mathcal{P}'_{\text{new}}\{i\}$ denotes the ith individual of the population.

4) Selection-II (Generation of New Solution Population $\mathcal{P}''_{new}$, Global Minimum F_{gmin}, and the Optimal Solution $\mathbf{G}_{opt}$):

$$\mathcal{P}''_{\text{new}} = \mathcal{P}'_{\text{new}}\left\{\mathcal{F}'_i < \mathcal{F}_i\right\} \bigcup \mathcal{P}\left\{\mathcal{F}'_i \geq \mathcal{F}_i\right\}, \quad i = 1, \ldots, N \tag{29}$$

$$\mathcal{F}_{\text{gmin}} = \min\left\{\mathcal{F}\left\{\mathcal{F}'_i \geq \mathcal{F}_i\right\} \bigcup \mathcal{F}'\left\{\mathcal{F}'_i < \mathcal{F}_i\right\}\right\}, \quad i = 1, \ldots, N \tag{30}$$

$$\mathcal{G}_{\text{opt}} = \mathcal{P}''_{\text{new}}\{\text{ind}_{\text{opt}}\} \tag{31}$$

where $\text{ind}_{\text{opt}} = \min\{\mathcal{F}\{\mathcal{F}'_i \geq \mathcal{F}_i\} \bigcup \mathcal{F}'\{\mathcal{F}'_i < \mathcal{F}_i\}\}$ denotes the index of the best individual and $\mathcal{G}_{\text{opt}}$ denotes the best individual with respect to the index ind_{opt}.

Specifically, the optimization process of the cost matrix $\mathcal{N}$ based on the evolutionary BSA algorithm is summarized in Algorithm 2. Finally, the implementation of the proposed ECSDL framework with alternating step optimization between the cost-sensitive subspace $\boldsymbol{W}_{\mathcal{N}}$ and the cost matrix $\mathcal{N}$ is summarized in Algorithm 3.

D. Classification

This paper tends to learn a cost-sensitive discriminative subspace $\boldsymbol{W}^*_{\mathcal{N}^*}$ for effective feature representation. In classification, we adopt the Euclidean-distance-induced NN classifier. The predicted label $\hat{k}_n$ of a test instance $\mathbf{x}_n$ can be computed using (19).

IV. Human Attractiveness Analysis for Vision Application

Human attractiveness analysis is an emerging subject in computer vision and biometric community. Ancient Greek scholars measure the vertical and horizontal distances among eyes, nose, mouth, and so on and propose some general rules such as golden ratio to evaluate the attractiveness of faces. Facial attractiveness assessment using geometric- and appearance-based features coupled with pattern recognition

Algorithm 2 Solve $\mathfrak{N}$ Under the Computed $W^*_{\mathfrak{N}}$

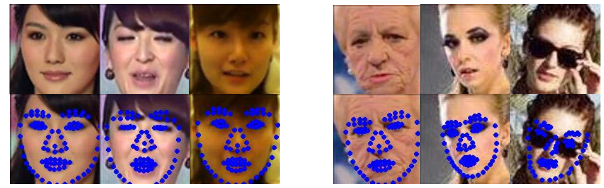
Input: The population size N, problem dimension D, lower and upper bounds l_d and l_u, the maximal iterations *epoch*, and $W^*_{\mathfrak{N}}$;
Procedure:
1. *Initialization*:
1.1. Population generation $\mathfrak{P}_{i,j} \leftarrow \mathrm{U}(l_d^j, l_u^j)$ using Eq.(20);
1.2. Objective function evaluation using Eq.(21);
while *iteration<epoch* **do**
2. *Selection-I*: *update step for historical population.*
2.1. Historical population $\mathfrak{Q}_{i,j} \leftarrow \mathrm{U}(l_d^j, l_u^j)$ using Eq.(22);
2.2. Redefine $\mathfrak{Q} \leftarrow \mathfrak{P}$ using 'if-then' rule in Eq.(23) for memory;
2.3. Permute $\mathfrak{Q}' \leftarrow$ *permuting*$(\mathfrak{Q})$ by shuffling Eq.(24);
3. *Recombination*: *update step for solution population.*
3.1. Generate crossover mapping matrix using Eq.(25);
3.2. Perform mutate using Eq.(26);
3.3. Boundary control with Eq.(27);
3.4. Objective function evaluation with the new population using Eq.(28);
4. *Selection-II: update step for new solution population, global minimum and optimal solution.*
4.1. Update population using Eq.(29);
4.2. Update the global minimum $\mathcal{F}_{best}$ using Eq.(30);
4.3. Update the optimal solution using Eq.(31)
end while
Output: $\mathfrak{N}^*$.

Algorithm 3 ECSDL

Input: Training set $\{\mathbf{x}_i\}_{i=1}^{N}$, the training target matrix $\mathbf{T}$;
Procedure:
1. Initialize the cost matrix $\mathfrak{N}$;
2. Compute $W^*_{\mathfrak{N}}$ by using Algorithm 1;
3. Search the optimal cost matrix $\mathfrak{N}^*$ by using Algorithm 2;
4. Compute the importance function $\wp^*(k)$, $k = 1, \cdots, c$ via Eq.(5);
5. Compute $\mathbf{S}_{b(\mathfrak{N}^*)}$ and $\mathbf{S}_{w(\mathfrak{N}^*)}$ using Eq.(6) and (7);
6. Perform Eigen-decomposition on the matrix $\mathbf{S}^{-1}_{b(\mathfrak{N}^*)}\mathbf{S}_{w(\mathfrak{N}^*)}$ in Eq.(16)
7. Get the optimal cost-sensitive subspace $W^*_{\mathfrak{N}^*} = [\mathbf{w}_1, \cdots, \mathbf{w}_d]$.
Output: $W^*_{\mathfrak{N}^*}$ and $\mathfrak{N}^*$.

techniques have been studied in [20]. We explore the human attractiveness analysis in this paper because it is recognized as a weak cost-sensitive learning task [21]. For example, to a certain extent, it may be less serious to classify an "unattractive" person as "attractive" than to classify an "attractive" person as "unattractive" in psychology.

Recently, a public multimodality beauty (M^{2}B) database that includes three modalities: facial, dressing, and vocal from eastern and western females is proposed [22]. In this section, we will exploit the proposed ECSDL method on the M^{2}B data for human attractiveness analysis including *facial*, *dressing*, and *vocal* attractiveness assessments.



(a) (b)

Fig. 1. Examples of face pair of (a) eastern and (b) western females.

A. M^{2}B Database

In the M^{2}B database, the facial, dressing, and vocal features were collected from 620 eastern females (i.e., Chinese, Korean, and Japanese) and 620 western females (i.e., Caucasian, consisting of Angles, Celtic, Latin, and Germanic). For *facial* attractiveness analysis, the geometric feature, i.e., shape context based on 87 landmark points, and the appearance feature based on local binary pattern (LBP), Gabor filter response, and color moment are exploited.

Specifically, the LBP descriptor is used for capturing the small texture details, and the LBP value of an image pixel (x, y) is represented as $\mathrm{LBP}_{P,R}(x, y) = \sum_{i=0}^{P-1} b_i 2^i$, where P denotes the number of sampling points and R denotes the radius of the circle to be sampled. A Gabor filter is used for texture representation, and color moment as a low-level color descriptor consisting of the first-order moment (mean of color values) and the second-order moment (variance of color values) of the image is also used.

For *dressing* attractiveness analysis, the dressing features are extracted from 20 upper body parts and 10 lower body parts using five kinds of feature descriptors, such as histogram-oriented gradient (HOG), LBP, color moment, color histogram, and skin descriptor. The HOG and LBP features are used to describe the dressing texture attributes such as collar or curling. The color moment, color histogram, and skin descriptor are used to describe color-related dressing attributes such as shirt color.

For *vocal* attractiveness analysis, the MIRToolbox was used to extract vocal features. The specific details of facial, dressing, and vocal feature extraction methods and the attractiveness assessment results for different modalities can be found in [22]. Finally, the facial, dressing, and vocal features are reduced to 300, 300, and 50 dimensions using PCA, respectively. Some examples of facial images from eastern and western females with landmark points are shown in Fig. 1. Some examples of dressing images are shown in Fig. 2. We observe from Fig. 1 that the facial images in the M^{2}B database contain abrupt features such as illumination, poses, occlusions, and expressions. Although these features contribute to facial attractiveness, in this paper, only fontal faces with restricted setting were used in the facial attractiveness analysis for validating the proposed approach. The attractiveness scores (ground truth) of facial, dressing, and vocal features for each person were normalized within [1], [10] from k-wise ratings of raters [22].

B. Parameters Setting

In evolutionary optimization of the cost matrix shown in Algorithm 2, both the maximum *population size* and the search

TABLE I

RANK-1 RECOGNITION ACCURACY OF FACIAL, DRESSING, AND VOCAL ATTRACTIVENESS USING SUBSPACE-BASED METHODS

Feature	Race	PCA-NN	CSPCA-NN	LDA-NN	CSLDA-NN	LPP-NN	CSLPP-NN	CSMFA-NN	ECSDL-NN
Geometric	E	29.50±1.85	30.27±1.96	27.45±3.59	28.64±2.47	28.86±2.28	29.90±3.05	29.09±3.51	**39.54±1.01**
	W	29.23±2.64	29.23±2.41	30.22±2.77	30.50±2.51	30.77±3.91	29.50±2.49	29.41±1.91	**42.73±0.91**
Appearance	E	31.41±2.38	32.68±2.22	28.45±0.84	29.41±2.19	30.36±2.56	30.41±2.75	31.82±2.48	**44.55±1.26**
	W	28.55±2.94	28.55±2.26	26.09±2.99	26.59±2.44	28.91±2.34	28.00±3.28	31.50±2.85	**44.55±1.80**
Dressing	E	35.45±3.60	30.41±2.78	32.14±2.17	33.14±2.83	38.68±2.96	39.95±1.47	37.86±4.48	**43.18±0.75**
	W	27.27±3.72	24.82±2.64	23.55±2.12	24.41±2.45	23.73±4.18	29.09±3.34	29.82±2.47	**38.18±1.71**
Vocal	E	37.77±2.25	38.36±3.86	39.77±3.46	39.05±2.81	43.09±3.32	41.59±3.12	40.63±2.67	**53.64±0.90**
	W	33.14±3.50	33.59±3.03	34.41±2.80	33.91±4.33	34.68±2.14	34.50±2.37	36.77±3.07	**44.09±1.51**

TABLE II

RANK-1 RECOGNITION ACCURACY OF FACIAL, DRESSING, AND VOCAL ATTRACTIVENESS USING GENERAL CLASSIFIERS

Attribute	Race	KNN	SVM	LSSVM	CISVM	CSOR	ECSDL-NN
Appearance	E	36.45±3.03	36.59±1.59	36.23±2.17	34.91±2.79	39.61±1.19	**44.55±1.26**
	W	37.64±3.57	38.64±2.37	39.59±3.11	37.20±3.01	42.78±1.64	**44.55±1.80**
Dressing	E	41.13±3.22	39.82±3.12	40.68±2.39	39.59±5.23	42.20±2.03	**43.18±0.75**
	W	35.68±2.35	36.18±2.05	33.50±2.75	35.45±3.17	35.71±2.15	**38.18±1.71**
Vocal	E	43.30±2.97	46.59±1.89	47.27±3.08	45.50±3.15	49.95±2.02	**53.64±0.90**
	W	38.23±3.90	39.05±3.25	37.18±3.09	37.82±3.70	40.67±2.30	**44.09±1.51**

(a) (b)

Fig. 2. Examples of (a) eastern and (b) western dressing images.

epochs are set as 100 and the lower and upper boundaries are set as -1 and 1, respectively. Notably, the population size and epochs are applications specific, but the complexity would be increased with large population size and more epochs.

C. Cost-Sensitive Attractiveness Assessment/Recognition

To qualitatively evaluate human attractiveness, the objective attractiveness scores are divided into five levels of 1, 2, 3, 4, and 5, with respect to the attractiveness degree of "poor," "fair," "good," "very good," and "excellent," respectively. In experiment, the attractiveness assessment of eastern (denoted by "E") and western (denoted by "W") females is studied separately. 400 females are randomly selected from 620 persons as training set, and the remaining 220 females are determined as testing set. Then, we run each procedure ten times in a "cross-validation" manner, and the average rank-1 recognition accuracy (i.e., the ratio between the number of correctly recognized samples and the number of total testing samples) is reported with a standard deviation for each method.

The compared methods are divided into two categories.

1) The comparisons with four subspace methods such as PCA, LDA, LPP, and MFA, and their cost-sensitive variants including CSPCA, CSLDA, CSLPP, and CSMFA are presented. As described in Table I, the proposed ECSDL clearly outperforms other subspace learning methods with 10% accuracy. This demonstrates that the proposed evolutionary cost matrix optimization and the discriminative subspace learning are very effective for cost-sensitive subspace learning and classification task.
2) The comparisons with generic classifiers such as KNN, SVM, and LSSVM are provided in Table II. We also compare with the CISVM [18] proposed for addressing the cost-sensitive matrix construction problem using the concept of "CI." Additionally, a CSOR [21] that was proposed for facial attractiveness is compared. In experiments, the number of NNs is empirically set as 30. From Table II, we observe that for different tasks, CISVM performs worse than other methods.

The possible reason may be that the CI is still predefined and task dependent, instead of a precise cost value. In addition, with the increase of the CI width, the training complexity of SVM increases. Although CSOR is improved compared with SVM by introducing a cost-sensitive element, the cost matrix is prior defined and a lack of flexible property when it is adapted to different tasks and new environments.

In human attractiveness assessment with five levels from 1 to 5, the *cumulative score* proposed in [23] is selected as a metric for evaluating the proposed cost-sensitive method in recognition. Specifically, the cumulative score is defined as

$$CumScore(\vartheta) = N_{e\leq\vartheta}/N_{\text{test}} \times 100\% \quad (32)$$

where ϑ denotes the tolerated error level and $N_{e\leq\vartheta}$ denotes the number of testing instances whose absolute error e is less than ϑ ($\vartheta = 0, 1, 2, \ldots, c-1$). Note that e denotes the absolute error between the predicted label and the true label. N_{test} denotes the number of total testing instances and c denotes the number of classes. Clearly, $CumScore(0)$ denotes the rank-1 recognition. The *CumScore* curves for facial, dressing, and vocal attractiveness assessment are illustrated in Figs. 3–5, respectively. We can see that the proposed

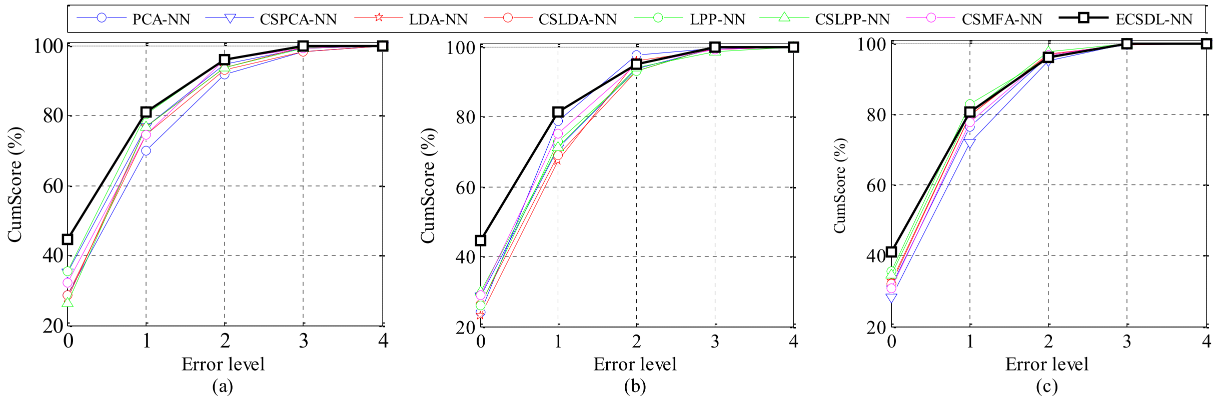


Fig. 3. Cumulative scores based on different methods for "facial" attractiveness recognition with respect to (a) eastern, (b) western, and (c) eastern + western.

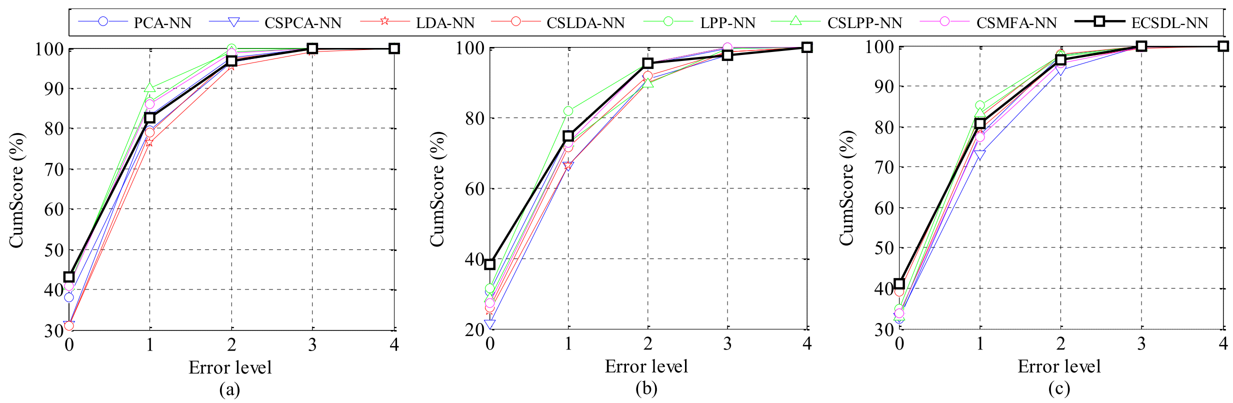


Fig. 4. Cumulative scores based on different methods for "dressing" attractiveness recognition with respect to (a) eastern, (b) western, and (c) eastern + western.

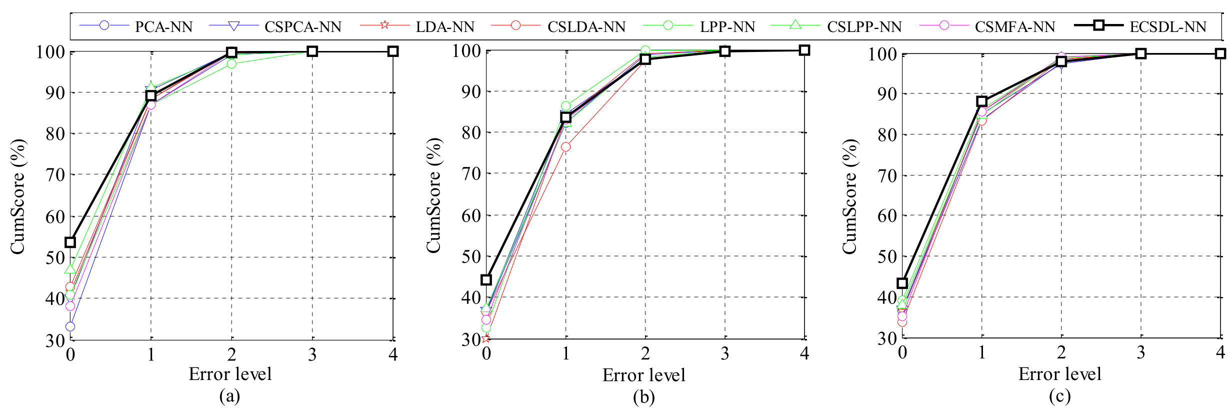


Fig. 5. Cumulative scores based on different methods for "vocal" attractiveness recognition with respect to (a) eastern, (b) western, and (c) eastern + western.

ECSDL show the best performance, particularly when $\vartheta = 0$ (i.e., rank-1 recognition).

V. Face Data Analysis for Vision Application

In this section, we conduct face recognition and face verification experiments using the proposed method. This section aims at testing the usefulness of the proposed cost-sensitive method. Two benchmark face data sets are used: 1) AR face database [24] that contains the faces of 100 persons (50 males and 50 females) and 2) the challenging LFW [25] that consists of 13 233 images of 5749 people in unrestricted environments.

A. Experiment on AR Data Set for Face Recognition

From the angle of access control system application, face recognition, as illustrated in an FR-ACS-based example, can be demonstrated as a cost-sensitive task [15]. In the implementation, we follow the same experimental setting as [26] in which seven facial images per person from Session 1 with illumination and expression changes were used for training and

TABLE III
RECOGNITION RATES OF COST-SENSITIVE SUBSPACE-BASED METHODS ON AR DATA

Methods	CSPCA	CSLPP	CSMFA	CSLDA	ECSDL
Recognition rate	68.8	45.5	69.5	86.4	**88.9**

TABLE IV
COMPARISONS WITH BASELINES AND STATE-OF-THE-ART COST-SENSITIVE FACE RECOGNITIONS ON AR DATA

Methods	NN	NS	SVM	CISVM	mckNN	mcSVM	ECSDL
Recognition rate	71.4	76.0	87.1	-	83.2	86.6	**88.9**

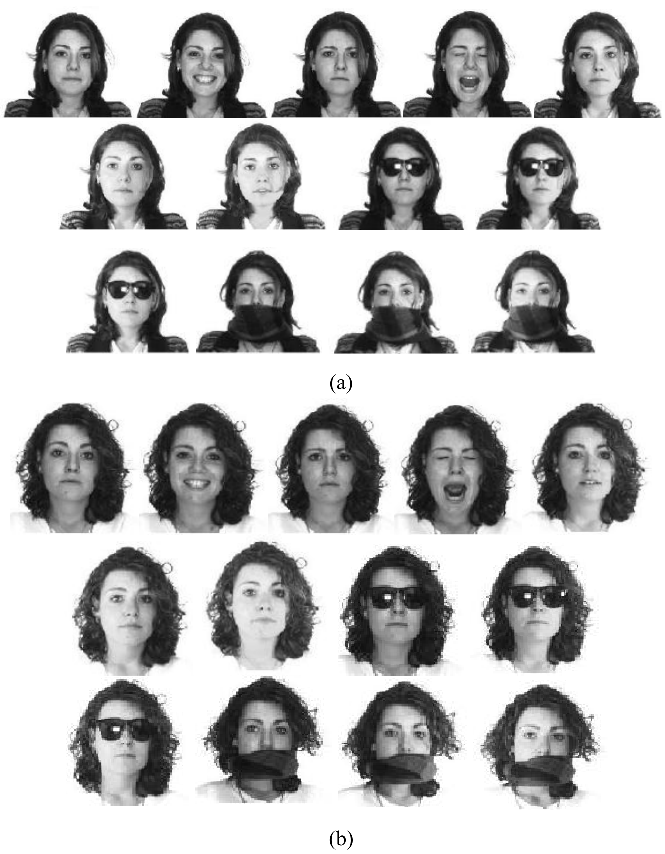


Fig. 6. One subject from two different sessions in AR database. (a) Session 1. (b) Session 2.

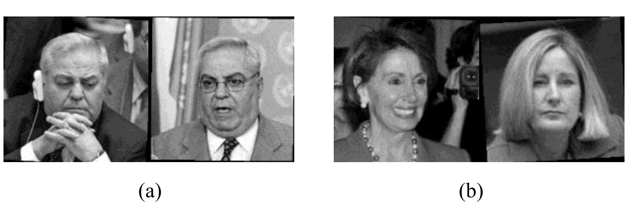


Fig. 7. Pairwise faces from LFW data set. (a) Same face pair. (b) Not same face pair.

the other seven images per person with the same condition from Session 2 were used for testing. The example images of one subject from two different sessions are shown in Fig. 6. The eigenface [2] with 300 dimensions after PCA dimension reduction is extracted as features in the experiment. An eigenface aims at extracting facial features based on PCA by transforming a face from pixel space into principal component space. The transformation is achieved by performing eigendecomposition on the covariance matrix of the training data. The eigenvectors are the so-called eigenface feature. For fair comparisons, we follow the same train/test split.

We have compared the proposed ECSDL approach with four cost-sensitive subspace-analysis-based methods (i.e., CSPCA, CSLPP, CSMFA, and CSLDA) in Table III. From Table III, we can see that the proposed ECSDL method outperforms other cost-sensitive subspace learning methods with 2% accuracy improved. CSLDA ranks the second best and demonstrates the importance of discriminative learning, but it fails in automatic learning of the cost matrix. During the subspace-based methods, CSLPP performs the worst. The reason may be that the manifold characteristic of low-dimensional embedding is not dominant in the AR database.

Additionally, three generic classifiers (e.g., NN, nearest subspace, and linear SVM) and two specialized cost-sensitive face recognition methods (e.g., mckNN [15] and mcSVM [27]) are also compared in Table IV. Some baseline results are from [15]. From Table IV, we can observe that the cost-sensitive face recognition methods such as multiclass cost-sensitive k-Nearest Neighbor (mckNN) and multiclass cost-sensitive SVM (mcSVM) outperform the conventional classifiers. Comparatively, mcSVM is superior to others (86.6%) except the proposed ECSDL method. This demonstrates that the proposed ECSDL in this paper can effectively improve face recognition.

Note that the result of CISVM is not provided because there was no report for its use in the face recognition task. With rigorous consideration, we have downloaded the released codes of CISVM and run the codes on AR data. We see that the obtained recognition accuracy is approximately 28%. Note that this paper focuses on cost-sensitive learning and subspace learning, and the state-of-the-art results based on the deep learning framework are not reported for fair comparison.

B. Experiment on LFW Data Set for Face Verification

Face verification is also a cost-sensitive problem, because we can imagine that a wrong match between two faces (stranger versus house owner) will cause a more serious loss than a wrong match between two faces (house owner versus house owner). To this end, we evaluate our method on the challenging LFW data set. Since the data set sampled from *Yahoo! News* shows large variations in pose, illumination, expression, and age, it is widely used for unrestricted face

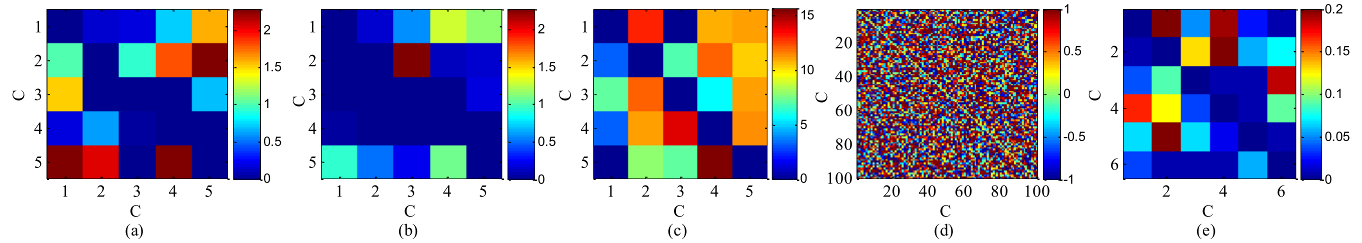


Fig. 8. Visualization of cost matrix observed in different tasks (C indicates the number of classes). Note that tasks 1, 2, and 3 denote the attractiveness assessment based on facial feature, dressing feature, and vocal feature, respectively, task 4 denotes the face recognition, and task 5 denotes the GR. (a) Cost matrix of task 1. (b) Cost matrix of task 2. (c) Cost matrix of task 3. (c) Cost matrix of task 4. (d) Cost matrix of task 4. (e) Cost matrix of task 5.

TABLE V

RECOGNITION RATES OF COST-SENSITIVE SUBSPACE-BASED METHODS ON LFW DATA

Method	CSPCA	CSLDA	CSLPP	CSMFA	ECSDL
LBP descriptor	82.87±1.18	82.45±1.69	84.30±1.45	53.18±1.70	**85.77±0.83**
SIFT descriptor	78.65±1.14	79.27±1.23	81.65±1.74	52.76±1.35	**83.43±1.50**

TABLE VI

COMPARISONS WITH STATE-OF-THE-ART METRIC LEARNING METHODS ON LFW

Method	SILD	ITML	LDML	CSML	KISSME	DML-eig	ECSDL
LBP descriptor	80.07±4.27	83.98±1.52	82.27±1.83	85.57±1.64	83.37±1.71	82.28±1.30	**85.77±0.83**
SIFT descriptor	80.85±1.93	81.45±1.45	81.05±1.52	-	83.08±1.77	81.27±7.27	**83.43±1.50**

verification and matching in the wild. Two face pairs are shown in Fig. 7.

The data set is organized into two views.

1) In *view* 1, a set consisting of 2200 pairs for training and 1000 pairs for testing is developed for model selection.
2) In *view* 2, 6000 pairs for tenfold cross validation are developed. In each fold, 600 pairs with 300 similar pairs and 300 dissimilar pairs are contained.

Note that the experimental setup for face verification is different from that for the standard face recognition. In experiment, the fair pairs are given and the decision on each pair is generally made as "same" (positive pair) or "not same" (negative pair) without needing the identity information of each person.

For this data set, state-of-the-art metric learning methods [28]–[34] are generally explored over intrapersonal subspace instead of the generic classifiers (e.g., SVM). In order to make the proposed ECSDL method suitable in LFW data, the feature vector that can reflect the similarity information is constructed for each pair. We do the experiments by following the standard protocol of LFW. The brief description is shown as follows.

For each aligned face, two facial descriptors: LBP and scale invariant feature transformation (SIFT) are used to extract features, respectively. The SIFT feature has the characteristic of scale, rotation, and translation invariance, which aims at extracting the key points of an image in the scale space. For this LFW database, the wild faces are unconstrained multiposes and different illuminations, and therefore, SIFT is a good candidate for feature description. Each face is then represented as a 300-D vector [28] after PCA dimension reduction. Due to the lack of label information, for evaluating the proposed method in this scenario, we represent a face pair using five similarity metrics: correlation coefficient, Euclidean distance, cosine distance, Mahalanobis distance, and bilinear similarity function with positive semidefinite matrix learned in [28]. Hence, a 5-D vector is formulated to represent each similar/dissimilar pair and a binary classifier is trained.

Following the tenfold cross-validation protocol on view 2, the mean accuracy of tenfold is reported with standard deviation. The results of cost-sensitive subspace methods are reported in Table V, from which we can observe that CSMFA shows the worst performance among all the methods.

The possible reason is that the constructed locality graph using k NNs of each input sample fails on the LFW database that consists of many pairwise faces, such that the intrasample information is lost. The proposed ECSDL outperforms other cost-sensitive subspace methods and the effectiveness is demonstrated further.

Moreover, we have also compared ECSDL with several state-of-the-art metric learning methods such as SILD [29], KISSME [30], CSML [31], ITML [32], LDML [33], and DML-eig [34]. The comparison results are shown in Table VI, from which we observe that our proposed ECSDL performs a significantly better recognition than metric learning methods for both descriptors. Besides, a new prospective is that group metrics can be integrated as input features for face verification by learning a binary classifier. Notably, we focus on cost-sensitive learning and subspace learning, and the deep neural networks that depend on large training data are not compared.

VI. E-NOSE DATA ANALYSIS FOR OLFACTION APPLICATION

Gas recognition (GR) is an important part in artificial olfaction. Generally, it aims at detecting the existence of

TABLE VII
RANK-1 RECOGNITION OF GASES USING SUBSPACE-LEARNING-BASED NN CLASSIFIERS

Method	PCA-NN	CSPCA-NN	LDA-NN	CSLDA-NN	LPP-NN	CSLPP-NN	CSMFA-NN	ECSDL-NN
HCHO	95.24	95.24	92.06	92.06	92.06	93.65	95.24	96.83
C_6H_6	87.50	87.50	83.33	79.17	91.67	91.67	87.50	91.67
C_7H_8	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
CO	95.00	95.00	70.00	85.00	85.00	85.00	95.00	90.00
NH_3	100.0	100.0	100.0	100.0	95.00	95.00	100.0	100.0
NO_2	84.62	84.62	69.23	69.23	76.92	76.92	84.62	100.0
ARR	93.73	93.73	85.77	87.58	90.11	90.37	93.73	**96.42**
TRR	94.44	94.44	88.27	89.51	91.35	91.98	94.44	**96.30**

TABLE VIII
RANK-1 RECOGNITION OF GASES USING GENERAL CLASSIFIERS

Method	SVM	PCA-SVM	KSVM	LDA	PCA-LDA	PLS-DA	KLDA	KPLS-DA	CISVM	ECSDL-NN
HCHO	98.41	98.41	98.41	88.89	82.54	93.65	95.24	98.41	93.65	96.83
C_6H_6	79.17	91.67	87.50	66.67	58.33	45.83	100.0	91.67	83.33	91.67
C_7H_8	100.0	100.0	100.0	90.91	86.36	68.18	95.45	95.45	72.73	100.0
CO	100.0	65.00	100.0	100.0	90.00	75.00	95.00	95.00	80.00	90.00
NH_3	90.00	100.0	95.00	90.00	90.00	70.00	95.00	90.00	95.00	100.0
NO_2	69.23	30.77	76.92	30.77	30.77	23.08	76.92	69.23	30.77	100.0
ARR	89.47	80.97	92.97	77.87	73.00	62.62	92.94	89.96	75.91	**96.42**
TRR	92.59	88.27	95.06	82.72	77.16	72.22	94.44	93.21	82.72	**96.30**

poisonous gases. If a kind of poisonous gas is wrongly recognized into a kind of nontoxic odor, it may cause harm to people's health. Therefore, GR can be viewed as a cost-sensitive problem.

The cost matrix is definitely defined according to (4). E-NOSE is an artificial olfaction system composed of a sensor array with partial specificity coupled with the pattern recognition algorithm [35]. The gaseous contaminant recognition is also recognized as a cost-sensitive recognition problem. In this section, we will explore the proposed methods on the E-NOSE database for a new application of GR. The E-NOSE database is provided in [13], which contains six kinds of gaseous contaminants, i.e., formaldehyde (HCHO), benzene (C_6H_6), toluene (C_7H_8), carbon monoxide (CO), ammonia (NH_3), and nitrogen dioxide (NO_2). The number of samples for each kind of gas is 188, 72, 66, 58, 60, and 38, respectively. In feature extraction, the steady-state point of each sensor is extracted as a feature, and as a result, a 6-D sensory feature vector is formulated for sample representation. The position of the steady-state point is set as the $4/5$ (240th point) of the whole response (300 points). The specific details of gas sensing and data acquisition can be referred to as [13].

In the implementation, two-thirds of samples of each class are selected as training set. The rank-1 recognition of each class, average recognition rate (ARR), and the total recognition rate (*TRR*) are reported. Notably, *ARR* is the ratio of the summation of all recognition rates and the number of classes, while *TRR* is the ratio between the number of correctly classified samples for all classes and the total number of samples. The comparisons with subspace-based learning methods are shown in Table VII, in which the rank-1 recognition results using the NN classifier are reported. We can observe that the proposed ECSDL performs the best recognition performance with 96.42% of *ARR* and 96.30% of *TRR*.

For comparison with existing methods (e.g., SVM, LDA, PLS-DA, and their kernel extensions) that have been used in the E-NOSE application, we report the recognition results in Table III. We can observe that the proposed ECSDL method shows the best performance. Note that the one-against-one scheme is used in SVM- and LDA-based methods.

VII. COMPUTATIONAL COMPLEXITY AND TIME ANALYSIS

For ECSDL in Algorithm 3, it involves the computational complexity of Algorithms 1 and 2. For Algorithm 1, it involves the computation of $\mathbf{S}_{b(\mathbf{N})}$, $\mathbf{S}_{w(\mathbf{N})}$ and the eigendecomposition of $\mathbf{S}_{b(\mathbf{N}^*)}^{-1}\mathbf{S}_{w(\mathbf{N}^*)}$, and thus the complexity is $O(D^3)$. For Algorithm 2, the complexity is related to the population size N and the number *epochs* for loop, and thus the extra complexity is $O(N \cdot \textit{epochs})$. Therefore, the computational complexity of ECSDL is $O(D^3) + O(N \cdot \textit{epochs})$. Note that the NN classifier is independent of the proposed method, and thus the computational complexity of NN classifier is excluded here.

With a naïve MATLAB implementation, the algorithms are run on a 2.5-GHz Windows machine with a 4-GB RAM. The computation time on the large-scale LFW data set is presented in Table VIII, from which we observe that CSPCA and CSLPP cost much more time than CSLDA. This is because the covariance matrix computation in CSPCA and the locality graph construction in CSLPP are time consuming. The CSMFA cost the most time (6731.9 s) among all the methods. The reason is that the computation of two locality graphs for intraclass and interclass is needed. The ECSDL method costs a comparatively little high computational time (2318.1 s), because of the large-scale evolutionary search process. The computational time depends on the population size and searching epochs.

TABLE IX
COMPUTATIONAL TIME ON LFW DATA SET FOR TRAINING AND TESTING IN ONEFOLD

Method	CSPCA	CSLDA	CSLPP	CSMFA	ECSDL
Time (s)	412.08	26.39	331.77	6731.9	2318.1

VIII. VISUALIZATION OF COST MATRIX

In order to visualize the learned cost matrix using the proposed model and have an insight into what the cost matrix looks like, we show the cost matrix of different tasks in experiments in Fig. 8. From the visualization of cost matrix, we know that the diagonal entries of each cost matrix are zero, which comply with the cost definition in (4). Additionally, different sizes of values are obtained in each cost matrix, which corresponds to the best classification performance. The nonuniform cost matrix also confirms the necessity of cost-sensitive learning for improving the classification ability. Additionally, we can see that the cost matrix is not symmetric, which shows that the cost of misclassifying A as B is different from the cost of misclassifying B as A. For example, in Fig. 8(a), the cost $\mathcal{N}_{12} = 0.17$ and the cost $\mathcal{N}_{21} = 1.02$. Note that classes 1 and 2 in task 1 (facial attractiveness analysis) denote "poor" and "fair," respectively. In our motivation, we expect that the cost $\mathcal{N}_{12}$ of classifying "poor" attractiveness as "fair" should be lower than that cost $\mathcal{N}_{21}$ of classifying "fair" attractiveness as "poor" (i.e., $\mathcal{N}_{12} < \mathcal{N}_{21}$), and clearly, the cost matrix complies with our expectation. Further, we consider the classification between class 1 "poor" and class 5 "excellent." We can observe that $\mathcal{N}_{15} < \mathcal{N}_{51}$, which implies that misclassifying "excellent" as "poor" leads to the highest cost and comply with our expectation. Notably, the true cost matrix is unknown and expected to be solved in our proposed ECSDL approach. The recognition accuracy as an effective measure has been shown from Tables I–VII and IX, which demonstrate the effectiveness of the proposed model.

IX. DISCUSSION

From a variety of benchmark data sets in vision and olfaction application, the generality of the proposed method is effectively and preliminarily revealed. From the perspective of algorithm, the complexity, computational cost, and the convergence of the proposed approach are optimistic. ECSDL is proposed under a cost-sensitive discriminative learning framework. EAs are widely used to solve different types of optimization problems for their rapid search in the whole solution space with heuristic and bioinspired update strategies. EAs have global exploration in the entire search space and local exploitation abilities to find the best solution near a new solution it has discovered, but they do not guarantee to find the global optimums of a problem. In this paper, the instinct optimization involves three bioinspired genetic operators, i.e., mutation, crossover, and selection. The optimal or near-optimal solutions of the proposed methods can be obtained with finite iterations and a low computational cost for real applications. The velocity depends much on the size of population. Another aspect that we would like to claim is the deep learning algorithms trained on a large-scale data. The proposed cost-sensitive learning method is proposed for general problems (i.e., small sample problems). The cost-sensitivity may be avoided if large-scale data are available.

In a real-world application scenario, similar to general-machine-learning-based classification, a batch of training data will be obtained or collected for model construction. Generally, the training process (model parameter tuning) is implemented offline and then the model parameters are used for prediction or classification. There is no exception for the proposed model and the training process including the subspace projection solver and the off-line cost matrix optimizer; then, the learned projection $\boldsymbol{W}_{\mathcal{N}}$ and the cost matrix $\boldsymbol{\mathcal{N}}$ are saved for testing. Note that the motivation of the proposed method is for cost-sensitive classification problems; however, for general classification problems, the method also works by appropriately defining the cost matrix as some particular matrix (e.g., identity matrix).

It is possible to manually assign trivial cost "0" to difficult cases. However, how to measure the difficulty becomes an open problem. That is, it is difficult to quantify how severe one type of mistake is against another one. Although users may know what type of mistake is more serious than another type, it is difficult to specify the cost value of one mistake. Cost-sensitive learning is closely related with the real classification scenario. Both discriminative subspace learning and classifier learning target at improving the classification performance. In this paper, cost-sensitive discriminative subspace learning is focused by considering the cost-sensitive nature of the feature subspace. Learning a robust cost-sensitive classifier by incorporating the weighted prediction error will be a future direction.

X. CONCLUSION

In this paper, we propose an ECSDL framework for dealing with the cost-sensitive classification tasks in real-world vision and olfaction applications. The misclassification loss is paid more attention than the single classification accuracy in modeling process. In cost-sensitive scenarios, high classification accuracy may not mean the best performance. Instead, a lower misclassification loss may be a more effective metric. The merits of this paper include the discriminative subspace $\boldsymbol{W}_{\boldsymbol{\mathcal{N}}}$ learning in Algorithm 1 for pursuit of the maximum class separability and the automatic cost matrix $\boldsymbol{\mathcal{N}}$ optimization in Algorithm 2 based on an evolutionary backtracking search algorithm. A unified ECSDL method in Algorithm 3 with a variable alternating optimization algorithm is proposed. Extensive experiments have been conducted on a variety of vision and olfaction application scenarios. The experimental results and comparisons with the state-of-the-art methods demonstrate the extremely prominent efficacy of the proposed approach for cost-sensitive recognition tasks.

In the future work, four aspects may be involved.

1) It is also challenging to make more insight into the cost matrix convex optimization based on gradient learning and make it really intelligent.

2) The proposed ECSDL method can be extended to its nonlinear version using Mercer kernel theorem, such that a kernelized version can be proposed for nonlinear subspace projection.
3) In terms of the representation ability of deep learning, the deep features of the database may be extracted for the proposed method.
4) The cost-sensitive classifier model can be directly constructed by optimizing the cost matrix such that the best classifier is solved.

ACKNOWLEDGMENT

The authors would like to thank the Editor in Chief, Associate Editor, and anonymous reviewers for their valuable comments, which greatly improved the quality of this paper.

REFERENCES

[1] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognit. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.
[2] P. N. Belhumeur, J. P. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.
[3] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using Laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.
[4] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.
[5] S. Yan, J. Liu, X. Tang, and T. S. Huang, "A parameter-free framework for general supervised subspace learning," *IEEE Trans. Inf. Forensics Security*, vol. 2, no. 1, pp. 69–76, Mar. 2007.
[6] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H.-J. Zhang, "Multilinear discriminant analysis for face recognition," *IEEE Trans. Image Process.*, vol. 16, no. 1, pp. 212–220, Jan. 2007.
[7] D. A. Humphreys, P. M. Harris, M. Rodríguez-Higuero, F. A. Mubarak, D. Zhao, and K. Ojasalo, "Principal component compression method for covariance matrices used for uncertainty propagation," *IEEE Trans. Instrum. Meas.*, vol. 64, no. 2, pp. 356–365, Feb. 2015.
[8] H. Liu, S. Li, and L. Fang, "Robust object tracking based on principal component analysis and local sparse representation," *IEEE Trans. Instrum. Meas.*, vol. 64, no. 11, pp. 2863–2875, Nov. 2015.
[9] M. Van and H. J. Kang, "Wavelet kernel local Fisher discriminant analysis with particle swarm optimization algorithm for bearing defect classification," *IEEE Trans. Instrum. Meas.*, vol. 64, no. 12, pp. 3588–3600, Dec. 2015.
[10] L. Fang and S. Li, "Face recognition by exploiting local Gabor features with multitask adaptive sparse representation," *IEEE Trans. Instrum. Meas.*, vol. 64, no. 10, pp. 2605–2615, Oct. 2015.
[11] S. Joardar, A. Chatterjee, and A. Rakshit, "A real-time palm dorsa subcutaneous vein pattern recognition system using collaborative representation-based classification," *IEEE Trans. Instrum. Meas.*, vol. 64, no. 4, pp. 959–966, Apr. 2015.
[12] L. Zhang and D. Zhang, "Domain adaptation extreme learning machines for drift compensation in E-nose systems," *IEEE Trans. Instrum. Meas.*, vol. 64, no. 7, pp. 1790–1801, Jul. 2015.
[13] L. Zhang and F. Tian, "Performance study of multilayer perceptrons in a low-cost electronic nose," *IEEE Trans. Instrum. Meas.*, vol. 63, no. 7, pp. 1670–1679, Jul. 2014.
[14] H. Liu, D. Guo, and F. Sun, "Object recognition using tactile measurements: Kernel sparse coding methods," *IEEE Trans. Instrum. Meas.*, vol. 65, no. 3, pp. 656–665, Mar. 2016.
[15] 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.
[16] 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.
[17] J. Lu and Y.-P. Tan, "Cost-sensitive subspace learning for face recognition," in *Proc. IEEE Conf. CVPR*, Jun. 2010, pp. 2661–2666.
[18] X.-Y. Liu and Z.-H. Zhou, "Learning with cost intervals," in *Proc. ACM SIGKDD*, Washington, DC, USA, Jul. 2010, pp. 403–412.
[19] P. Civicioglu, "Backtracking search optimization algorithm for numerical optimization problems," *Appl. Math. Comput.*, vol. 219, no. 15, pp. 8121–8144, Apr. 2013.
[20] D. Zhang, Q. Zhao, and F. Chen, "Quantitative analysis of human facial beauty using geometric features," *Pattern Recognit.*, vol. 44, no. 4, pp. 940–950, Apr. 2011.
[21] H. Yan, "Cost-sensitive ordinal regression for fully automatic facial beauty assessment," *Neurocomputing*, vol. 129, pp. 334–342, Apr. 2014.
[22] T. V. Nguyen, S. Liu, B. Ni, J. Tan, Y. Rui, and S. Yan, "Sense beauty via face, dressing, and/or voice," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 239–248.
[23] X. Geng, Z.-H. Zhou, and K. Smith-Miles, "Automatic age estimation based on facial aging patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2234–2240, Dec. 2007.
[24] A. Martinez and R. Benavente, "The AR face database," CVC, Tech. Rep. #24, 1998.
[25] G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Amherst, MA, USA, Tech. Rep., 2007.
[26] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 543–550.
[27] 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.
[28] Q. Cao, Y. Ying, and P. Li, "Similarity metric learning for face recognition," in *Proc. IEEE ICCV*, Dec. 2013, pp. 2408–2415.
[29] M. Kan, S. Shan, D. Xu, and X. Chen, "Side-Information based linear discriminant analysis for face recognition," in *Proc. BMVC*, 2011.
[30] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 2288–2295.
[31] H. V. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," in *Proc. ACCV*, Nov. 2010, pp. 709–720.
[32] J. Davis, B. Kulis, P. Jain, S. Sra, and I. Dhillon, "Information-theoretic metric learning," in *Proc. ICML*, Jun. 2007, pp. 209–216.
[33] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? Metric learning approaches for face identification," in *Proc. ICCV*, Sep. 2009, pp. 498–505.
[34] Y. Ying and P. Li, "Distance metric learning with eigenvalue optimization," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 1–26, Jan. 2012.
[35] J. W. Gardner and P. N. Bartlett, *Electronic Noses: Principles and Applications*. London, U.K.: Oxford Univ. Press, 1999.
[36] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, Nov. 2013.
[37] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.
[38] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[39] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE Trans. Neural Netw.*, vol. 8, no. 1, pp. 98–113, Jan. 1997.
[40] G. B. Huang, H. Lee, and E. Learned-Miller, "Learning hierarchical representations for face verification with convolutional deep belief networks," in *Proc. IEEE Int. Conf. CVPR*, Jun. 2012, pp. 2518–2525.
[41] Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Int. Conf. CVPR*, Jun. 2014, pp. 1701–1708.
[42] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.
[43] A. Karpathy, G. Toderici, S. Shetty, and T. Leung, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Int. Conf. CVPR*, Jun. 2014, pp. 1725–1732.
[44] L. Zhang and D. Zhang, "Evolutionary cost-sensitive extreme learning machine," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2016.2607757.

Lei Zhang (M'14) received the Ph.D. degree in circuits and systems from the College of Communication Engineering, Chongqing University, Chongqing, China, in 2013.

He was selected as a Hong Kong Scholar in China in 2013, and was a Post-Doctoral Fellow with The Hong Kong Polytechnic University, Hong Kong, from 2013 to 2015. He is currently a Professor/Distinguished Research Fellow with Chongqing University. He has authored more than 50 scientific papers in top journals, including the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT, the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, the IEEE SENSORS JOURNAL, *Information Fusion*, *Sensors & Actuators B*, and *Analytica Chimica Acta*. His current research interests include electronic olfaction, machine learning, pattern recognition, computer vision, and intelligent systems.

Dr. Zhang was a recipient of the Outstanding Reviewer Award of the *Sensor Review* journal in 2016, the Outstanding Doctoral Dissertation Award of Chongqing in 2015, the Hong Kong Scholar Award in 2014, the Academy Award for Youth Innovation of Chongqing University in 2013, and the New Academic Researcher Award for Doctoral Candidates from the Ministry of Education, China, in 2012.

David Zhang (F'09) received the bachelor's degree in computer science from Peking University, Beijing, China, the M.Sc. and Ph.D. degrees in computer science from the Harbin Institute of Technology (HIT), Harbin, China, in 1982 and 1985, respectively, and the second Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 1994.

From 1986 to 1988, he was a Post-Doctoral Fellow with Tsinghua University, Beijing, and then an Associate Professor with the Academia Sinica, Beijing. Since 2005, he has been a Chair Professor with The Hong Kong Polytechnic University, Hong Kong, where he is the Founding Director of the Biometrics Research Center (UGC/CRC) supported by the Hong Kong SAR Government in 1998. He also serves as a Visiting Chair Professor at Tsinghua University, and an Adjunct Professor at Peking University, Shanghai Jiao Tong University, Shanghai, China, HIT, and the University of Waterloo. He has authored more than 10 books, over 300 international journal papers, and 30 patents from the U.S., Japan, Hong Kong, and China.

Prof. Zhang is a Croucher Senior Research Fellow, Distinguished Speaker of the IEEE Computer Society, and a fellow of IAPR. He is the Founder and Editor-in-Chief of the *International Journal of Image and Graphics*; the Book Editor of *Biometrics* (Springer); an Organizer of the International Conference on Biometrics Authentication; and an Associate Editor of more than 10 international journals including IEEE TRANSACTIONS.