ℓ_1-Norm Heteroscedastic Discriminant Analysis Under Mixture of Gaussian Distributions

Wenming Zheng, *Member, IEEE*, Cheng Lu, Zhouchen Lin, *Fellow, IEEE*, Tong Zhang, Zhen Cui, and Wankou Yang

***Abstract*— Fisher's criterion is one of the most popular discriminant criteria for feature extraction. It is defined as the generalized Rayleigh quotient of the between-class scatter distance to the within-class scatter distance. Consequently, Fisher's criterion does not take advantage of the discriminant information in the class covariance differences, and hence, its discriminant ability largely depends on the class mean differences. If the class mean distances are relatively large compared with the within-class scatter distance, Fisher's criterion-based discriminant analysis methods may achieve a good discriminant performance. Otherwise, it may not deliver good results. Moreover, we observe that the between-class distance of Fisher's criterion is based on the ℓ_2-norm, which would be disadvantageous to separate the classes with smaller class mean distances. To overcome the drawback of Fisher's criterion, in this paper, we first derive a new discriminant criterion, expressed as a *mixture of absolute generalized Rayleigh quotients*, based on a Bayes error upper bound estimation, where mixture of Gaussians is adopted to approximate the real distribution of data samples. Then, the criterion is further modified by replacing ℓ_2-norm with ℓ_1 one to better describe the between-class scatter distance, such that it would be more effective to separate the different classes. Moreover, we propose a novel ℓ_1-norm heteroscedastic discriminant analysis method based on the new discriminant analysis (L1-HDA/GM) for heteroscedastic feature extraction, in which the optimization problem of L1-HDA/GM can be efficiently solved by using the eigenvalue decomposition approach. Finally, we conduct extensive experiments on four real data sets and demonstrate that the proposed method achieves much competitive results compared with the state-of-the-art methods.**

***Index Terms*— Feature extraction, Fisher's discriminant criterion, heteroscedastic discriminant criterion, ℓ_1-norm heteroscedastic discriminant analysis (L1-HDA), Rayleigh quotient.**

I. Introduction

LINEAR feature extraction plays a crucial role in statistical pattern recognition [1]–[3]. The goal of linear feature extraction can be regarded as seeking a transformation matrix that transforms the input data from the original high-dimensional space to a low-dimensional space while preserving some useful information. Fisher's linear discriminant analysis (FLDA) [4] is one of the most popular linear feature extraction methods, which aims to find a set of optimal discriminant vectors, such that the projections of the training samples onto these vectors have maximal between-class scatter distance and minimal within-class scatter distance. This is realized by solving a series of discriminant vectors that maximize Fisher's discriminant criterion, defined as the generalized Rayleigh quotient of the between-class scatter distance to the within-class scatter distance. Over the past several decades, Fisher's criterion-based discriminative feature extraction methods had been successfully applied to face recognition [5], image retrieval [6], and speech recognition [7]. More recently, Yang *et al.* [8] adopted the Fisher's criterion to enhance the discriminative ability of the sparse coefficient matrix in the sparse representation model [9]. Although the Fisher's criterion had been shown to be very effective in practical applications, it should be noted that this criterion was developed under the homoscedastic distributions of the class data samples. Since the Fisher's criterion is characterized by the ratio of the between-class scatter distance to the within-class scatter distance, it may not deliver a good discriminant performance when the class mean distances are relatively small compared with the within-class scatter distance. Considering the electroencephalogram (EEG) feature extraction as an example, we cannot determine what features are the most discriminative ones according to Fisher's criterion because the EEG signal conditioned on each class is often assumed to have a zero mean [10], and hence, Fisher's ratio will always be zero. In such a case, only the class covariance matrices can be

Manuscript received July 30, 2017; revised February 7, 2018; accepted July 25, 2018. Date of publication August 29, 2018; date of current version September 18, 2019. This work was supported in part by the National Basic Research Program of China under Grant 2015CB351704 and Grant 2015CB352502, in part by the National Natural Science Foundation of China under Grant 61572009, Grant 61625301, Grant 61731018, and Grant 61772276, in part by the Jiangsu Provincial Key Research and Development Program under Grant BE2016616, and in part by Qualcomm and Microsoft Research Asia. *(Corresponding author: Wenming Zheng.)*

W. Zheng is with the Key Laboratory of Child Development and Learning Science (Ministry of Education), School of Biological Sciences and Medical Engineering, Southeast University, Nanjing 210096, China (e-mail: wenming_zheng@seu.edu.cn).

C. Lu and T. Zhang are with the Key Laboratory of Child Development and Learning Science (Ministry of Education), School of Information Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: cheng.lu@seu.edu.cn; tongzhang@seu.edu.cn).

Z. Lin is with the Key Laboratory of Machine Perception (Ministry of Education), School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China, and also with the Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: zlin@pku.edu.cn).

Z. Cui is with the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information (Ministry of Education), School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: zhen.cui@njust.edu.cn).

W. Yang is with the School of Automation, Southeast University, Nanjing 210096, China (e-mail: youngwankou@yeah.net).

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

Digital Object Identifier 10.1109/TNNLS.2018.2863264

utilized to extract the discriminant features [11]. Consequently, how to define a good discriminant criterion for extracting the useful discriminant features from the class covariance matrices is the major goal of this paper.

In order to utilize the discriminant information from both class means and class covariance matrices, many heteroscedastic discriminant criteria have been proposed during the past several years [12]–[15], [17], [18], which result in various heteroscedastic discriminant analysis (HDA) methods. Here, we divide them into the following three categories. The first one is derived under the maximum-likelihood framework [12]–[14]. A representative method of this category was proposed by Kumar and Andreou [12] for speech recognition. The second category is derived based on Chernoff distance or Bhattacharyya distance [1], [16], where a representative method, denoted as HDA/Chernoff, is based on the Chernoff criterion [15]. A similar method to HDA/Chernoff is the approximate information discriminant analysis (AIDA) method [14], which uses the so-called μ-measure [17] as the discriminant criterion. The third category investigates hybrid linear feature extraction scheme for the HDA (HDA/HLFE) [18], [19], in which the discriminative information are, respectively, extracted from the class means and class covariance matrices. Since HDA/HLFE is derived under the assumption of single Gaussian distribution of each class data samples, it should be noted that HDA/HLFE may not be well suitable for the cases where the class data samples abide by Gaussian mixture distribution. In addition, it is also notable that the discriminant vectors of HDA/HLFE are learned in two separated subspaces, such that the learned discriminant vectors may not be optimal in terms of Bayes error since the Bayes error is characterized by both class means and class covariance matrices simultaneously. For all of the aforementioned HDA methods, a common limitation of these methods is the suffering of the so-called small sample size problem [20], i.e., all these methods require that the number of samples in each class be larger than the dimension of the data space in order to guarantee the nonsingularity of the class covariance matrices.

In addition to the HDA methods, there are other discriminant analysis approaches that have been proposed in recent years to overcome the drawbacks of FLDA, e.g., multiview learning (or multimodal learning) methods [49], [50], subclass methods [21], [22], kernel-based methods [23], [24], or deep neural network method [51]. The multiview learning (or multimodal learning) methods are mainly related with the feature extraction problems of learning from the data represented by multiple distinct feature sets [52]. The subclass methods, e.g., the subclass discriminant analysis (SDA) [21], deal with the discriminative feature extraction by dividing each class samples into several subclasses, which enables this method more powerful than the FLDA method in extracting discriminative features. The kernel-based discriminant analysis (KDA) methods [23] are the nonlinear extension of FLDA via kernel trick [25] to solve the drawbacks of FLDA. In KDA, the input data samples are mapped by a nonlinear mapping from the input data space to a high-dimensional reproducing kernel Hilbert space (RKHS), such that the nonseparable data samples of the input data space become separable in RKHS. As a consequence, performing feature extraction in RKHS using FLDA results in the nonlinear feature extraction in the original input data space. Similar to the kernel-based learning methods, the deep neural network methods can also extract the nonlinear features via nonlinear neural network learning.

Although the aforementioned methods are proposed to overcome the drawbacks of FLDA, most of them are developed under the Fisher's criterion, i.e., minimizing the within-class scatter distance and maximizing the between-class scatter distance. Hence, some of the limitations of Fisher's criterion, such as the difficulty of extracting the discriminant information lying in the class covariance differences, may still exist to some extent for these methods.

In this paper, we develop a new discriminant criterion under the distributions of Gaussian and mixture of Gaussians, respectively, for heteroscedastic discriminant problems, which can be expressed as a *mixture of absolute generalized Rayleigh quotients* (MAGRQs). Preliminary applications of this paper to nonfrontal facial expression recognition and EEG classification had investigated in [26]–[28].

To show the physical meaning of our MAGRQ criterion, let us first consider a special two-class heteroscedastic case as shown in Fig. 1, in which the first row illustrates three examples of two-class homoscedastic data sets (denoted by red and green colors), whereas the second and third rows illustrate another six examples of two-class data sets with the same class means but different covariance matrices. From Fig. 1, we can see that the separability of the two-class data sets is closely related with both class means and class covariance matrices. In particularly, it is notable that even the between-class distances are the same, e.g., Fig. 1 (d)–(i), the two-class data sets associated with the largest covariance matrices difference could be best separated. Especially, in Fig. 1 (g)–(i), the class means are almost overlapped, and hence, the traditional FLDA would not be applicable, whereas the HDA method could largely separate the two-class data sets. Fig. 2 shows a special case of two-class heteroscedastic discriminant problem, where the class means are equal (zero) but the class covariance matrices are different. It is obvious that Fisher's criterion cannot be used in this scenario because of the zero class means. Now, let $\mathbf{v}$ denote a projection vector, such that the projections of two data samples $\mathbf{x}$ and $\mathbf{y}$ onto this projection vector are $\mathbf{v}^T\mathbf{x}$ and $\mathbf{v}^T\mathbf{y}$, where we suppose that $\mathbf{x}$ is from class 1 and $\mathbf{y}$ is from class 2. Intuitively, to best distinguish the data samples between the two classes, we should minimize their overlapping parts as much as possible. To this end, we may expect that one class has smaller scatter distances, whereas the other one has larger scatter distances, which means that we have to seek a projection vector $\mathbf{v}$ such that the projection of one class has a smaller variance, whereas the other one has a larger variance. This can be modeled as the following maximization problem:

$$\begin{aligned}\max_{\mathbf{v}^T\mathbf{v}=1} |\mathrm{var}(\mathbf{v}^T\mathbf{x}) - \mathrm{var}(\mathbf{v}^T\mathbf{y})| &= |\mathbf{v}^T\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{v} - \mathbf{v}^T\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{v}| \\ &= |\mathbf{v}^T(\boldsymbol{\Sigma}_{\mathbf{x}} - \boldsymbol{\Sigma}_{\mathbf{y}})\mathbf{v}| \qquad (1)\end{aligned}$$

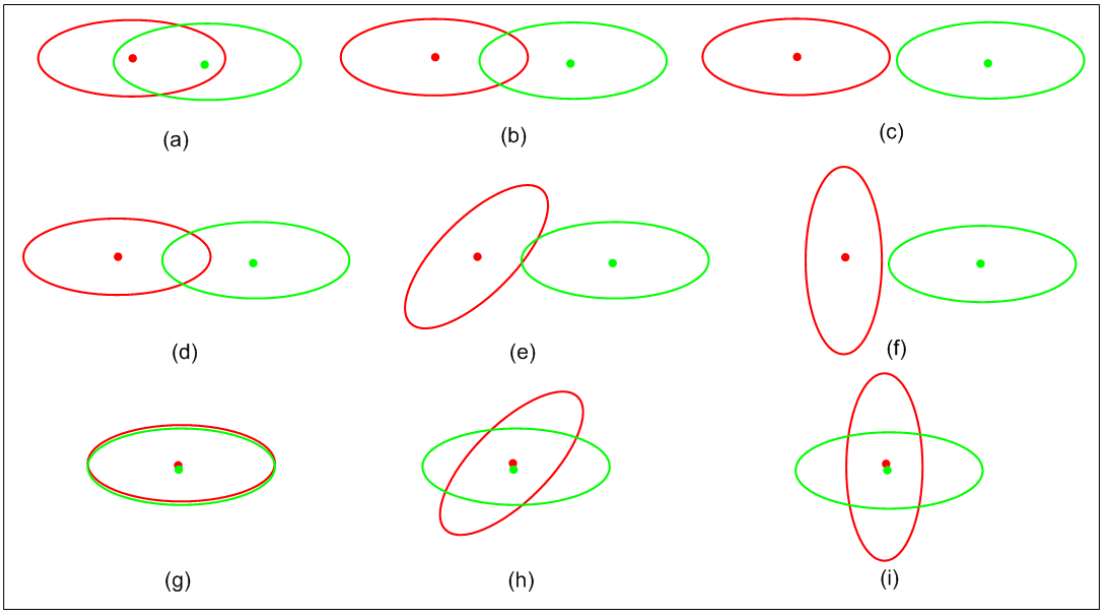


Fig. 1. Separability between two classes of data sets depicted by red color and green color, respectively. (a)–(c) Examples that the separability of two classes improves with the increases of the class mean distances. (d)–(i) Examples that the separability of two classes improves with the increase of the difference of class covariance matrices.

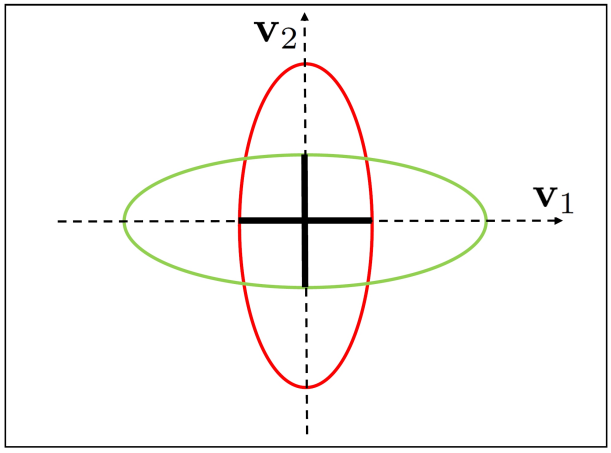


Fig. 2. Example where the class means are equal but the class covariance matrices are different. In this case, the FLDA method is not applicable because the between-class scatter matrix becomes a zero matrix.

where $\mathbf{\Sigma_x}$ and $\mathbf{\Sigma_y}$ denote the covariance matrix of classes 1 and 2, respectively. According to (1), we can obtain that the two most discriminative vectors for distinguishing between the two classes in Fig. 2 should be $\mathbf{v}_1$ and $\mathbf{v}_2$. This is because the projections of the data samples onto these two projection vectors will have the minimal overlapping parts (indicated by thick lines).

Malina [29] proposed an extended Fisher's criterion that is expressed as the similar form of MAGRQ. Unfortunately, Malina's criterion is limited to two-class feature extraction problems, and it is proposed empirically and hence lacks a rigorous theoretical justification. In contrast to Malina's criterion, our MAGRQ criterion is obtained based on a rigorous theoretical derivation. More specifically, we develop an upper bound of Bayes error under single Gaussian distribution assumption of each class data set and then extend to the case of mixture of Gaussian distributions. We also show that minimizing the upper bounds of Bayes error in both cases will result in the similar MAGRQ discriminant criterion, where a larger value of the MAGRQ criterion would lead to a smaller bound of the Bayes error. In addition, it seems that our MAGRQ criterion is also related with the multiview or multimodal learning problems such as the works addressed in [49] and [50], and there are significantly different between them. Specifically, the multiview learning or multimodal learning is mainly targeted at the problems of learning from the data of multiple distinct feature sets. In contrast to these methods, the proposed MAGRQ criterion is developed under a single feature set. Moreover, the multiview learning or multimodal learning methods of [49] and [50] are developed without considering the discriminant information lying in the class covariance differences, whereas the proposed MAGRQ criterion aims to extract this kind of discriminant information.

In dealing with the discriminant analysis problem, it is well known that both between-class scatter distance and within-class scatter distance can be formulated as the ℓ_2-norm operation [30]. Since the ℓ_2-norm is more sensitive to the influence of outliers, discriminant analysis-based ℓ_1-norm had received increasing interests of researchers [30]–[34], [34] in order to boost the robustness of the discriminant analysis methods. Wang *et al.* [30] first introduced the ℓ_1-norm distance metric for learning robust common spatial filters from EEG data samples contaminated by noises. The basic idea was further adopted to deal with the robust discriminative feature extraction of FLDA by Zhong and Zhang [31], Zheng *et al.* [32], and Wang *et al.* [33], respectively, which was referred to as the L1-FLDA method here.

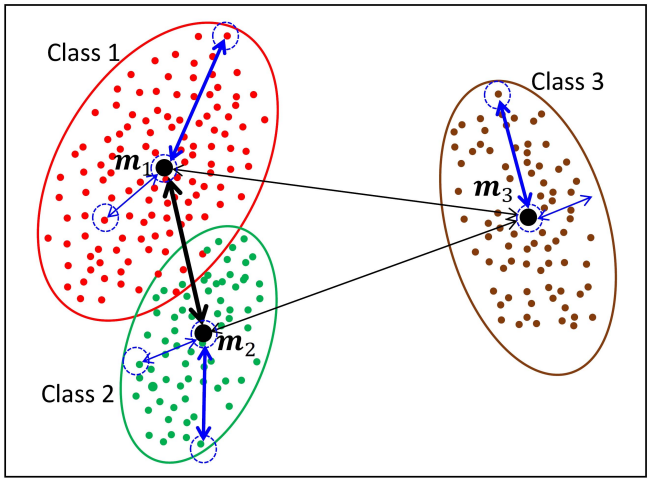


Fig. 3. Example of a set of data set with three classes to illustrate the scenarios that more attentions should be focused on, in which the distances indicted by thicker lines imply that they are more important than those indicated by thinner lines in order to separate the different classes.

Despite the success of L1-FLDA in robust discriminative feature extraction, it is interesting to see that replacing ℓ_2-norm with ℓ_1-norm in the between-class scatter distance would be advantageous to increase the discrimination ability of FLDA, whereas it would not be a good choice to replace the ℓ_2-norm with ℓ_1-norm for the within-class scatter distance. This is because the use of ℓ_1-norm tends to suppressing the contribution of the well-separated classes (with longer between-class scatter distance) and hence can emphasize more on the classes (with smaller between-class scatter distance) that are difficult to be separated. For the within-class scatter distance, we expect to minimize the within-scatter distance so as to achieve better discrimination, which means that we should focus more on the classes with longer within-scatter scatter distances rather than on those classes with smaller within-class scatter distances. In this sense, using the ℓ_2-norm would be more advantageous than the ℓ_1 one to describe the within-class scatter distance. Fig. 3 shows an example of a set of data samples with three classes to illustrate the scenarios that more attentions should be focused on in order to achieve better discrimination, in which the distances indicted by thicker lines imply that they are more important than those indicated by thinner lines in order to separate the different classes. Consequently, to emphasize more on the classes with smaller between-class scatter distance, the ℓ_1-norm could be adopted to describe the between-class scatter distance. On the contrary, to emphasize more on the classes with larger within-class scatter distance, the ℓ_2-norm could be adopted to describe the within-class scatter distance. According to the above-mentioned analysis, we extend the MAGRQ criterion by replacing the ℓ_2-norm with the ℓ_1 one in the between-class scatter distance and hereafter propose the ℓ_1-norm-based MAGRQ (L1-MAGRQ) criterion.

Based on the aforementioned L1-MAGRQ criterion, in this paper, we propose a novel HDA method under the mixture of Gaussian distribution (L1-HDA/GM) of each class data samples. Moreover, we also propose an efficient algorithm to solve the optimal discriminant vector sets of L1-HDA/GM, in which only the principal eigenvalue decomposition problems are involved, which can be efficiently solved by using the power iteration approach and the rank-one-update (ROU) technique [35]. In addition, although L1-HDA/GM can be seen as the extension of our preliminary works in [26]–[28], it improves the previous works by using ℓ_1-norm to replace the ℓ_2 one for describing the between-class scatter distance. Specifically, under the ℓ_1-norm distance metric, the feature extraction of L1-HDA/GM will pay more attention to the nonseparated pairwise classes, which makes it more powerful in extracting the discriminative features.

The remainder of this paper is organized as follows. In Section II, we briefly introduce the Bayes error upper bound under both Gaussian and mixture of Gaussian distributions. In Section III, we develop the MAGRQ criterion based on the Bayes error upper bound estimation and then propose the simplified version of the L1-HDA/GM method as for the case when the number of Gaussian components is fixed at 1. In Section IV, we propose the complete L1-HDA/GM method. The experiments are presented in Section V, and Section VI concludes this paper.

II. Bayes Error Upper Bound Under Gaussian and Mixture of Gaussian Distributions

In this section, we briefly introduce an upper bound of the Bayes error under the single Gaussian distribution assumption and then extend it to the case of mixture of Gaussian distributions, which are the basis of deriving our MAGRQ criterion in Sections III and IV, respectively.

A. Bayes Error Upper Bound Under Single Gaussian Distribution

Suppose that we are given a set of d-dimensional vector set $\mathbf{X} = \{\mathbf{x}_i^j | i = 1, \ldots, c;\ j = 1, \ldots, N_i\}$, where $\mathbf{x}_i^j \in \mathbb{R}^d$ be a sample vector and c and N_i denote the number of classes and the number of data samples in the ith class, respectively. Let $p_i(\mathbf{x})$ and P_i denote the distribution and the prior probability of the ith class, respectively. Assume that the distribution of the ith class is Gaussian, i.e., $p_i(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{m}_i, \mathbf{\Sigma}_i)$, where $\mathcal{N}(\mathbf{x}|\mathbf{m}_i, \mathbf{\Sigma}_i)$ is expressed as

$$\mathcal{N}(\mathbf{x}|\mathbf{m}_i, \mathbf{\Sigma}_i) = \frac{1}{(2\pi)^{\frac{d}{2}}|\mathbf{\Sigma}_i|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T \times \mathbf{\Sigma}_i^{-1}(\mathbf{x}-\mathbf{m}_i)\right\}$$

$\mathbf{m}_i$ and $\mathbf{\Sigma}_i$ denote the class mean and the class covariance matrix, respectively. Then, the Bayes error between the classes i and j can be expressed as [1]

$$\varepsilon = \int \min\left(P_i p_i(\mathbf{x}), P_j p_j(\mathbf{x})\right) d\mathbf{x}. \tag{2}$$

By applying the following inequality to (2):

$$\min(a, b) \le \sqrt{ab} \quad \forall a, b \ge 0, \tag{3}$$

we obtain that the Bayes error can be bounded as the following form [1]:

$$\varepsilon \le \int \sqrt{P_i P_j p_i(\mathbf{x}) p_j(\mathbf{x})}\, d\mathbf{x} \triangleq \sqrt{P_i P_j}\,\varepsilon_{ij} \tag{4}$$

where

$$\varepsilon_{ij} = \int \sqrt{p_i(\mathbf{x}) p_j(\mathbf{x})} d\mathbf{x}. \tag{5}$$

Substituting the expressions of $p_i(\mathbf{x})$ into (5), we obtain that ε_{ij} can be calculated by

$$\varepsilon_{ij} = \exp\left\{-\frac{1}{8}\Delta\mathbf{m}_{ij}^T \bar{\boldsymbol{\Sigma}}_{ij}^{-1} \Delta\mathbf{m}_{ij}\right\} \left(\frac{\sqrt{|\boldsymbol{\Sigma}_i||\boldsymbol{\Sigma}_j|}}{|\bar{\boldsymbol{\Sigma}}_{ij}|}\right)^{\frac{1}{2}} \tag{6}$$

where $\bar{\boldsymbol{\Sigma}}_{ij} = (1/2)(\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j)$ and $\Delta\mathbf{m}_{ij} = \mathbf{m}_i - \mathbf{m}_j$.

B. Bayes Error Upper Bound Under Mixture of Gaussian Distributions

The aforementioned Bayes error upper bound is obtained under the assumption of single Gaussian distribution of $p_i(\mathbf{x})$. Assume that the class probability density function $p_i(\mathbf{x})$ is a mixture of Gaussians, i.e., $p_i(\mathbf{x})$ can be expressed as the form

$$p_i(\mathbf{x}) = \sum_{r=1}^{K_i} \pi_{ir} \mathcal{N}(\mathbf{x}|\mathbf{m}_{ir}, \boldsymbol{\Sigma}_{ir}) \tag{7}$$

where $0 \le \pi_{ir} \le 1$ ($\sum_{r=1}^{K_i} \pi_{ir} = 1$) are called the mixing coefficients and K_i is the number of Gaussian mixture components.

Let $\mathcal{N}(\mathbf{x}|\mathbf{m}_{ir}, \boldsymbol{\Sigma}_{ir}) \triangleq \mathcal{N}_{ir}$. Then, from (2), we obtain that the Bayes error between the classes i and j can be bounded by

$$\begin{aligned}\varepsilon &= \int \min(P_i p_i(\mathbf{x}), P_j p_j(\mathbf{x})) d\mathbf{x} \\ &\le \sum_{r=1}^{K_i}\sum_{l=1}^{K_j} \int \min\{P_i \pi_{ir} \mathcal{N}_{ir}, P_j \pi_{jl} \mathcal{N}_{jl}\} d\mathbf{x} \\ &\le \sum_{r=1}^{K_1}\sum_{l=1}^{K_j} \sqrt{P_i \pi_{ir} P_j \pi_{jl}} \varepsilon_{ij}^{rl}\end{aligned} \tag{8}$$

where

$$\varepsilon_{ij}^{rl} = \exp\left\{-\frac{1}{8}\Delta\mathbf{m}_{ij}^{rl\,T} \left(\bar{\boldsymbol{\Sigma}}_{ij}^{rl}\right)^{-1} \Delta\mathbf{m}_{ij}^{rl}\right\} \left(\frac{\sqrt{|\boldsymbol{\Sigma}_{ir}||\boldsymbol{\Sigma}_{jl}|}}{|\bar{\boldsymbol{\Sigma}}_{ij}^{rl}|}\right)^{\frac{1}{2}} \tag{9}$$

where $\bar{\boldsymbol{\Sigma}}_{ij}^{rl} = (1/2)(\boldsymbol{\Sigma}_{ir} + \boldsymbol{\Sigma}_{jl})$ and $\Delta\mathbf{m}_{ij}^{rl} = \mathbf{m}_{ir} - \mathbf{m}_{jl}$.

In what follows, we will limit our attention to derive the MAGRQ criterion based on the Bayes error bound in (4) and (9), respectively. We first derive the MAGRQ criterion under single Gaussian distribution and then extend to the case of mixture of Gaussian distributions.

III. MAGRQ Criterion for HDA Under Single Gaussian Distribution

In this section, we develop the MAGRQ criterion based on the Bayes error upper bound estimation and then propose a novel HDA method based on this criterion.

A. MAGRQ Criterion Under Single Gaussian Distribution

Assume that the data samples of each class abide by the single Gaussian distribution; then, we obtain that, when projecting the samples to 1-D by a vector $\boldsymbol{\omega} \in \mathbb{R}^d$, the distribution of the projected samples in the ith class data set will become $\tilde{p}_i(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\omega}^T\mathbf{m}_i, \boldsymbol{\omega}^T\boldsymbol{\Sigma}_i\boldsymbol{\omega})$, and the upper bound ε_{ij} becomes

$$\begin{aligned}\varepsilon_{ij}(\omega) &= \exp\left\{-\frac{1}{8}\frac{(\boldsymbol{\omega}^T\Delta\mathbf{m}_{ij})^2}{\boldsymbol{\omega}^T\bar{\boldsymbol{\Sigma}}_{ij}\boldsymbol{\omega}}\right\}\left(\frac{\boldsymbol{\omega}^T\boldsymbol{\Sigma}_i\boldsymbol{\omega}\boldsymbol{\omega}^T\boldsymbol{\Sigma}_j\boldsymbol{\omega}}{(\boldsymbol{\omega}^T\bar{\boldsymbol{\Sigma}}_{ij}\boldsymbol{\omega})^2}\right)^{\frac{1}{4}} \\ &= \exp\left\{-\frac{1}{8}\frac{(\boldsymbol{\omega}^T\Delta\mathbf{m}_{ij})^2}{\boldsymbol{\omega}^T\bar{\boldsymbol{\Sigma}}_{ij}\boldsymbol{\omega}}\right\}\left(1 - \left(\frac{\omega^T\Delta\boldsymbol{\Sigma}_{ij}\omega}{\boldsymbol{\omega}^T\bar{\boldsymbol{\Sigma}}_{ij}\boldsymbol{\omega}}\right)^2\right)^{\frac{1}{4}}\end{aligned} \tag{10}$$

where $\Delta\mathbf{m}_{ij} = \mathbf{m}_i - \mathbf{m}_j$ and $\Delta\boldsymbol{\Sigma}_{ij} = (1/2)(\boldsymbol{\Sigma}_i - \boldsymbol{\Sigma}_j)$.

To minimize the Bayes error, we should minimize its upper bound. Hence, based on (4) and (10), we should maximize both $((\boldsymbol{\omega}^T\Delta\mathbf{m}_{ij})^2/\boldsymbol{\omega}^T\bar{\boldsymbol{\Sigma}}_{ij}\boldsymbol{\omega})$ and $(|\omega^T\Delta\boldsymbol{\Sigma}_{ij}\omega|/\boldsymbol{\omega}^T\bar{\boldsymbol{\Sigma}}_{ij}\boldsymbol{\omega})$, which results in the following two-class heteroscedastic discriminant criterion:

$$J_{ij}(\omega) = \frac{(\boldsymbol{\omega}^T\Delta\mathbf{m}_{ij})^2 + |\omega^T\Delta\boldsymbol{\Sigma}_{ij}\omega|}{\boldsymbol{\omega}^T\bar{\boldsymbol{\Sigma}}_{ij}\boldsymbol{\omega}}. \tag{11}$$

We call the criterion of (11) as the *pairwise* MAGRQ criterion. This criterion is the Malina's discriminant criterion [29] for two-class feature extraction. From the definition of $J_{ij}(\omega)$ in (11), we can see that the two-class MAGRQ criterion can be seen as the mixture of Fisher's criterion and Fukunaga–Koontz criterion [36], in which the first part corresponds to Fisher's criterion, whereas the later one corresponds to the Fukunaga–Koontz criterion.

On the other hand, from the expression of (11), we obtain that the physical meaning of the MAGRQ criterion can be explained as the simultaneous optimization of the following two parts:

$$\begin{cases}\max(\boldsymbol{\omega}^T\Delta\mathbf{m}_{ij})^2 + |\omega^T\Delta\boldsymbol{\Sigma}_{ij}\omega| \\ \min \boldsymbol{\omega}^T\bar{\boldsymbol{\Sigma}}_{ij}\boldsymbol{\omega}.\end{cases} \tag{12}$$

For multiclass cases, the maximization problem of (12) can be extended by maximizing the pairwise summation of the two parts of (12), that is

$$\begin{cases}\max \sum_{i,j} P_i P_j[(\boldsymbol{\omega}^T\Delta\mathbf{m}_{ij})^2 + |\omega^T\Delta\boldsymbol{\Sigma}_{ij}\omega|] \\ \min \sum_{i,j} P_i P_j \boldsymbol{\omega}^T\bar{\boldsymbol{\Sigma}}_{ij}\boldsymbol{\omega} = \min 2\boldsymbol{\omega}^T\bar{\boldsymbol{\Sigma}}\boldsymbol{\omega}\end{cases} \tag{13}$$

where $\bar{\boldsymbol{\Sigma}} = \sum_{i=1}^{c} P_i\boldsymbol{\Sigma}_i$.

From (13), we define the multiclass MAGRQ criterion as the following form:

$$J(\omega) = \frac{\|\boldsymbol{\omega}^T\mathbf{B}\|_2^2 + \sum_{i<j} P_{ij}|\omega^T\Delta\boldsymbol{\Sigma}_{ij}\omega|}{\boldsymbol{\omega}^T\bar{\boldsymbol{\Sigma}}\boldsymbol{\omega}} \tag{14}$$

where $P_{ij} = P_i P_j$, and

$$\begin{aligned}\mathbf{B} = [&\sqrt{P_{12}}\Delta\mathbf{m}_{12}, \ldots, \sqrt{P_{1c}}\Delta\mathbf{m}_{1c} \\ &\times\sqrt{P_{23}}\Delta\mathbf{m}_{23}, \ldots, \sqrt{P_{(c-1)c}}\Delta\mathbf{m}_{(c-1)c}].\end{aligned} \tag{15}$$

The between-class scatter distance of the multiclass MAGRQ criterion in (14) is based on ℓ_2-norm, and hence, it is referred to as the ℓ_2-norm-based MAGRQ (L2-MAGRQ) criterion.

On the other hand, as what we have pointed out in Section II, using ℓ_1-norm would be more advantageous than ℓ_2-norm in separating the different classes. Hence, we use ℓ_1-norm to replace the ℓ_2 one for describing the between-class scatter distance of $J(\omega)$, which means that we use $\|\boldsymbol{\omega}^T\mathbf{B}\|_1^2$ to replace $\|\boldsymbol{\omega}^T\mathbf{B}\|_2^2$ in the nominator part of $J(\omega)$, resulting in the following ℓ_1-norm-based MAGRQ (L1-MAGRQ) criterion:

$$J_1(\omega) = \frac{\|\boldsymbol{\omega}^T\mathbf{B}\|_1^2 + \sum_{i<j} P_{ij}|\omega^T \Delta\boldsymbol{\Sigma}_{ij}\omega|}{\boldsymbol{\omega}^T\bar{\boldsymbol{\Sigma}}\boldsymbol{\omega}}. \tag{16}$$

Based on the L2-MAGRQ criterion defined in (14) and L1-MAGRQ criterion defined in (16), we can develop two HDA methods under the Gaussian distribution, which are, respectively, denoted by L2-HDA/G and L1-HDA/G. In what follows, we will first provide the detailed algorithm description of the L1-HDA/G method in Section III-B and then address the L2-HDA/G algorithm based on the L1-HDA/G algorithm in Section III-C.

B. L1-HDA/G Algorithm

Suppose that we want to obtain k discriminant vectors, denoted as $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_k$, of L1-HDA/G. Then, we subsequently define the k discriminant vectors as follows.

Let $\omega_1, \ldots, \omega_r$ be the first r discriminant vectors. Then, the $(r+1)$th discriminant vector is defined by

$$\omega_{r+1} = \arg\max_{\omega} J_1(\boldsymbol{\omega}), \quad \text{s.t. } \omega^T\mathbf{S}_t\omega_j = 0 \quad \forall\ j \le r \tag{17}$$

where $\mathbf{S}_t$ is the covariance matrix of all data samples, such that the discriminant vectors are statistically uncorrelated [37].

Let $\boldsymbol{\omega} = \bar{\boldsymbol{\Sigma}}^{-(1/2)}\boldsymbol{\alpha}$ and

$$\begin{cases} \Delta\hat{\boldsymbol{\Sigma}}_{ij} = P_{ij}\bar{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\Delta\boldsymbol{\Sigma}_{ij}\bar{\boldsymbol{\Sigma}}^{-\frac{1}{2}} \\ \hat{\mathbf{B}} = \bar{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\mathbf{B}. \end{cases} \tag{18}$$

Then, we obtain that solving the optimization problem (17) is equivalent to solving the following optimization problem:

$$\alpha_{r+1} = \arg\max_{\alpha} \hat{J}_1(\boldsymbol{\alpha}), \quad \text{s.t. } \alpha^T\mathbf{U}_r = \mathbf{0}^T \tag{19}$$

where $\mathbf{U}_r = [\hat{\mathbf{S}}_t\alpha_1, \ldots, \hat{\mathbf{S}}_t\alpha_r]$, $\hat{\mathbf{S}}_t = \bar{\boldsymbol{\Sigma}}^{-(1/2)}\mathbf{S}_t\bar{\boldsymbol{\Sigma}}^{-(1/2)}$, and

$$\hat{J}_1(\alpha) = \frac{\|\boldsymbol{\alpha}^T\hat{\mathbf{B}}\|_1^2 + \sum_{i<j}|\alpha^T\Delta\hat{\boldsymbol{\Sigma}}_{ij}\alpha|}{\boldsymbol{\alpha}^T\boldsymbol{\alpha}}. \tag{20}$$

The absolute value signs in the expression of $\hat{J}_1(\alpha)$ make the optimization of (20) difficult. So we introduce two $c \times c$ skew symmetric sign matrices $\mathbf{U} = ((\mathbf{U})_{ij})_{c\times c}$ and $\mathbf{V} = ((\mathbf{V})_{ij})_{c\times c}$, where $(\mathbf{U})_{ij}, (\mathbf{V})_{ij} \in \{+1, -1\}$, where $(\mathbf{U})_{ij}$ and $(\mathbf{V})_{ij}$ denote the ith row and jth column element of $\mathbf{U}$ and $\mathbf{V}$, respectively. Let $\mathbf{u}$ denote the vector by concatenating entries of $\mathbf{U}$ according to the following form:

$$\mathbf{u} = [(\mathbf{U})_{12}, \ldots, (\mathbf{U})_{1c}, (\mathbf{U})_{23}, \ldots, (\mathbf{U})_{(c-1)c}]^T.$$

Denote $\boldsymbol{\Omega}$ as the set of all sign matrices and define

$$\mathbf{T}(\mathbf{U}, \mathbf{V}) = \mathbf{B}\mathbf{u}\mathbf{u}^T\mathbf{B}^T + \sum_{i<j}(\mathbf{V})_{ij}\Delta\hat{\boldsymbol{\Sigma}}_{ij}. \tag{21}$$

Then, we obtain that

$$\begin{aligned} \boldsymbol{\alpha}^T\mathbf{T}(\mathbf{U}, \mathbf{V})\boldsymbol{\alpha} &= (\boldsymbol{\alpha}^T\mathbf{B}\mathbf{u})^2 + \sum_{i<j}(\mathbf{V})_{ij}\boldsymbol{\alpha}^T\Delta\hat{\boldsymbol{\Sigma}}_{ij}\boldsymbol{\alpha} \\ &\le \|\boldsymbol{\alpha}^T\mathbf{B}\|_1^2 + \sum_{i<j}|\boldsymbol{\alpha}^T\Delta\hat{\boldsymbol{\Sigma}}_{ij}\boldsymbol{\alpha}|. \end{aligned} \tag{22}$$

From (22), we obtain that the optimization problem of (20) can be formulated as the following one:

$$\hat{J}_1(\boldsymbol{\alpha}) = \max_{\mathbf{U},\mathbf{V}\in\boldsymbol{\Omega}, \|\alpha\|=1} \boldsymbol{\alpha}^T\mathbf{T}(\mathbf{U}, \mathbf{V})\boldsymbol{\alpha}. \tag{23}$$

From (23), we obtain that

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \hat{J}_1(\boldsymbol{\alpha}) &= \max_{\|\boldsymbol{\alpha}\|=1} \max_{\mathbf{U},\mathbf{V}\in\boldsymbol{\Omega}} \boldsymbol{\alpha}^T\mathbf{T}(\mathbf{U}, \mathbf{V})\boldsymbol{\alpha} \\ &= \max_{\mathbf{U},\mathbf{V}\in\boldsymbol{\Omega}} \max_{\|\boldsymbol{\alpha}\|=1} \boldsymbol{\alpha}^T\mathbf{T}(\mathbf{U}, \mathbf{V})\boldsymbol{\alpha}. \end{aligned} \tag{24}$$

By observing (24), we can see that: if the sign matrices $\mathbf{U}$ and $\mathbf{V}$ are fixed, then the optimal discriminant vector is the *normalized* (we will not reemphasize this in the sequel) eigenvector associated with the largest eigenvalue of the matrix $\mathbf{T}(\mathbf{U}, \mathbf{V})$. Solving the principal eigenvector of $\mathbf{T}(\mathbf{U}, \mathbf{V})$ can be easily realized via the power iteration method. So our problem of maximizing $\hat{J}_1(\boldsymbol{\alpha})$ is changed to finding the optimal sign matrices $\mathbf{U}$ and $\mathbf{V}$, such that the largest eigenvalue of $\mathbf{T}(\mathbf{U}, \mathbf{V})$ is maximized.

In what follows, we will propose a greedy algorithm to find the suboptimal sign matrices $\mathbf{U}$ and $\mathbf{V}$. To begin with, we introduce the following theorem.

Theorem 1: Let $\alpha^{(1)}$ be the principal eigenvector of $\mathbf{T}(\mathbf{U}_1, \mathbf{V}_1)$. Define $\mathbf{U}_2$ and $\mathbf{V}_2$ as

$$\begin{cases} (\mathbf{U}_2)_{ij} = \text{sign}(\alpha^{(1)^T}\Delta\hat{\mathbf{m}}_{ij}), \\ (\mathbf{V}_2)_{ij} = \text{sign}(\alpha^{(1)^T}\Delta\hat{\boldsymbol{\Sigma}}_{ij}\alpha^{(1)}) \end{cases} \tag{25}$$

where

$$\text{sign}(a) = \begin{cases} 1, & \text{if } a \ge 0 \\ -1, & \text{Others.} \end{cases}$$

Suppose that $\alpha^{(2)}$ is the principal eigenvector of $\mathbf{T}(\mathbf{U}_2, \mathbf{V}_2)$. Then, we have

$$\alpha^{(2)^T}\mathbf{T}(\mathbf{U}_2, \mathbf{V}_2)\alpha^{(2)} \ge \alpha^{(1)^T}\mathbf{T}(\mathbf{U}_1, \mathbf{V}_1)\alpha^{(1)}. \tag{26}$$

Proof: See Appendix A. □

Thanks to Theorem 1, we are able to improve the sign matrices step by step.

To solve the discriminant vector $\boldsymbol{\alpha}_{r+1}$, we introduce Propositions 1 and 2 in the following. Their proofs can be easily obtained from [38].

Proposition 1: Let $\mathbf{Q}_r\mathbf{R}_r$ be the QR decomposition of $\mathbf{U}_r$, where $\mathbf{R}_r$ is an $r \times r$ upper triangular matrix. Then, $\boldsymbol{\alpha}_{r+1}$ defined in (19) is the principal eigenvector corresponding to the largest eigenvalue of the following matrix $(\mathbf{I}_d - \mathbf{Q}_r\mathbf{Q}_r^T)\mathbf{T}(\mathbf{U}, \mathbf{V})(\mathbf{I}_d - \mathbf{Q}_r\mathbf{Q}_r^T)$.

Proposition 2: Suppose that $\mathbf{Q}_r\mathbf{R}_r$ is the QR decomposition of $\mathbf{U}_r$. Let $\mathbf{U}_{r+1} = (\mathbf{U}_r \ \ \hat{\mathbf{S}}_t\boldsymbol{\alpha}_{r+1})$, $\mathbf{q} = \hat{\mathbf{S}}_t\boldsymbol{\alpha}_{r+1} - \mathbf{Q}_r(\mathbf{Q}_r^T\hat{\mathbf{S}}_t\boldsymbol{\alpha}_{r+1})$, and $\mathbf{Q}_{r+1} = (\mathbf{Q}_r \ \ (\mathbf{q}/\|\mathbf{q}\|))$. Then, $\mathbf{Q}_{r+1}\begin{pmatrix} \mathbf{R}_r & \mathbf{Q}_r^T\hat{\mathbf{S}}_t\boldsymbol{\alpha}_{r+1} \\ \mathbf{0} & \|\mathbf{q}\| \end{pmatrix}$ is the QR decomposition of $\mathbf{U}_{r+1}$.

The above-mentioned two propositions provide an efficient approach for solving (19): Proposition 1 makes it possible to use the power method to solve (19), while Proposition 2 makes it possible to update $\mathbf{Q}_{r+1}$ from $\mathbf{Q}_r$ by adding a single column. Moreover, it should be noted that

$$\begin{aligned}\mathbf{I}_d - \mathbf{Q}_{r+1}\mathbf{Q}_{r+1}^T &= \prod_{i=1}^{r+1}\left(\mathbf{I}_d - \mathbf{q}_i\mathbf{q}_i^T\right) \\ &= \left(\mathbf{I}_d - \mathbf{Q}_r\mathbf{Q}_r^T\right)\left(\mathbf{I}_d - \mathbf{q}_{r+1}\mathbf{q}_{r+1}^T\right) \quad (27)\end{aligned}$$

where $\mathbf{q}_i$ is the ith column of $\mathbf{Q}_r$. Equation (27) makes it possible to update $(\mathbf{I}_d-\mathbf{Q}_{r+1}\mathbf{Q}_{r+1}^T)\mathbf{T}(\mathbf{U}_r, \mathbf{V}_r)(\mathbf{I}_d-\mathbf{Q}_{r+1}\mathbf{Q}_{r+1}^T)$ from $(\mathbf{I}_d - \mathbf{Q}_r\mathbf{Q}_r^T)\mathbf{T}(\mathbf{U}_r, \mathbf{V}_r)(\mathbf{I}_d - \mathbf{Q}_r\mathbf{Q}_r^T)$ by the ROU technique.

Here, it should be noted that the initial setting of the sign matrices $\mathbf{U}$ and $\mathbf{V}$ in $\mathbf{T}(\mathbf{U}, \mathbf{V})$ may influence the optimality of the solution. Consequently, to obtain a better solution, we may initialize the sign matrices $\mathbf{U}$ and $\mathbf{V}$ based on the optimal discriminant vectors solved by other discriminant analysis algorithms. For example, suppose that α is the optimal discriminant vector solved by the HDA/Chernoff algorithm [15], then the initial sign matrix of $\mathbf{U}$ and $\mathbf{V}$ can be obtained as follows:

$$(\mathbf{U})_{pq} \leftarrow \text{sign}(\alpha^T \Delta\hat{\mathbf{m}}_{pq}) \quad (28)$$

$$(\mathbf{V})_{pq} \leftarrow \text{sign}(\alpha^T \Delta\hat{\mathbf{\Sigma}}_{pq}\alpha). \quad (29)$$

We summarize the algorithm for solving the first k discriminant vectors of L1-HDA/G in Algorithm 1.

C. L2-HDA/G Algorithm

In the aforementioned section, we had developed a set of optimal discriminative vectors of L1-HDA/G based on the L1-MAGRQ criterion. Similarly, if the L2-MAGRQ criterion is adopted, then we can obtain an optimal set of L2-HDA/G discriminative vectors. Specifically, the optimal discriminative vectors of L2-HDA/G can be subsequently obtained: suppose that we have obtained the first r optimal discriminant vectors of L2-HDA/G, denoted as $\omega_1, \ldots, \omega_r$, then the $(r+1)$th discriminant vector is defined by

$$\omega_{r+1} = \arg\max_{\omega} J(\boldsymbol{\omega}), \quad \text{s.t. } \omega^T\mathbf{S}_t\omega_j = 0 \quad \forall j \le r. \quad (30)$$

The optimization problem of (30) is equivalent of the following one:

$$\alpha_{r+1} = \arg\max_{\alpha^T\mathbf{U}=\mathbf{0}^T} \hat{J}(\boldsymbol{\alpha}) \quad (31)$$

where

$$\hat{J}(\alpha) = \frac{\|\boldsymbol{\alpha}^T\hat{\mathbf{B}}\|_2^2 + \sum_{i<j}|\alpha^T\Delta\hat{\mathbf{\Sigma}}_{ij}\alpha|}{\boldsymbol{\alpha}^T\boldsymbol{\alpha}} \quad (32)$$

in which $\mathbf{U}_r$ and $\hat{\mathbf{B}}$ are defined in Section III-B.

Noting that the ℓ_2-norm metric can be easily computed without the absolute value operation, the expression of $\mathbf{T}(\mathbf{U}, \mathbf{V})$ in (21) can be replaced by $\mathbf{T}(\mathbf{V})$ defined as follows:

$$\mathbf{T}(\mathbf{V}) = \mathbf{B}\mathbf{B}^T + \sum_{i<j}(\mathbf{V})_{ij}\Delta\hat{\mathbf{\Sigma}}_{ij}. \quad (33)$$

As a result, the L2-HDA/G algorithm can be obtained with a simple modification of the L1-HDA/G algorithm shown in Algorithm 1, which can be summarized in Algorithm 2.

Algorithm 1 Solving Optimal Vectors $\boldsymbol{\omega}_i$ $(i = 1, \ldots, k)$ of L1-HDA/G

Input:

- Input data set $\{\mathbf{x}_i^j | i = 1, \ldots, c;\ j = 1, \ldots, N_i\}$, class label vector $\mathbf{L}$, where $N_1 + \cdots + N_c = N$.

Initialization:

- Compute $\mathbf{\Sigma}_i$, $\bar{\mathbf{\Sigma}}_{ij}$, $\bar{\mathbf{\Sigma}}$, $\mathbf{B}$, $\mathbf{S}^t$, $\hat{\mathbf{\Sigma}}_i = \bar{\mathbf{\Sigma}}^{-(1/2}\mathbf{\Sigma}_i\bar{\mathbf{\Sigma}}^{-\frac{1}{2}}$, $\hat{\mathbf{B}} = \bar{\mathbf{\Sigma}}^{-\frac{1}{2}}\mathbf{B}$, $\hat{\mathbf{S}}_t = \bar{\mathbf{\Sigma}}^{-\frac{1}{2}}\mathbf{S}_t\bar{\mathbf{\Sigma}}^{-\frac{1}{2}}$, $\mathbf{m}_i$, $\hat{\mathbf{m}}_i = \bar{\mathbf{\Sigma}}^{-\frac{1}{2}}\mathbf{m}_i$;
- Initialize the discriminant vectors α_i;

For $i = 1, 2, \ldots, k$**, Do**

1) Compute $\Delta\hat{\mathbf{\Sigma}}_{pq} \leftarrow \frac{\hat{\mathbf{\Sigma}}_p - \hat{\mathbf{\Sigma}}_q}{2}$, $(p < q)$;
2) Set $\mathbf{U}$, $\mathbf{V} \leftarrow$ zero matrix, and compute
$$\begin{cases}(\mathbf{U}_1)_{pq} \leftarrow \text{sign}(\alpha_i{}^T\Delta\hat{\mathbf{m}}_{pq}); \\ (\mathbf{V}_1)_{pq} \leftarrow \text{sign}(\alpha_i{}^T\Delta\hat{\mathbf{\Sigma}}_{pq}\alpha_i).\end{cases}$$
3) **While** $\mathbf{U} \neq \mathbf{U}_1$ **and** $\mathbf{V} \neq \mathbf{V}_1$**, Do**
   a) Set $\mathbf{U} \leftarrow \mathbf{U}_1$ and $\mathbf{V} \leftarrow \mathbf{V}_1$, compute $\mathbf{T}(\mathbf{U}, \mathbf{V})$ and the principal eigenvector, α_i, of $\mathbf{T}(\mathbf{U}, \mathbf{V})$;
   b) Compute $\begin{cases}(\mathbf{U}_1)_{pq} \leftarrow \text{sign}(\alpha_i{}^T\Delta\hat{\mathbf{m}}_{pq}); \\ (\mathbf{V}_1)_{pq} \leftarrow \text{sign}(\alpha_i{}^T\Delta\hat{\mathbf{\Sigma}}_{pq}\alpha_i).\end{cases}$
4) Update $\mathbf{Q}_i$: $\mathbf{Q}_i \leftarrow (\mathbf{Q}_{i-1} \quad \frac{\mathbf{q}_i}{\|\mathbf{q}_i\|_2})$, where
$$\begin{cases}\mathbf{q}_1 \leftarrow \hat{\mathbf{S}}_t\boldsymbol{\alpha}_1 \text{ and } \mathbf{Q}_1 \leftarrow \frac{\mathbf{q}_1}{\|\mathbf{q}_1\|_2}, & \text{if i=1;} \\ \mathbf{q}_i \leftarrow \hat{\mathbf{S}}_t\boldsymbol{\alpha}_i - \mathbf{Q}_{i-1}(\mathbf{Q}_{i-1}^T\hat{\mathbf{S}}_t\boldsymbol{\alpha}_i), & \text{otherwise.}\end{cases}$$
5) Update $\hat{\mathbf{\Sigma}}_p$ and $\hat{\mathbf{B}}$:
$$\hat{\mathbf{\Sigma}}_p \leftarrow (\mathbf{I} - \mathbf{q}_i\mathbf{q}_i^T)\hat{\mathbf{\Sigma}}_p(\mathbf{I} - \mathbf{q}_i\mathbf{q}_i^T), \hat{\mathbf{B}} \leftarrow (\mathbf{I} - \mathbf{q}_i\mathbf{q}_i^T)\hat{\mathbf{B}};$$
6) Compute $\boldsymbol{\omega}_i = \bar{\mathbf{\Sigma}}^{-\frac{1}{2}}\boldsymbol{\alpha}_i$, and set $\boldsymbol{\omega}_i \leftarrow \boldsymbol{\omega}_i/\|\boldsymbol{\omega}_i\|$;

Output: $\boldsymbol{\omega}_1, \ldots, \omega_k$.

D. Computational Analysis of L1-HDA/G and L2-HDA/G

According to the detailed algorithm description of L1-HDA/G shown in Algorithm 1, we can obtain the computational complexity of L1-HDA/G algorithm. Specifically, in the initialization part, the computational complexity of calculating the class covariance matrices is $O(cd^2N)$, and the computational complexity of calculating the transformation matrices ($\hat{\mathbf{\Sigma}}_i$, $\hat{\mathbf{B}}$, and $\hat{\mathbf{S}}_t$) is $O(cd^3)$. In calculating each discriminative vector ω_i, the computational complexity of steps 1) and 2) are $O(d^2)$ and $O(c^2d^2)$, respectively. The computational complexity of calculating $\mathbf{T}(\mathbf{U}, \mathbf{V})$ in step 3) is $O(c^2d)+O(c^2d^2)$, and the complexity of solving the principal eigenvector of $\mathbf{T}(\mathbf{U}, \mathbf{V})$ is only $O(d^2)$ (e.g., using the power method). The complexity of updating $\mathbf{U}$ and $\mathbf{V}$ in step 3) is $O(c^2d)+O(c^2d^2)$. In addition, it is easy to check that the computational complexity of steps 4)–6) are $O(d^2)$, $O(d^2)$, and $O(d^2)$, respectively. According to the aforementioned analysis, we summarize the computational complexity of Algorithm 1 in Table I.

In contrast to the L1-HDA/G algorithm, the major difference of L2-HDA/G lies in the calculation of $\mathbf{T}(\mathbf{U}, \mathbf{V})$ (replaced by $\mathbf{T}(\mathbf{V})$ in L2-HDA/G). The computational complexity of

TABLE I
COMPARISON OF COMPUTATIONAL COMPLEXITY BETWEEN THE L1-HDA/G AND L2-HDA/G ALGORITHMS

Algorithm	Computational complexity of each step in the algorithm						
	Initialization	1)	2)	3)	4)	5)	6)
Algorithm 1	$O(cd^2N)$+$O(cd^3)$	$O(c^2)$	$O(c^2d^2)$	$O(c^2d) + O(c^2d^2)$	$O(d^2)$	$O(d^2)$	$O(d^2)$
Algorithm 2	$O(cd^2N)$+$O(cd^3)$	$O(c^2)$	$O(c^2d^2)$	$O(d^3)$+$O(c^2d^2)$	$O(d^2)$	$O(d^2)$	$O(d^2)$

Algorithm 2 Solving Optimal Vectors $\boldsymbol{\omega}_i$ $(i = 1, \ldots, k)$ of L2-HDA/G

Input:

- Input data set $\{\mathbf{x}_i^j | i = 1, \ldots, c;\ j = 1, \ldots, N_i\}$, class label vector $\mathbf{L}$, where $N_1 + \cdots + N_c = N$.

Initialization:

- Compute $\boldsymbol{\Sigma}_i$, $\bar{\boldsymbol{\Sigma}}_{ij}$, $\bar{\boldsymbol{\Sigma}}$, $\mathbf{B}$, $\mathbf{S}^t$, $\hat{\boldsymbol{\Sigma}}_i = \bar{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_i\bar{\boldsymbol{\Sigma}}^{-\frac{1}{2}}$, $\hat{\mathbf{B}} = \bar{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\mathbf{B}$, $\hat{\mathbf{S}}_t = \bar{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\mathbf{S}_t\bar{\boldsymbol{\Sigma}}^{-\frac{1}{2}}$, $\mathbf{m}_i$, $\hat{\mathbf{m}}_i = \bar{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\mathbf{m}_i$;
- Initialize the discriminant vectors α_i;

For $i = 1, 2, \ldots, k$**, Do**

1) Compute $\Delta\hat{\boldsymbol{\Sigma}}_{pq} \leftarrow \frac{\hat{\boldsymbol{\Sigma}}_p - \hat{\boldsymbol{\Sigma}}_q}{2}$, $(p < q)$;
2) Set $\mathbf{V} \leftarrow$ zero matrix, and compute
$$(\mathbf{V}_1)_{pq} \leftarrow \text{sign}(\alpha_i{}^T\Delta\hat{\boldsymbol{\Sigma}}_{pq}\alpha_i).$$
3) **While** $\mathbf{V} \neq \mathbf{V}_1$**, Do**
   a) Set $\mathbf{V} \leftarrow \mathbf{V}_1$ and compute $\mathbf{T}(\mathbf{V})$ and the principal eigenvector, α_i, of $\mathbf{T}(\mathbf{V})$;
   b) Compute $(\mathbf{V}_1)_{pq} \leftarrow \text{sign}(\alpha_i{}^T\Delta\hat{\boldsymbol{\Sigma}}_{pq}\alpha_i)$.
4) Update $\mathbf{Q}_i$: $\mathbf{Q}_i \leftarrow (\mathbf{Q}_{i-1} \quad \frac{\mathbf{q}_i}{\|\mathbf{q}_i\|_2})$, where
$$\begin{cases} \mathbf{q}_1 \leftarrow \hat{\mathbf{S}}_t\boldsymbol{\alpha}_1 \text{ and } \mathbf{Q}_1 \leftarrow \frac{\mathbf{q}_1}{\|\mathbf{q}_1\|_2}, & \text{if i=1;} \\ \mathbf{q}_i \leftarrow \hat{\mathbf{S}}_t\boldsymbol{\alpha}_i - \mathbf{Q}_{i-1}(\mathbf{Q}_{i-1}^T\hat{\mathbf{S}}_t\boldsymbol{\alpha}_i), & \text{otherwise.} \end{cases}$$
5) Update $\hat{\boldsymbol{\Sigma}}_p$ and $\hat{\mathbf{B}}$:
$$\hat{\boldsymbol{\Sigma}}_p \leftarrow (\mathbf{I} - \mathbf{q}_i\mathbf{q}_i^T)\hat{\boldsymbol{\Sigma}}_p(\mathbf{I} - \mathbf{q}_i\mathbf{q}_i^T), \hat{\mathbf{B}} \leftarrow (\mathbf{I} - \mathbf{q}_i\mathbf{q}_i^T)\hat{\mathbf{B}};$$
6) Compute $\boldsymbol{\omega}_i = \bar{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{\alpha}_i$, and set $\boldsymbol{\omega}_i \leftarrow \boldsymbol{\omega}_i/\|\boldsymbol{\omega}_i\|$;

Output: $\boldsymbol{\omega}_1, \ldots, \omega_k$.

calculating $\mathbf{T}(\mathbf{V})$ is $O(d^3) + O(c^2d^2)$, which would be a bit more than calculating the value of $\mathbf{T}(\mathbf{U}, \mathbf{V})$ in the L1-HDA/G algorithm. The detailed computational complexity of L2-HDA/G is also summarized in Table I.

IV. L1-MAGRQ CRITERION UNDER MIXTURE OF GAUSSIAN DISTRIBUTIONS AND L1-HDA/GM

In this section, we generalize the L2-MAGRQ criterion and the L1-MAGRQ criterion from the single Gaussian distribution to the mixture of Gaussian distributions. Then, we propose the L2-HDA/GM method and the L1-HDA/GM method. If the data samples of each class abide by the mixture of Gaussian distributions, then we have the following theorem with respect to the projected samples.

Theorem 2: Suppose that the distribution function of the ith class is a mixture of Gaussians, that is
$$p_i(\mathbf{x}) = \sum_{r=1}^{K_i}\pi_{ir}\mathcal{N}(\mathbf{x}|\mathbf{m}_{ir}, \boldsymbol{\Sigma}_{ir}).$$
Then, the class distribution function $\tilde{p}_i(\omega^T\mathbf{x})$ of the projected samples $\omega^T\mathbf{x}$ is also a mixture of Gaussians, that is
$$\tilde{p}_i(\omega^T\mathbf{x}) = \sum_{r=1}^{K_i}\pi_{ir}\mathcal{N}(\omega^T\mathbf{x}|\omega^T\mathbf{m}_{ir}, \omega^T\boldsymbol{\Sigma}_{ir}\omega). \tag{34}$$

Proof: See Appendix B. □

Thanks to Theorem 2, we obtain that the two-class Bayes error bound expressed in (8) can be replaced by
$$\varepsilon \leq \sum_{r=1}^{K_1}\sum_{l=1}^{K_j}\sqrt{P_i\pi_{ir}P_j\pi_{jl}}\varepsilon_{ij}^{rl}(\omega) \tag{35}$$
where $\varepsilon_{ij}^{rl}(\omega)$ is formulated as
$$\varepsilon_{ij}^{rl}(\omega) = \exp\left\{-\frac{1}{8}\frac{\left(\boldsymbol{\omega}^T\Delta\mathbf{m}_{ij}^{rl}\right)^2}{\boldsymbol{\omega}^T\bar{\boldsymbol{\Sigma}}_{ij}^{rl}\boldsymbol{\omega}}\right\}\left(1 - \left(\frac{\omega^T\Delta\boldsymbol{\Sigma}_{ij}^{rl}\omega}{\boldsymbol{\omega}^T\bar{\boldsymbol{\Sigma}}_{ij}^{rl}\boldsymbol{\omega}}\right)^2\right)^{\frac{1}{4}}. \tag{36}$$

Similar to the derivation in Section III, from the Bayes error upper bound shown in (35) and (36), we obtain the following two-class MAGRQ criterion under the mixture of Gaussian distributions:
$$J_{ij}(\omega) = \sum_{r,l}\pi_{ir}\pi_{jl}\frac{\left(\boldsymbol{\omega}^T\Delta\mathbf{m}_{ij}^{rl}\right)^2 + \left|\omega^T\Delta\boldsymbol{\Sigma}_{ij}^{rl}\omega\right|}{\boldsymbol{\omega}^T\bar{\boldsymbol{\Sigma}}_{ij}^{rl}\boldsymbol{\omega}}. \tag{37}$$
Note that, in real applications, the number of samples is often insufficient for estimating a mixture of Gaussians with different $\boldsymbol{\Sigma}_{ir}$ values. To remedy this issue, we may assume that the matrices $\{\boldsymbol{\Sigma}_{ir}\}$ are identical, say equal to $\boldsymbol{\Sigma}_i$. In this case, the two-class MAGRQ criterion in (37) can be expressed as the following form as for the case of the mixture of Gaussian distribution:
$$J_{ij}(\omega) = \frac{\sum_{r,l}\pi_{ir}\pi_{jl}\left(\boldsymbol{\omega}^T\Delta\mathbf{m}_{ij}^{rl}\right)^2}{\boldsymbol{\omega}^T\bar{\boldsymbol{\Sigma}}_{ij}\boldsymbol{\omega}} + \frac{|\omega^T\Delta\boldsymbol{\Sigma}_{ij}\omega|}{\boldsymbol{\omega}^T\bar{\boldsymbol{\Sigma}}_{ij}\boldsymbol{\omega}} \tag{38}$$
where $\Delta\boldsymbol{\Sigma}_{ij}$ and $\bar{\boldsymbol{\Sigma}}_{ij}$ are the same as those defined in Section III-A.

For multiclass case, we define the following L2-MAGRQ criterion based on the two-class MAGRQ criterion of (38):
$$J(\omega) \triangleq \frac{\|\boldsymbol{\omega}^T\tilde{\mathbf{B}}\|_2^2}{\boldsymbol{\omega}^T\bar{\boldsymbol{\Sigma}}\boldsymbol{\omega}} + \frac{\sum_{i<j}P_{ij}|\boldsymbol{\omega}^T\Delta\boldsymbol{\Sigma}_{ij}\boldsymbol{\omega}|}{\boldsymbol{\omega}^T\bar{\boldsymbol{\Sigma}}\boldsymbol{\omega}} \tag{39}$$

where

$$\tilde{\mathbf{B}} = \left[\sqrt{P_{12}}\mathbf{B}_{12}, \ldots, \sqrt{P_{1c}}\mathbf{B}_{1c}, \sqrt{P_{23}}\mathbf{B}_{23}, \ldots, \sqrt{P_{(c-1)c}}\mathbf{B}_{(c-1)c}\right] \quad (40)$$

$$\mathbf{B}_{ij} = \left[\sqrt{\pi_{i1}\pi_{j1}}\Delta\mathbf{m}_{ij}^{11}, \sqrt{\pi_{i1}\pi_{j2}}\Delta\mathbf{m}_{ij}^{12}, \ldots, \sqrt{\pi_{iK_i}\pi_{jK_j}}\Delta\mathbf{m}_{ij}^{N_iN_j}\right]. \quad (41)$$

In this case, we can obtain the following L1-MAGRQ criterion under the mixture of Gaussian distribution:

$$J_1(\omega) \triangleq \frac{\|\boldsymbol{\omega}^T\tilde{\mathbf{B}}\|_1^2}{\boldsymbol{\omega}^T\bar{\boldsymbol{\Sigma}}\boldsymbol{\omega}} + \frac{\sum_{i<j} P_{ij}|\boldsymbol{\omega}^T\Delta\boldsymbol{\Sigma}_{ij}\boldsymbol{\omega}|}{\boldsymbol{\omega}^T\bar{\boldsymbol{\Sigma}}\boldsymbol{\omega}}. \quad (42)$$

Finally, based on the L1-MAGRQ criterion and the L2-MAGRQ criterion defined in (42) and (39), we can define the optimal discriminant vector set of L1-HDA/GM and L2-HDA/GM, respectively, which are similar to those defined in (17) and (30), respectively, and the solution methods are also similar to those shown in Algorithms 1 and 2, respectively.

V. Experiments

In this section, we will evaluate the discriminant performance of the proposed methods on four real-data databases, i.e., the Multi-PIE and the BU-3DFE facial expression databases [41], [44], the EEG database in "BCI Competition 2005" (data set IIIa) [46], and the UCI database [55]. Since L1-HDA/G is only a special case of L1-HDA/GM as for the case when the number of Gaussian components equals to 1, in the following experiments, we only adopt the L1-HDA/GM method to conduct the experiments. Moreover, we also use the L2-HDA/GM method to conduct the same experiments. For comparison purpose, several state-of-the-art discriminant analysis methods are adopted to conduct the same experiments, which include the FLDA method [5], the AIDA method [14], the HDA/Chernoff method [15], the HDA/HLFE method [18], the SDA method [21], and the multi-view deep network (MvDN) method [51]. In addition, we also conduct the experiment without any feature extraction and refer it as the baseline method. Noting that the experiments aim to evaluate the discriminative feature extraction performance of the various methods, in the following experiments, we only adopt simple classifiers, such as K-nearest neighbor and the linear classifier, to produce the classification results in order to compare the discriminative power of the extracted features.[1] In real applications, one may use more complex classifiers, such as support vector machine [39] and Adaboost [40], to further enhance the classification performance.

A. Experiment on Multi-PIE Facial Expression Database

In this experiment, we will use the famous Multi-PIE database [41] to evaluate the performance of the various methods. This is a multiview facial expression database consisting of 755 370 facial images of 337 subjects. Similar to [42], we choose 4200 facial images from 100 subjects as those previously used in [42] to conduct the experiment, in which each subject contains 42 facial images that cover seven facial views (0°, 15°, 30°, 45°, 60°, 75°, and 90°) and six facial expressions (disgust, neutral, scream, smile, squint, and surprise). Fig. 4 shows 42 facial images covering the six facial expressions and seven facial views of one subject in the Multi-PIE database.

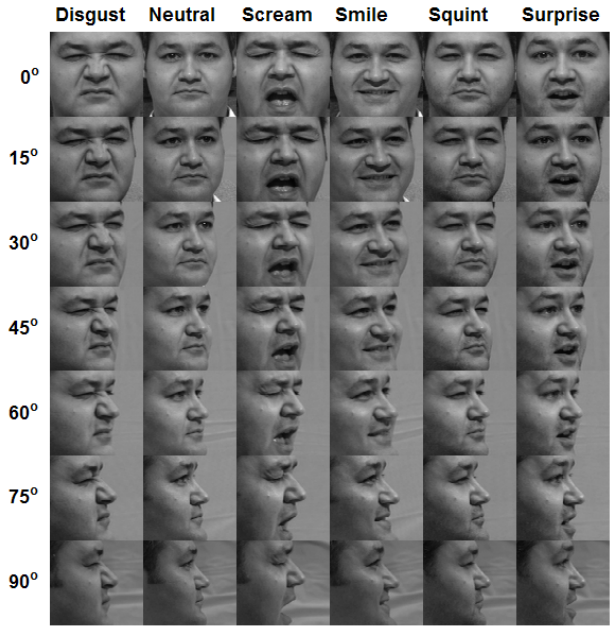


Fig. 4. Examples of the 42 facial images covering six facial expressions and seven facial views of one subject in the Multi-PIE database.

We explore two kinds of facial feature extraction schemes to evaluate the proposed methods. The first one is to adopt local binary patterns (LBPs) [43] to extract 5015 features from each facial image, and the second one is to extract deep learning (DL) features learned via deep neural networks as used in face recognition [54]. The details of extracting both kinds of features are summarized as follows.

1) To extract the LBP facial features, we use the multiscale face region division scheme proposed in [42] to obtain 85 facial regions and then extract a 59-D LBP feature vectors from each region, which results in 85 LBP feature vectors for each facial image. Finally, we concatenate all the 85 LBP feature vectors into a 5015-D feature vector.
2) To extract the DL features, we focus our attentions to the existing DL neural network model that had been successfully used in extracting facial features. For this purpose, we first utilize the real-world affective faces (RAFs) database [53] to fine-tune the deep-face VGG model (VGG-Face) [54]. Then, based on the fine-tuned VGG-Face model, we further extract the deep facial expression features by fitting the facial images of Multi-PIE database into the model. In this way, we finally obtain a set of facial features with dimensionality of 4096 taken from fc7 layer of the VGG-Face model.

In order to visualize the data distribution associated with each facial expression, we project the LBP feature points

[1]Otherwise, one may not be able to separate the contribution from feature extraction and that from classifier.

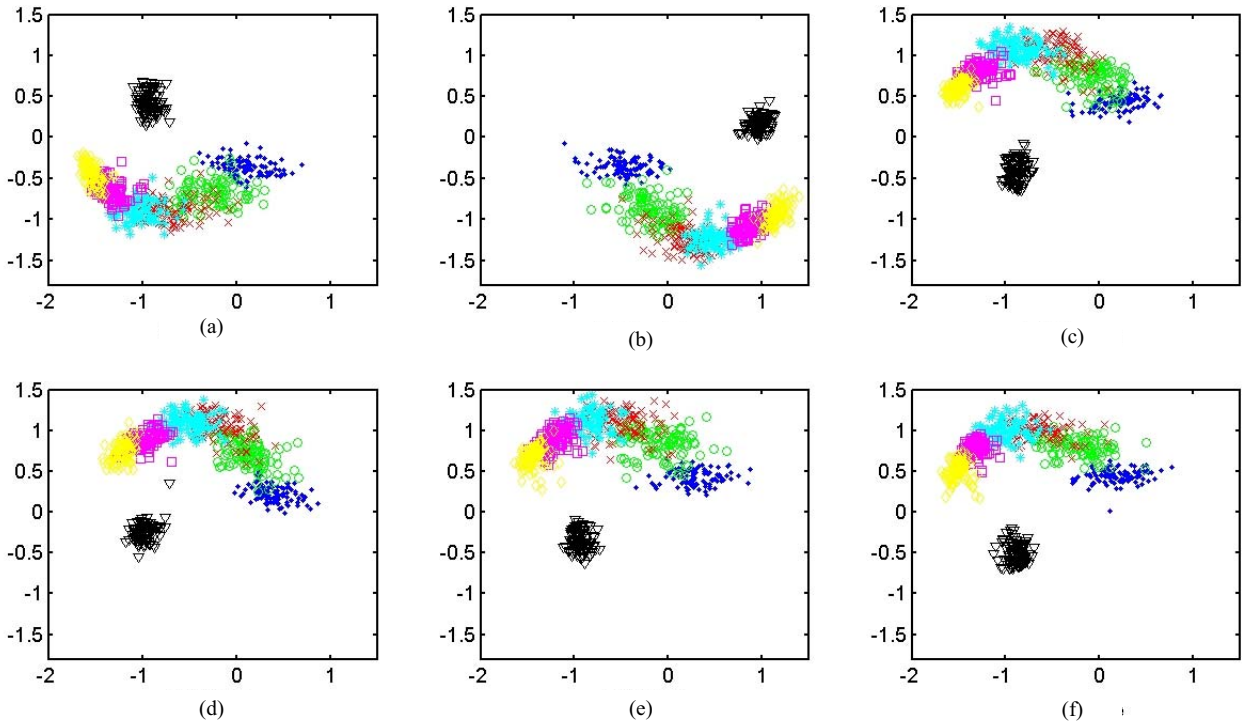


Fig. 5. Distributions of LBP facial features projected by the first two principal components of PCA with respect to all the six facial expression classes of Multi-PIE database, where the distribution of the data points of each facial expression demonstrates a multimodal form. (a) Disgust. (b) Neutral. (c) Scream. (d) Smile. (e) Squint. (f) Surprise.

associated with the same facial expression onto the first two principal components obtained by the principal component analysis (PCA) and then depict the distribution of the projection points. Fig. 5 shows the distributions of projection points with respect to all the six facial expression classes. From Fig. 5, we can see that, due to the multiview property of the facial images, the distribution of the data points associated with each facial expression demonstrates a multimodal form with seven clusters. In this case, it would be very possible that using single Gaussian function could not be enough to characterize the distribution of the multiview facial feature points. For this reason, for both the L1-HDA/GM and L2-HDA/GM methods, we use the mixture of Gaussians to describe the distribution of the facial feature points, in which the number of Gaussian components is set to be the number of facial views (=7) and each Gaussian component corresponds to the data samples of one facial view.

To evaluate the recognition performance of the various methods, we adopted the experimental protocol used in [42] to carry out this experiment. According to this protocol, the cross-validation strategy is used to design the experiment, in which we randomly partition the 100 subjects into two subsets, in which the first subset contains the facial images of 80 subjects and the second subset contains the facial images of 20 subjects. Then, we choose the first subset as the training data set and the second subset as the testing data set. Consequently, the training data set contains a total of 3360 facial images, whereas the testing one contains 840 facial images. Then, we train the discriminant vectors of the various discriminant algorithms on the training data set and evaluate the performance using the testing data set. Moreover, considering that the dimensionality of the feature space is relatively larger than the number of each class samples, a PCA operation on the training data set is used to reduce the dimensionality of the feature vectors, such that the covariance matrix of each class data samples is nonsingular. We conduct 10 trials of experiments in this database, and in each of the 10 trials, new training and testing data sets are chosen to evaluate the recognition performance of the various methods. Finally, we average the results of all the experimental trials to obtain the final recognition rate. Table II shows the average recognition rates with respect to each facial expression and the overall recognition rates of the various methods on the Multi-PIE facial expression database.

From Table II, we observe the following three major points.

1) The average recognition accuracies of using LBP features are higher than those of using DL features. This is most likely due to the fact that the VGG-face model is fine-tuned using other facial expression database, i.e., the RAF database, instead of the Multi-PIE database. Since the RAF database is irrelevant to the facial expression images to be tested, the extracted facial features may not well capture the discriminative information of the facial images of Multi-PIE database.
2) The L1-HDA/GM method and the L2-HDA/GM method achieve much competitive recognition rates for most of the linear discriminant analysis methods, where the highest overall recognition accuracy (=81.65%) is

TABLE II
AVERAGE CLASSIFICATION RATES (%) OF VARIOUS METHODS WITH RESPECT TO EACH FACIAL EXPRESSION ON THE MULTI-PIE DATABASE

		Disgust	Neutral	Scream	Smile	Squint	Surprise	Overall
# samples		700	700	700	700	700	700	4200
# subjects		100	100	100	100	100	100	100
dimensionality of LBP feature		5015	5015	5015	5015	5015	5015	5015
dimensionality of deep learning feature		4096	4096	4096	4096	4096	4096	4096
LBP Feature	Baseline	58.93	71.14	76.93	68.29	41.07	80.64	66.17
	FLDA	73.57	74.21	90.71	79.29	66.21	89.86	78.98
	AIDA [14]	63.43	74.43	85.64	76.14	65.86	91.36	76.14
	HDA/HLFE [18]	69.29	78.64	90.00	78.36	66.43	87.00	78.29
	HDA/Chernoff [15]	64.93	76.21	86.07	76.07	66.57	90.29	76.69
	SDA [21]	73.14	77.79	90.79	79.79	73.86	91.71	81.18
	MvDN [52]	66.00	77.50	91.00	75.00	80.50	95.50	80.92
	L2-HDA/GM	74.00	78.50	90.64	80.57	73.07	92.07	81.48
	L1-HDA/GM	73.86	78.64	89.93	80.86	73.86	92.79	**81.65**
Deep Learning Feature	Baseline	29.36	79.29	78.00	65.71	49.21	85.14	64.45
	FLDA	68.07	68.00	92.14	72.29	60.29	88.14	74.82
	AIDA [14]	32.29	58.21	71.00	48.14	35.79	69.64	52.51
	HDA/HLFE [18]	63.43	66.71	91.50	73.86	59.93	88.50	73.99
	HDA/Chernoff [15]	55.86	62.43	87.57	67.00	48.36	82.14	67.23
	SDA [21]	66.5	70.14	91.86	71.00	64.64	89.50	75.61
	MvDN [52]	62.00	68.00	91.50	79.00	76.50	91.50	**78.42**
	L2-HDA/GM	67.43	70.79	92.07	73.29	65.29	90.07	76.49
	L1-HDA/GM	68.79	70.00	92.14	73.93	64.50	89.57	76.49

TABLE III
COMPUTATIONAL EFFICIENCY OF THE VARIOUS METHODS IN TERMS OF THE CPU RUNNING TIME (second) IN THE TRAINING STAGE ON THE MULTI-PIE DATABASE

Feature	Baseline	FLDA	AIDA	HDA/HLFE	HDA/Chernoff	SDA	MvDN	L2-HDA/GM	L1-HDA/GM
LBP Feature	0.01	0.45	0.52	1.38	2.43	0.65	1209	11.16	9.62
DL Feature	0.02	3.71	14.29	61.95	132.90	5.19	2566	77.05	65.92

achieved by L1-HDA/GM. The better recognition accuracies may attribute to the use the mixture of Gaussians to approximate the multimodal distribution of the feature vectors.

3) The SDA method and the MvDN method also achieve the competitive recognition performance compared with the other methods. This is most likely due to the fact that SDA is actually a special case of our L2-HDA/GM method when the class covariance matrices of data samples are equal. In addition, it is interesting to see that from Fig. 5, the scatter of the data points is similar, and hence, the corresponding class covariance matrices would be similar. As a result, the difference between our L2-HDA/GM method and SDA in this experiment is trivial, and hence, they achieve similar better recognition results. As for the MvDN method, we note that it is a nonlinear feature extraction method and hence its better recognition performance may largely attribute to the nonlinear feature learning trick.

Moreover, to evaluate the computational efficiency of the various methods in the training stage, we also compare the CPU running time (second) among the various methods, where the calculation of CPU time is based on the computation platform of MATLAB2017B software with CPU i7-7700k and 16-GB memory. Table III summarizes the results of the various methods under the two kinds of facial features, i.e., LBP feature versus DL feature, where the parameter scale of the MvDN method is as high as 6×10^5. From Table III, we can see that the computational complexity of the proposed methods is very competitive to the state-of-the-art HDA methods, such as HDA/HLFE and HDA/Chernoff, which are much less than the MvDN method. In addition, from Table III, we can also see that the CPU running time of L1-HDA/G is a bit less than L2-HDA/G, which coincides with the computational analysis in Section III-D.

B. Experiment on BU-3DFE Facial Expression Database

In this experiment, we will evaluate the discriminative performance of the proposed methods on the BU-3DFE database, which was developed by Yin *et al.* [44] at Binghamton University. The BU-3DFE database contains 2400 3-D facial expression models of 100 subjects, which cover six basic facial expressions (anger, disgust, fear, happy, sad, and surprise) with four levels of intensities. Based on the 3-D facial expression model, a set of 12000 multiview 2-D facial images are generated [26], which covers five yaw facial views (0°, 30°, 45°, 60°, and 90°). Fig. 6 shows the examples of 30 facial images of one subject corresponding to six basic facial expressions and five facial views in the BU-3DFE database.

TABLE IV
AVERAGE CLASSIFICATION RATES (%) OF VARIOUS METHODS WITH RESPECT TO EACH FACIAL EXPRESSION ON THE BU-3DFE DATABASE

		Happy	Sad	Angry	Fear	Surprise	Disgust	Overall
# samples		2000	2000	2000	2000	2000	2000	12000
# subjects		100	100	100	100	100	100	100
dimensionality		10624	10624	10624	10624	10624	10624	10624
dimensionality of deep learning feature		4096	4096	4096	4096	4096	4096	4096
LBP Feature	Baseline	88.25	62.28	75.78	30.75	82.53	54.83	65.73
	FLDA	84.55	75.50	75.15	60.60	89.40	73.20	76.40
	AIDA [14]	85.95	75.35	75.63	56.00	88.85	72.40	75.70
	HDA/HLFE [18]	87.75	77.83	77.35	55.15	89.25	71.35	76.45
	HDA/Chernoff [15]	86.30	74.65	73.50	56.23	87.38	66.93	74.16
	SDA [21]	86.70	78.28	77.68	62.28	90.83	74.23	78.33
	MvDN [52]	84.00	73.75	77.00	66.25	88.13	71.88	77.10
	L2-HDA/GM	86.53	78.38	77.63	62.55	90.78	74.53	**78.40**
	L1-HDA/GM	85.88	78.23	77.88	62.98	90.98	74.25	78.36
Deep Learning Feature	Baseline	65.90	44.38	57.40	26.05	67.70	38.05	49.91
	FLDA	71.25	52.55	58.23	4.88	78.83	56.65	60.73
	AIDA [14]	67.83	38.60	50.43	34.25	68.45	35.50	49.18
	HDA/HLFE [18]	65.33	51.80	50.73	44.68	74.88	54.75	57.03
	HDA/Chernoff [15]	71.83	40.78	53.40	31.30	73.25	43.05	52.27
	SDA [21]	76.95	56.78	61.40	49.93	83.15	63.25	65.24
	MvDN [52]	60.87	67.00	58.50	77.75	55.125	78.00	**66.21**
	L2-HDA/GM	77.13	56.58	62.53	49.50	82.95	62.50	65.20
	L1-HDA/GM	76.63	56.65	61.88	50.05	82.33	62.48	65.00

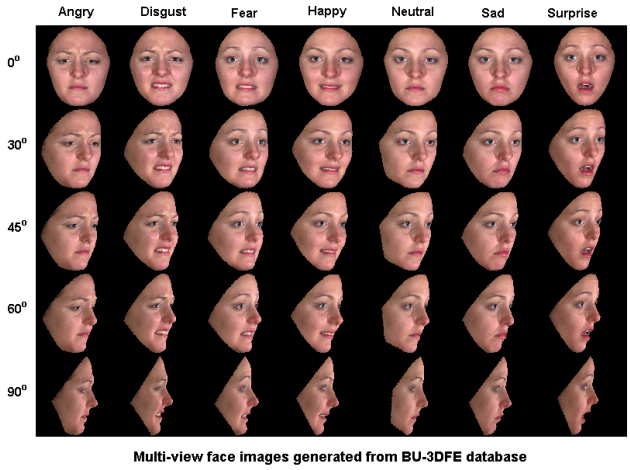


Fig. 6. Examples of the 2-D facial images of one subject in the BU-3DFE database with respect to the six facial expressions and five facial views.

In this database, we also explore two kinds of facial feature extraction schemes to evaluate the proposed methods. One is to extract the scale invariant feature transform (SIFT) [45] feature, and the other one is to extract the DL feature, in which the DL feature extraction procedure is similar to the one used in the Multi-PIE database. To extract the SIFT features, we utilized the 83 landmark points obtained by projecting 83 landmark 3-D points located on each 3-D facial expression model onto a 2-D image and extract a set of 83 SIFT feature vectors with 128 dimensionality from each facial image. Then, we concatenate all 83 SIFT feature vectors into a 10 624-D feature vector to describe the facial image. In addition, to visualize the distribution of the feature points associated with each facial expression, we reduce the dimensionality of the facial feature vectors of the same facial expression from the 10 624-D space to 2-D subspace using PCA and then depict the distribution of the data points. Fig. 7 shows the distributions of projection points with respect to all the six facial expressions. From Fig. 7, we observe that the distribution of the data points of each facial expression demonstrates a multimodal form with five clusters. Consequently, for both the L1-HDA/GM and L2-HDA/GM methods, the mixture of Gaussians with five components is used to describe the distribution of the facial feature points for the proposed methods, and each Gaussian component also corresponds to the data samples of one facial view.

In the experiments, we use the same cross-validation experimental setting as what we have done in Section V-A, i.e., there are 10 trials of experiments that are conducted, and in each trial of experiments, we partition the whole data set into a training set and a testing set, where the training set contains a total of 9600 facial images of 80 subjects, whereas the testing one contains 2400 facial images of 20 subjects. In addition, PCA is also used to reduce the dimensionality of the feature vectors such that the class covariance matrices become nonsingular. Finally, the experimental results of all the 10 trials are averaged as the overall recognition rate. Table IV shows the recognition results of the various methods on the BU-3DFE facial expression database.

From Table IV, we observe the similar experimental results as those obtained on the Multi-PIE database. That is, the average recognition accuracies of using SIFT features are higher than those of using DL features, and both the L1-HDA/GM and L2-HDA/GM methods achieve higher recognition accuracies than most of the other discriminant analysis methods.

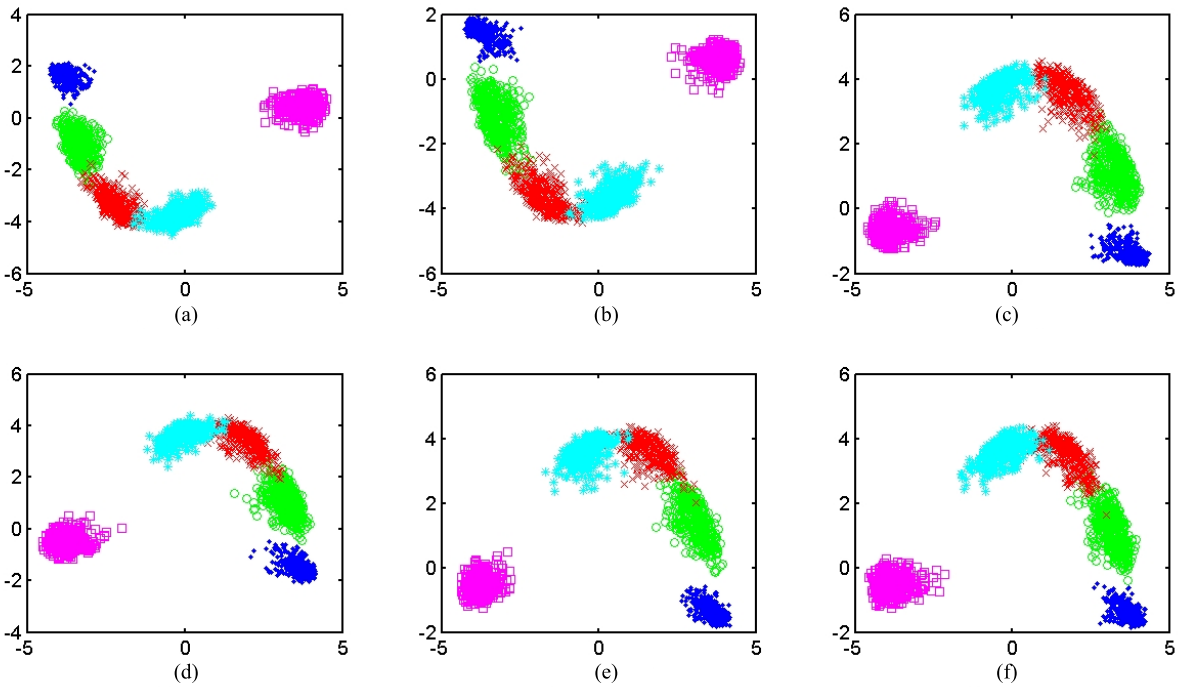


Fig. 7. Distributions of SIFT facial features projected by the first two principal components of PCA with respect to all the six facial expression classes of the BU-3DFE database. (a) Happy. (b) Sad. (c) Angry. (d) Fear. (e) Surprise. (f) Disgust.

Similar to the experiments on the Multi-PIE database, the major reason of the better recognition results of the proposed methods may attribute to the use the mixture of Gaussians to approximate the multimodal distribution of the feature vectors. Again, we see that the SDA method and the MvDN method also achieve better recognition results compared with the other state-of-the-art linear discriminant analysis methods because of the similar reason of the proposed methods.

C. Experiment on UCI Data Sets

In the above-mentioned two experiments, we note that the L1-HDA/GM method does not achieve significant improvement in contrast to the L2-HDA/GM method. This is probably due to the fact that the separabilities between the pairwise classes in both BU-3DFE and Multi-PIE databases are similar, such that the advantages of L1-HDA/GM cannot be well reflected. To further compare the recognition performances between L1-HDA/GM and L2-HDA/GM, in this section, we will conduct more experiments on more databases, in which the UCI database [55] previously used in [15] is adopted for evaluating the discriminant performance. There are totally nine data sets that are explored in the experiments and are listed as follows:

1) Wisconsin breast cancer;
2) BUPA liver disorder;
3) Pima Indians diabetes;
4) Wisconsin diagnostic breast cancer;
5) Cleveland heart-disease;
6) thyroid gland;
7) Landsat satellite;
8) multifeature digit (Zernike moments);
9) vowel context.

For each one of the nine data sets, we randomly divide it into two subsets, with an approximately equal size of samples. Then, we choose one subset for training the algorithms and use the other one for testing the discriminant performance. We swap the training subset and the testing subset, such that each subset is used as the training data set once. For each training data set, we utilize the nearest neighbor clustering to divide the data samples belonging to the same class into K subclasses. In this case, we can use the mixture of Gaussian model with K components to describe the distribution of the data samples of each class.

Similar to [15], before the experiments, a PCA is performed on the training set to reduce the dimensionality of the data samples, such that the covariance matrix of each data set is nonsingular. Throughout the experiments, we use the quadratic classifier for classifying the testing data. For each one of the nine data sets, the average error rate is used to evaluate the various discriminant methods. Table V summarizes the main properties of the nine UCI data sets and the average error rates of various feature extraction methods, where "#PC" in the fourth row shows how many principal components we use after the PCA processing. From Table V, we can observe that L1-HDA/GM outperforms L2-HDA/GM in all these nine data sets. Another observation from Table V is that, for both L1-HDA/GM and L2-HDA/GM, using the mixture of Gaussian model to describe the data samples of each class could achieve better than using the single Gaussian model.

D. Experiment on EEG Data Sets

In this experiment, we will evaluate the effectiveness of the proposed HDA in dealing with the feature extraction problem

TABLE V
UCI BENCHMARK DATA SETS AND THE AVERAGE ERROR RATES (%) OF SEVERAL METHODS

Data set		WBC	BUPA	PID	WDBC	CHD	TG	LS	MD	VC
# samples		682	345	768	569	297	215	6435	2000	990
# subjects		2	2	2	2	2	3	6	10	11
dimensionality		9	6	8	30	13	5	36	47	10
# PC		9	6	8	7	13	5	36	33	10
baseline	/	4.48	40.12	31.69	9.15	41.62	5.57	12.15	23.98	9.58
FLDA	/	4.56	43.14	29.35	6.14	24.22	4.54	11.71	19.90	1.16
AIDA	/	4.4	36.28	32.71	6.74	24.89	6.18	18.04	23.86	1.33
HDA/HLFE	/	4.56	43.14	29.35	6.14	22.00	4.54	13.71	19.90	1.16
HDA/Chernoff	/	4.42	39.28	31.13	5.88	23.05	3.63	13.22	23.03	1.38
MvDN	/	3.89	35.00	29.01	6.01	21.13	4.15	8.99	16.24	1.44
SDA [21]	$K=1$	3.77	39.71	28.31	7.19	24.33	3.63	13.51	20.30	1.52
	$K=2$	4.20	36.57	29.61	6.49	24.67	4.55	13.12	19.85	1.52
	$K=3$	4.49	43.14	31.82	6.84	24.67	4.55	13.18	19.50	1.52
	$K=4$	4.78	37.43	33.38	6.84	26.00	4.55	13.51	18.70	1.52
	$K=5$	4.78	37.43	32.08	6.84	27.00	4.55	14.24	18.70	1.52
L2-HDA/GM	$K=1$	5.21	36.57	41.10	6.32	27.00	6.59	13.65	19.40	1.11
	$K=2$	4.35	36.57	34.68	6.32	23.00	5.45	11.74	18.35	1.11
	$K=3$	4.35	36.57	31.75	6.32	23.00	5.45	11.43	17.90	1.11
	$K=4$	4.35	36.57	31.43	6.32	22.17	5.45	11.13	17.90	1.11
	$K=5$	4.35	36.57	31.17	6.32	22.17	5.45	11.13	17.90	1.11
L1-HDA/GM	$K=1$	4.86	43.71	30.13	6.93	22.67	4.09	14.40	19.75	1.06
	$K=2$	3.77	36.14	30.13	6.93	20.67	3.64	11.98	18.98	1.06
	$K=3$	3.77	34.14	29.74	5.79	20.67	3.64	10.83	17.80	1.06
	$K=4$	3.77	34.14	29.22	5.79	20.67	3.64	10.43	17.55	1.06
	$K=5$	**3.77**	**34.14**	**28.44**	**5.79**	**20.67**	**3.64**	**10.43**	**17.55**	**1.06**

as for the case when the class means are the same. To this end, we focus on an EEG classification problem whose target is to recognize the motor imagery tasks based on the EEG signal, i.e., recognizing what kind of motor imagery task that the EEG signal corresponds to.

The data set used in this experiment is from the "BCI competition 2005" (data set IIIa) [46]. This data set consists of recordings from three subjects (k3b, k6b, and l1b), which performed four different motor imagery tasks (left/right hand, one foot, or tongue) according to a cue. During the experiments, the EEG signal is recorded in 60 channels, using the left mastoid as reference and the right mastoid as ground. The EEG was sampled at 250 Hz and was filtered between 1 and 50 Hz with the notch filter on. Each trial lasted for 7 s, with the motor imagery performed during the last 4 s of each trial. For subjects k6b and l1b, a total of 60 trials per condition were recorded. For subject k3b, a total of 90 trials per condition were recorded. In addition, similar to the method in [10], we discard the four trials of subject k6b with missing data. The EEG data samples associated with the same class are preprocessed, such that their class means equal to zero vector [27]. Fig. 8 shows examples of the preprocessed EEG data samples of one trial with respect to different EEG classes and subjects, in which the four figures of each row show the data point distributions corresponding to four EEG classes of one subject, while the three figures of each column show the data point distribution corresponding to the same class of three subjects, respectively. From Fig. 8, we can see that the sample mean in each class is a zero point, and hence, only the class covariance differences can be utilized to extract the discriminative features.

Similar to [27], for each trial of the EEG data, we only use parts of the sample points, i.e., from Nos. 1001 to 1750, as the experiment data. Consequently, each trial contains 750 data points. In the experiment, we adopt a twofold cross-validation strategy to perform the experiment, i.e., we divide all the EEG trials into two groups and select one as the training data set and the other one as the testing data set, and then we swap the training and testing data set to repeat the experiment. Since the class means of the EEG data samples equal to the zero vector, the traditional Fisher's criterion-based approach, such as FLDA and SDA, cannot be applied in this experiment. Consequently, in this experiment, we only evaluate the discriminative feature extraction performance of the following four HDA methods, i.e., the AIDA method, the HDA/Chernoff method, the HDA/HLFE method, and the proposed L1-HDA/GM method.

As for the EEG classification, we extract the same log-transformation variance used in [11] to represent the final EEG feature of each trial. Then, we use the linear classifier to perform the EEG classification. Fig. 9 shows the average recognition rates of the four methods on the three subjects, from which we can clearly see that the proposed L1-HDA/GM method achieves much competitive experimental results compared with the best experimental results obtained by the other state-of-the-art methods.

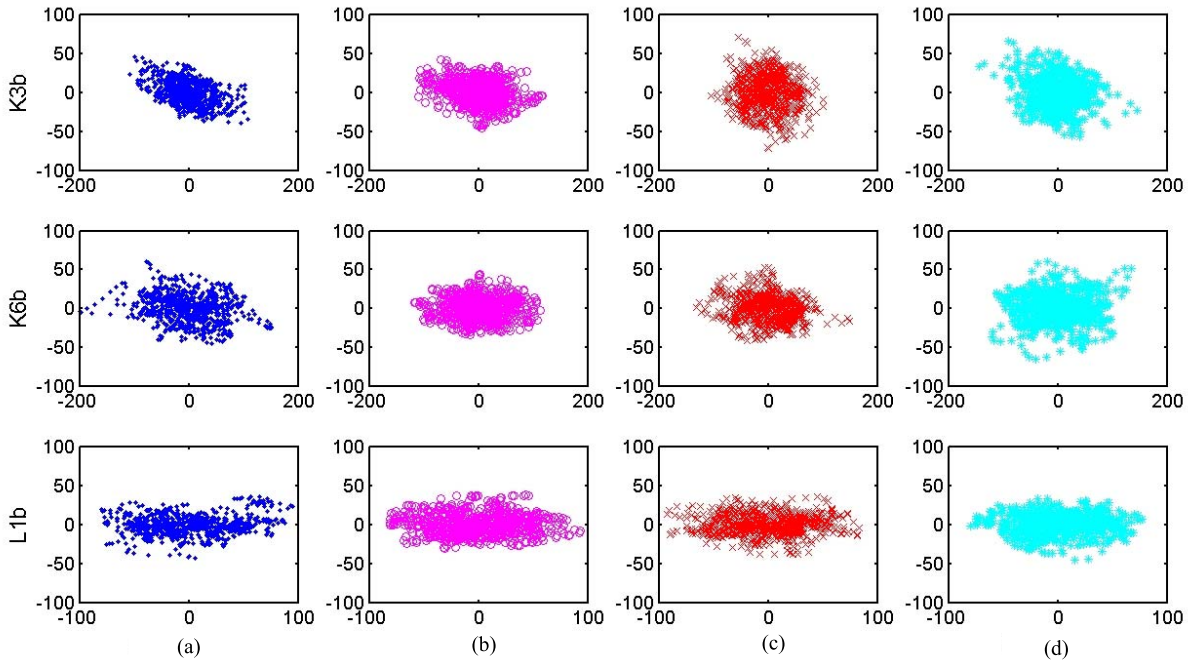


Fig. 8. Examples of the EEG data samples of one trial with respect to four EEG classes and three subjects. The four figures of each row show the data point distributions corresponding to the four EEG classes of one subject, whereas the three figures of each column show the data point distribution corresponding to the same class of the three subjects (K3b, K6b, and L1b). (a) Class 1. (b) Class 2. (c) Class 3. (d) Class 4.

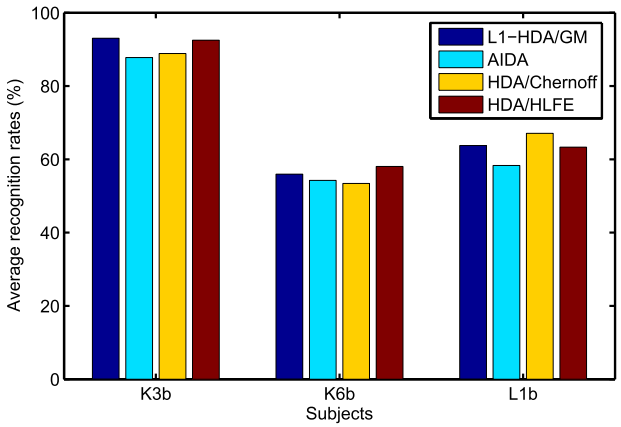


Fig. 9. Average recognition rates of various methods on the EEG data sets of three subjects.

VI. Conclusion and Discussion

In this paper, we have proposed a novel L2-MAGRQ criterion based on the Bayes error upper bound estimation. This criterion can be seen as a generalization of the traditional Fisher's criterion aiming to overcome the limitations of Fisher's criterion in the case of heteroscedastic distribution of the data samples in each class. The L2-MAGRQ criterion is further modified by replacing the ℓ_2-norm operation in the between-class scatter distance with ℓ_1-norm, resulting in the L1-MAGRQ criterion. Two kinds of HDA methods, L1-HDA/G (L2-HDA/G) and L1-HDA/GM (L2-HDA/GM), based on the L1-MAGRQ (L2-MAGRQ) criterion are, respectively, proposed for discriminative feature extraction, in which L1-HDA/G (L2-HDA/G) corresponds to the case where the distribution of each class data set is Gaussian, whereas L1-HDA/GM (L2-HDA/GM) corresponds to the mixture of Gaussian distributions [it is notable that L1-HDA/G (L2-HDA/G) is actually a special case of L1-HDA/GM (L2-HDA/GM) when the mixture of Gaussian distributions reduces to the Gaussian distribution]. Moreover, we also propose an efficient algorithm to solve the optimal discriminant vectors of L1-HDA/GM (L2-HDA/GM). Although the algorithm presented in this paper for solving the optimal discriminant vectors of L1-HDA/GM (L2-HDA/GM) is a greedy algorithm, it is easier to develop a nongreedy algorithm for L1-HDA/GM (L2-HDA/GM) by referring the method proposed in [47] and [48]. The experiments on four real databases had been conducted to evaluate the discriminative performance of the proposed methods. The experimental results demonstrate that the proposed L1-HDA/GM method achieves better recognition performance than most of the state-of-the-art methods, which may attribute to the use of mixture of Gaussian distributions and the Bayes error estimation. In addition, in the multiview facial expression recognition experiments, we see that the SDA method also achieves the similar better experimental results as ours. This is most likely due to the fact that SDA is a special case of L2-HDA/GM as for the case of similar class covariance matrices.

In addition, the proposed L1-HDA/GM (L2-HDA/GM) methods can also be used to improve the current graph-based subspace learning performance. For example, Peng *et al.* [56] constructed an ℓ_2-norm-based sparse similarity graph for robust subspace learning under the sparse representation framework, in which the sparse relationships among the data points are preserved in the low-dimensional subspace. It is notable that the feature extraction part of the method proposed by Peng *et al.* [56] is actually an unsupervised subspace

learning approach, which could not fully utilize the class label information of the data points to improve the discriminative ability of the extracted features. By adopting the proposed HDA methods, however, we may be able to learn a more discriminative subspace for the feature extraction purpose.

In addition, in our experiments, we can see that the MvDN method proposed in [51] achieved very competitive results with our L1-HDA/GM (L2-HDA/GM) methods. This is very likely due to the fact that MvDN is actually a nonlinear feature extraction method implemented by using the deep neural networks approach (hence, the computational complexity of MvDN would be larger than the other methods, see Table III for more details). In contrast to MvDN, both the proposed L1-HDA/GM and L2-HDA/GM methods are linear feature extraction methods. Nevertheless, it is notable that we are able to adopt the similar nonlinear learning trick of using deep neural network to realize the proposed L1-HDA/GM (L2-HDA/GM) algorithm to further improve the discriminative feature extraction performance, which would be our future work.

APPENDIX A
PROOF OF THEOREM 1

Proof: Suppose that $\alpha^{(2)}$ is the principal eigenvector of $\mathbf{T}(\mathbf{U}_2, \mathbf{V}_2)$, that is

$$\alpha^{(2)} = \arg\max_{\|\alpha\|=1} \alpha^T \mathbf{T}(\mathbf{U}_2, \mathbf{V}_2)\alpha. \tag{43}$$

Then, by the definition of $\alpha^{(2)}$, we have

$$\alpha^{(2)^T} \mathbf{T}(\mathbf{U}_2, \mathbf{V}_2)\alpha^{(2)} \geq \alpha^{(1)^T} \mathbf{T}(\mathbf{U}_2, \mathbf{V}_2)\alpha^{(1)}. \tag{44}$$

On the other hand, we have

$$\alpha^{(1)^T} \mathbf{T}(\mathbf{U}_2, \mathbf{V}_2)\alpha^{(1)} = \left(\sum_{i<j} (\mathbf{U}_2)_{ij} \alpha^{(1)^T} \Delta\hat{\mathbf{m}}_{ij}\right)^2 + \sum_{i<j} (\mathbf{V}_2)_{ij} \alpha^{(1)^T} \Delta\hat{\mathbf{\Sigma}}_{ij} \alpha^{(1)}. \tag{45}$$

From (25) and (45), we have

$$\begin{aligned}
&\alpha^{(1)^T} \mathbf{T}(\mathbf{U}_2, \mathbf{V}_2)\alpha^{(1)} \\
&= \left(\sum_{i<j} |\alpha^{(1)^T} \Delta\hat{\mathbf{m}}_{ij}|\right)^2 + \sum_{i<j} |\alpha^{(1)^T} \Delta\hat{\mathbf{\Sigma}}_{ij} \alpha^{(1)}| \\
&\geq \left(\sum_{i<j} (\mathbf{U}_1)_{ij} \alpha^{(1)^T} \Delta\hat{\mathbf{m}}_{ij}\right)^2 + \sum_{i<j} (\mathbf{V}_1)_{ij} \alpha^{(1)^T} \Delta\hat{\mathbf{\Sigma}}_{ij} \alpha^{(1)} \\
&= \alpha^{(1)^T} \mathbf{T}(\mathbf{U}_1, \mathbf{V}_1)\alpha^{(1)}.
\end{aligned} \tag{46}$$

Combine (44) and (46), we have

$$\alpha^{(2)^T} \mathbf{T}(\mathbf{U}_2, \mathbf{V}_2)\alpha^{(2)} \geq \alpha^{(1)^T} \mathbf{T}(\mathbf{U}_1, \mathbf{V}_1)\alpha^{(1)}. \tag{47}$$

APPENDIX B
PROOF OF THEOREM 2

For simplicity of the derivation, we denote the class distribution function of $\mathbf{x} \in \mathbf{X}_i$ by $p_x(\mathbf{x})$ and the distribution function of $y = \omega^T\mathbf{x}$ by $p_y(y)$, i.e., $p_x(\mathbf{x}) = p_i(\mathbf{x}|\mathbf{x} \in \mathbf{X}_i)$ and $p_y(y = \omega^T\mathbf{x}) = \tilde{p}_i(\omega^T\mathbf{x}|\mathbf{x} \in \mathbf{X}_i)$. Then, the characteristic function of $\mathbf{x}$ is

$$\begin{aligned}
\phi_x(\mathbf{t}) &= E(e^{j\mathbf{t}^T\mathbf{x}}) = \sum_{r=1}^{K_i} \pi_{ir} \int_{\mathbf{x}} e^{j\mathbf{t}^T\mathbf{x}} \mathcal{N}(\mathbf{m}_{ir}, \mathbf{\Sigma}_{ir}) d\mathbf{x} \\
&= \sum_{r=1}^{K_i} \pi_{ir} \exp\left(j\mathbf{t}^T\mathbf{m}_{ir} - \frac{1}{2}\mathbf{t}^T\mathbf{\Sigma}_{ir}\mathbf{t}\right)
\end{aligned} \tag{48}$$

where $j^2 = -1$.

Then, the characteristic function of y is

$$\begin{aligned}
\phi_y(\xi) &= E\{e^{j\xi y}\} = E\{e^{j\xi\omega^T\mathbf{x}}\} = \phi_x(\xi\omega) \\
&= \sum_{r=1}^{K_i} \pi_{ir} \exp\left(j\xi\omega^T\mathbf{m}_{ir} - \frac{1}{2}\xi^2\omega^T\mathbf{\Sigma}_{ir}\omega\right).
\end{aligned} \tag{49}$$

So the density distribution function of y is

$$\begin{aligned}
&p_y(\eta) \\
&= \frac{1}{2\pi}\int_\xi \phi_y(\xi) e^{-j\xi\eta} d\xi \\
&= \sum_{r=1}^{K_i} \pi_{ir} \frac{1}{2\pi}\int_\xi e^{-j\xi\eta} \exp\left(j\xi\omega^T\mathbf{m}_{ir} - \frac{1}{2}\xi^2\omega^T\mathbf{\Sigma}_{ir}\omega\right) d\xi \\
&= \sum_{r=1}^{K_i} \pi_{ir} \frac{1}{2\pi}\int_\xi \exp\left\{-\frac{1}{2}[\xi^2\omega^T\mathbf{\Sigma}_{ir}\omega + 2j\xi(\eta - \omega^T\mathbf{m}_{ir})]\right\} d\xi \\
&= \sum_{r=1}^{K_i} \pi_{ir} \frac{1}{2\pi}\int_\xi \exp\left\{-\frac{1}{2}\left[(\omega^T\mathbf{\Sigma}_{ir}\omega)\left(\xi + \frac{j(\eta - \omega^T\mathbf{m}_{ir})}{\omega^T\mathbf{\Sigma}_{ir}\omega}\right)^2 + \frac{(\eta - \omega^T\mathbf{m}_{ir})^2}{\omega^T\mathbf{\Sigma}_{ir}\omega}\right]\right\} d\xi \\
&= \sum_{r=1}^{K_i} \pi_{ir} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{(\eta - \omega^T\mathbf{m}_{ir})^2}{\omega^T\mathbf{\Sigma}_{ir}\omega}\right\} \\
&\quad \times \frac{1}{\sqrt{2\pi}}\int_\xi \exp\left\{-\frac{1}{2}\left[(\omega^T\mathbf{\Sigma}_{ir}\omega)\left(\xi + \frac{j(\eta - \omega^T\mathbf{m}_{ir})}{\omega^T\mathbf{\Sigma}_{ir}\omega}\right)^2\right]\right\} d\xi \\
&= \sum_{r=1}^{K_i} \pi_{ir} \frac{1}{\sqrt{2\pi(\omega^T\mathbf{\Sigma}_{ir}\omega)}} \exp\left\{-\frac{1}{2}\frac{(\eta - \omega^T\mathbf{m}_{ir})^2}{\omega^T\mathbf{\Sigma}_{ir}\omega}\right\} \\
&= \sum_{r=1}^{K_i} \pi_{ir} \mathcal{N}(\eta|\omega^T\mathbf{m}_{ir}, \omega^T\mathbf{\Sigma}_{ir}\omega).
\end{aligned} \tag{50}$$

This completes the proof of Theorem 2.

REFERENCES

[1] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York, NY, USA: Academic, 1990.

[2] H. Zhu, F. Meng, J. Cai, and S. Lu, "Beyond pixels: A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation," *J. Vis. Commun. Image Represent.*, vol. 34, pp. 12–27, Jan. 2016.

[3] H. Zhu *et al.*, "YoTube: Searching action proposal via recurrent and static regression networks," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2609–2622, Jun. 2018.

[4] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York, NY, USA: Wiley, 1973.

[5] 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.

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

[7] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Mar. 1992, pp. 13–16.

[8] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Sparse representation based Fisher discrimination dictionary learning for image classification," *Int. J. Comput. Vis.*, vol. 109, no. 3, pp. 209–232, Sep. 2014.

[9] X. Peng, H. Tang, L. Zhang, Z. Yi, and S. Xiao, "A unified framework for representation-based subspace clustering of out-of-sample and large-scale data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 12, pp. 2499–2512, Dec. 2016.

[10] M. Grosse-Wentrup and M. Buss, "Multiclass common spatial patterns and information theoretic feature extraction," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 8, pp. 1991–2000, Aug. 2008.

[11] H. Ramoser, J. Müller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," *IEEE Trans. Rehabil. Eng.*, vol. 8, no. 4, pp. 441–446, Dec. 2000.

[12] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Commun.*, vol. 26, no. 4, pp. 283–297, 1998.

[13] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Jun. 2000, pp. 129–132.

[14] K. Das and Z. Nenadic, "Approximate information discriminant analysis: A computationally simple heteroscedastic feature extraction technique," *Pattern Recognit.*, vol. 41, no. 5, pp. 1548–1557, 2008.

[15] R. P. W. Duin and M. Loog, "Linear dimensionality reduction via a heteroscedastic extension of LDA: The Chernoff criterion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 6, pp. 732–739, Jun. 2004.

[16] Y.-K. Noh, J. Hamm, F. C. Park, B.-T. Zhang, and D. D. Lee, "Fluid dynamic models for bhattacharyya-based discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 92–105, Jan. 2017.

[17] Z. Nenadic, "Information discriminant analysis: Feature extraction with an information-theoretic objective," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 8, pp. 1394–1407, Aug. 2007.

[18] P.-F. Hsieh, D.-S. Wang, and C.-W. Hsu, "A linear feature extraction for multiclass classification problems based on class mean and covariance discriminant information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 223–235, Feb. 2006.

[19] P.-F. Hsieh and D. Landgrebe, "Linear feature extraction for multiclass problems," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, vol. 4, Jul. 1998, pp. 2050–2052.

[20] W. Zheng, L. Zhao, and C. Zou, "An efficient algorithm to solve the small sample size problem for LDA," *Pattern Recognit.*, vol. 37, no. 5, pp. 1077–1079, 2004.

[21] M. Zhu and A. M. Martínez, "Subclass discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 8, pp. 1274–1286, Aug. 2006.

[22] H. Wan, H. Wang, G. Guo, and X. Wei, "Separability-oriented subclass discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 409–422, Feb. 2018.

[23] S. Mika, G. Ratsch, J. Weston, B. Schölkopf, and K. R. Mullers, "Fisher discriminant analysis with kernels," in *Proc. IEEE Int. Workshop Neural Netw. Signal Process. IX*, Aug. 1999, pp. 41–48.

[24] M.-H. Yang, "Kernel eigenfaces vs. kernel fisherfaces: Face recognition using kernel methods," in *Proc. 5th IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2002, pp. 215–220.

[25] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[26] W. Zheng, H. Tang, Z. Lin, and T. S. Huang, "A novel approach to expression recognition from non-frontal face images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Sep./Oct. 2009, pp. 1901–1908.

[27] W. Zheng and Z. Lin, "Optimizing multi-class spatio-spectral filters via Bayes error estimation for EEG classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 2268–2276.

[28] W. Zheng, H. Tang, Z. Lin, and T. S. Huang, "Emotion recognition from arbitrary view facial images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2010, pp. 490–503.

[29] W. Malina, "On an extended fisher criterion for feature selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-3, no. 5, pp. 611–614, Sep. 1981.

[30] H. Wang, Q. Tang, and W. Zheng, "L1-norm-based common spatial patterns," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 3, pp. 653–662, Mar. 2012.

[31] F. Zhong and J. Zhang, "Linear discriminant analysis based on ℓ_1-norm maximization," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3018–3027, Aug. 2013.

[32] W. Zheng, Z. Lin, and H. Wang, "L1-norm kernel discriminant analysis via Bayes error bound optimization for robust feature extraction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 4, pp. 793–805, Apr. 2014.

[33] H. Wang, X. Lu, Z. Hu, and W. Zheng, "Fisher discriminant analysis with ℓ_1-norm," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 828–842, Jun. 2014.

[34] Q. Ye, J. Yang, F. Liu, C. Zhao, N. Ye, and T. Yin, "ℓ_1-norm distance linear discriminant analysis based on an effective iterative algorithm," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 1, pp. 114–129, Jan. 2016.

[35] W. Zheng, Z. Lin, and X. Tang, "A rank-one update algorithm for fast solving kernel Foley–Sammon optimal discriminant vectors," *IEEE Trans. Neural Netw.*, vol. 21, no. 3, pp. 393–403, Mar. 2010.

[36] K. Fukunaga and W. L. G. Koontz, "Application of the Karhunen–Loève expansion to feature selection and ordering," *IEEE Trans. Comput.*, vol. C-19, no. 4, pp. 311–318, Apr. 1970.

[37] Z. Jin, J.-Y. Yang, Z.-S. Hu, and Z. Lou, "Face recognition based on the uncorrelated discriminant transformation," *Pattern Recognit.*, vol. 34, no. 7, pp. 1405–1416, 2001.

[38] W. Zheng, "Heteroscedastic feature extraction for texture classification," *IEEE Signal Process. Lett.*, vol. 16, no. 9, pp. 766–769, Sep. 2009.

[39] C. Cortes and V. Vapnik, "Support-vector network," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[40] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. CVPR*, Dec. 2001, pp. 511–518.

[41] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, 2010.

[42] W. Zheng, "Multi-view facial expression recognition based on group sparse reduced-rank regression," *IEEE Trans. Affective Comput.*, vol. 5, no. 1, pp. 71–85, Jan. 2014.

[43] T. Ojala, M. Pietikainen, and I. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[44] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3D facial expression database for facial behavior research," in *Proc. 7th Int. Conf. Autom. Face Gesture Recognit.*, Apr. 2006, pp. 211–216.

[45] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[46] G. Blankertz *et al.*, "The BCI competition III: Validating alternative approaches to actual BCI problems," *IEEE Trans. Rehabil. Eng.*, vol. 14, no. 2, pp. 153–159, Jun. 2006.

[47] F. Nie, H. Huang, C. Ding, D. Luo, and H. Wang, "Robust principal component analysis with non-greedy ℓ_1-norm maximization," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, 2011, pp. 1433–1438.

[48] Y. Liu, Q. Gao, S. Miao, X. Gao, F. Nie, and Y. Li, "A non-greedy algorithm for L1-norm LDA," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 684–695, Feb. 2017.

[49] T. Diethe, D. R. Hardoon, and J. Shawe-Taylor, "Constructing nonlinear discriminants from multiple data views," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2010, pp. 328–343.

[50] H. Wang, C. Ding, and H. Huang, "Multi-label linear discriminant analysis," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2010, pp. 126–139.

[51] M. Kan, S. Shan, and X. Chen, "Multi-view deep network for cross-view classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4847–4855.

[52] S. Sun, "A survey of multi-view machine learning," *Neural Comput. Appl.*, vol. 23, nos. 7–8, pp. 2031–2038, 2013.

[53] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2584–2593.

[54] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis.*, 2015, pp. 1–12.

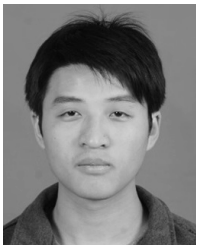
[55] (2004). *UCI Repository of Machine Learning Database*. [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository

[56] X. Peng, Z. Yu, Z. Yi, and H. Tang, "Constructing the L2-graph for robust subspace learning and subspace clustering," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 1053–1066, Apr. 2016.

Wenming Zheng (M'08) received the B.S. degree in computer science from Fuzhou University, Fuzhou, China, in 1997, the M.S. degree in computer science from Huaqiao University, Quanzhou, China, in 2001, and the Ph.D. degree in signal processing from Southeast University, Nanjing, China, in 2004.

Since 2004, he has been with the Research Center for Learning Science, Southeast University, where he is currently a Professor with the Key Laboratory of Child Development and Learning Science, Ministry of Education. His current research interests include affective computing, pattern recognition, machine learning, and computer vision.

Dr. Zheng is an Associate Editor of the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING and *Neurocomputing* and an Editorial Board member of *The Visual Computer*.



Cheng Lu received the B.S. and M.S. degrees from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2013 and 2017, respectively. He is currently pursuing the Ph.D. degree with the School of Information Science and Engineering, Southeast University, Nanjing, China, under the supervision of Prof. W. Zheng.

His current research interests include speech emotion recognition, machine learning, and pattern recognition.

Zhouchen Lin (M'00–SM'08–F'18) is currently a Professor with the Key Laboratory of Machine Perception, School of Electronics Engineering and Computer Science, Peking University, Beijing, China. His current research interests include computer vision, image processing, machine learning, pattern recognition, and numerical optimization.

Dr. Lin is an International Association of Pattern Recognition Fellow. He is an Area Chair of Conference on Computer Vision and Pattern Recognition 2014/2016/2019, International Conference on Computer Vision 2015, Conference on Neural Information Processing Systems 2015/2018, and Conference of Association for the Advancement of Artificial Intelligence (AAAI) 2019 and a Senior Program Committee Member of AAAI 2016/2017/2018 and International Joint Conference on Artificial Intelligence 2016/2018. He is an Associate Editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and the *International Journal of Computer Vision*.

Tong Zhang received the B.S. degree from the Department of Information Science and Technology, Southeast University, Nanjing, China, in 2011, and the M.S. degree from the Research Center for Learning Science, Southeast University, in 2014, where he is currently pursuing the Ph.D. degree in information and communication engineering.

His current research interests include pattern recognition, machine learning, and computer vision.

Zhen Cui received the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2014.

He was a Research Fellow with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, from 2014 to 2015. He was also a Research Assistant with Nanyang Technological University, Singapore, in 2012. He is currently a Professor with the Nanjing University of Science and Technology, Nanjing, China. His current research interests include computer vision, pattern recognition, and machine learning, especially focusing on deep learning, manifold learning, sparse coding, face detection/alignment/recognition, object tracking, image super-resolution, and emotion analysis.

Wankou Yang received the B.S., M.S., and Ph.D. degrees from the School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing, China, in 2002, 2004, and 2009, respectively.

He was with the School of Automation, Southeast University, Nanjing, as a Post-Doctoral Fellow from 2009 to 2011 and an Assistant Professor from 2011 to 2016, where he is currently an Associate Professor. His current research interests include pattern recognition and computer vision.