Locality-Constrained Discriminative Matrix Regression for Robust Face Identification

Chao Zhang, *Graduate Student Member, IEEE*, Huaxiong Li, *Member, IEEE*, Yuhua Qian, *Member, IEEE*, Chunlin Chen, *Member, IEEE*, and Xianzhong Zhou, *Member, IEEE*

***Abstract*— Regression-based methods have been widely applied in face identification, which attempts to approximately represent a query sample as a linear combination of all training samples. Recently, a matrix regression model based on nuclear norm has been proposed and shown strong robustness to structural noises. However, it may ignore two important issues: the label information and local relationship of data. In this article, a novel robust representation method called locality-constrained discriminative matrix regression (LDMR) is proposed, which takes label information and locality structure into account. Instead of focusing on the representation coefficients, LDMR directly imposes constraints on representation components by fully considering the label information, which has a closer connection to identification process. The locality structure characterized by subspace distances is used to learn class weights, and the correct class is forced to make more contribution to representation. Furthermore, the class weights are also incorporated into a competitive constraint on the representation components, which reduces the pairwise correlations between different classes and enhances the competitive relationships among all classes. An iterative optimization algorithm is presented to solve LDMR. Experiments on several benchmark data sets demonstrate that LDMR outperforms some state-of-the-art regression-based methods.**

***Index Terms*— Class competitions, locality constraints, matrix regression, robust face identification (FI).**

I. Introduction

FACE identification (FI) is one of the most intensively investigated topics for researchers in the fields of pattern recognition and computer vision. In the past 20 years, we have witnessed emerging FI methods, including subspace analysis methods [1]–[3], regression-based methods [4]–[6], and convolutional neural network (CNN)-based methods [7]–[13]. The CNN-based methods have recently captured much attention [14]–[16], and many successful deep CNNs are proposed for face recognition, such as FaceNet [7], CosFace [8], and ArcFace [9]–[11]. However, deep learning methods generally require a large amount of training data and high computational power. Besides, the unclear theoretical understanding of deep learning models makes it difficult to determine the optimal architecture and optimization algorithm. The subspace analysis methods can learn discriminative information in data, but they are incapable to well deal with the complex variations in images such as facial occlusions [3], which are commonplace in epidemic.

Recently, regression-based approaches attracted much interests due to the mathematically interpretability and great success in FI [17]–[21], image processing [22], visual tracking [23], and so on. Regression-based methods on FI assume that a query sample can be approximately represented by a linear combination of all training samples, and it is classified into the class that yields the minimal reconstruction residual. Wright *et al.* [18] presented a sparse representation classification (SRC) method. SRC aims at using a sparse linear combination of all training samples to represent the query sample. The ℓ^1-norm regularizer is applied in SRC to achieve the sparsity of representation coefficients and sparsity models attracted broad interests [24], [25]. Zhang *et al.* [20] believed that the underlying reason, which truly improves the recognition performance, is the collaboration between classes rather than sparsity. They proposed a collaborative representation classifier (CRC) using an ℓ^2-norm regularizer, which achieved much lower computational costs than SRC. To further improve the robustness to noises and occlusions, Naseem *et al.* [26] and Cai *et al.* [27] proposed robust linear regression classification (RLRC) and robust collaborative representation classification (RCRC), which are the extensions of LRC and CRC, respectively. Xu *et al.* [28] presented a two-phase test sample representation (TPTSR) method for face recognition. TPTSR performs twice linear representation for FI, which selects some neighbors of test sample in the first phase and performs identification using these neighbors. To alleviate the influence of outlier features or pixels, Yang *et al.* [29] proposed a novel regularized robust coding (RRC) method by introducing adaptive and iterative pixel weights learning mechanism. Cai *et al.* [27] presented a probabilistic CRC (ProCRC) based on the maximum likelihood of each class.

Manuscript received May 20, 2020; revised August 22, 2020; accepted November 25, 2020. Date of publication December 17, 2020; date of current version March 1, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61672332, Grant 62073160, Grant 71671086, Grant 61876079, and Grant 71732003; and in part by the National Key Research and Development Program of China under Grant 2018YFB1402600 and 2016YFD0702100. *(Corresponding author: Huaxiong Li.)*

Chao Zhang, Huaxiong Li, Chunlin Chen, and Xianzhong Zhou are with the Department of Control and Systems Engineering, Nanjing University, Nanjing 210093, China (e-mail: chzhang@smail.nju.edu.cn; huaxiongli@nju.edu.cn; clchen@nju.edu.cn; zhouxz@nju.edu.cn).

Yuhua Qian is with the Institute of Big Data Science and Industry, Shanxi University, Taiyuan 030006, China, and also with the Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006, China (e-mail: jinchengqyh@126.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2020.3041636.

Digital Object Identifier 10.1109/TNNLS.2020.3041636

It should be noted that all these abovementioned methods belong to the 1-D regression model, which measures the representation error in vector. They have a common underlying assumption that the pixelwise errors are independent [30], [31]. However, the true distribution of representation error is much sophisticated in real world. Yang *et al.* [30] preserved the 2-D structure of error image and proposed a nuclear norm-based matrix regression (NMR) method. Based on the observation that the error caused by contiguous noises is generally low-rank or approximately low-rank, NMR uses the nuclear norm as loss metric, which has shown robustness and efficiency for FI, especially to structural noises. Based on NMR, Luo *et al.* [32] took the sparse and low-rank structure of representation error into account simultaneously to handle the mixed noises. In [33] and [34], different nonconvex relaxations of the rank minimization problem are used to describe the low-rank structure of error image. However, NMR may have two disadvantages: the important label information is not utilized, and the locality structure of data is ignored, which are two important issues in FI.

Many researchers tried to search approaches to exploit the label information and local structure to improve the FI performance. In [35]–[38], group sparse representation methods are proposed to utilize the label information of training samples and enforce the sparsity of representation coefficients at the group level. Lai and Jiang [38] focused on the interclass sparsity of coefficients and proposed a classwise sparse classifier (CSC). Wang *et al.* [39] proposed a hierarchal images classification method, called locality-constrained linear coding (LLC) by using the distance information between data. Fan *et al.* [40] determined the weights of training samples by the Euclidean distance from query sample to all training samples and presented a weighted SRC (WSRC) method. Similarly, many other research works use the prior knowledge of distance information to characterize the locality structure of data, such as weighted CRC (WCRC) [41], locality-constrained least-squares regression [42], and so on [43]–[45]. All these locality involved approaches use the prior knowledge, which is derived from the nearest neighbor (NN) method, to learn the locality constraints. Peng *et al.* [46] proposed a locality-constrained collaborative representation for image classification. It is based on the observation that samples and their neighbors have similar codes and also use the distance information. Wen *et al.* [47] imposed an adaptive weighted constraint upon representation error to select important features in representation. Zheng *et al.* [37] proposed an Iterative Re-constrained Group Sparse Classifier (IRGSC), which combines the locality structure of data and features selection mechanism. These research works demonstrate that the label information and locality constraints generally improve the FI performance.

In this work, we take the label information, locality structure of samples, and class competition into account simultaneously and propose a locality-constrained discriminative matrix regression (LDMR) method to deal with structural noises in face images, such as masks, sunglasses, and scarves occlusion. Instead of regularizing on the representation coefficients [36], [39], [40], [42], LDMR directly constrains the classwise representation components, which has a closer connection to the identification process. The locality structure is characterized by the distances between test sample and subspaces spanned by multiclass training samples. Different classes are assigned with different weights, and the training samples in the same class are treated equally. This mechanism enforces the representation model to pay more attention to interclass difference and less sensitive to outliers. To further improve the discrimination of representation, a weighted pairwise class competition term is incorporated into LDMR, which reduces the correlations and enhances the competitions among all classes. The main contributions of this article are summarized as follows.

1) LDMR directly constrains the representation components instead of representation coefficients by fully considering the label information, which bridges the representation and identification phases. The locality structure of data characterized by the distances between test sample and different classes is utilized to learn more discriminative representation.
2) To further improve the discrimination of representation, a weighted pairwise class competitive constraint is incorporated into LDMR. This constraint has an impact on all pairs of classes, which reduces the correlations and enhances the competitions among all classes.
3) An iterative optimization based on alternating direction method of multipliers (ADMM) framework is presented to solve the LDMR model efficiently. Extensive experiments are performed on several popular face data sets to demonstrate the effectiveness and robustness of LDMR compared with some state-of-the-art regression-based methods.

The rest of this article is organized as follows. Section II introduces the linear representation-based models and NMR method briefly. Section III illustrates our proposed method in detail. Section IV reports the experiment results and analysis. Section V concludes this article.

II. Regression-Based Models

In this section, we recall the linear representation-based models and NMR model in brief, which are the most-related foundations of our work.

A. Linear Representation-Based Models

Given a training data matrix $\mathbf{D} \in \mathbf{R}^{m\times n}$ with n samples and a query sample $\mathbf{y} \in \mathbf{R}^{m}$. Denote $\mathbf{x} \in \mathbf{R}^{n}$ as the target representation vector. The basic framework of linear representation-based models can be unified as follows:

$$\min_{\mathbf{x}} \ \|\mathbf{y}-\mathbf{D}\mathbf{x}\|_2^2 + \alpha\|\mathbf{x}\|_p^p \tag{1}$$

where α is a balance parameter and $p = 1, 2$ in SRC [18] and CRC [20], respectively. Problem (1) with $p = 1$ is called Lasso regression [48] and Ridge regression with $p = 2$ [30]. Solving the Lasso regression is not easy and time-consuming [18], whereas Ridge regression be solved efficiently by a closed-form solution.

Based on the basic framework (1), sample weights learning, also called locality constraints, and feature weights learning are introduced to learn more discriminative representation. Generally, these models can be described as

$$\min_{\mathbf{x}} \ \|\mathbf{s} \odot (\mathbf{y} - \mathbf{D}\mathbf{x})\|_q^q + \alpha \|\mathbf{w} \odot \mathbf{x}\|_p^p \tag{2}$$

where $\odot$ denotes the elementwise product. Vector $\mathbf{s}$ and $\mathbf{w}$ can be induced from prior knowledge, which use the locality structure of data to learn a discriminative representation vector. When $\mathbf{s} = \mathbf{w} = \mathbf{1}$, problem (2) is degraded to SRC with $p = 1$ and CRC with $p = 2$. In WSRC [40], $\mathbf{s} = \mathbf{1}$ and $\mathbf{w}$ is determined by the distances from test sample to all training samples, which encourages the neighbors to contribute more to representation. In [29], RRC adaptively learns the feature weights $\mathbf{s}$ to improve the role of important features and eliminate the influence of outliers. IRGSC combines iterative sample weights learning and feature weights learning together to learns a more flexible $\mathbf{s}$ [37].

B. Nuclear Norm-Based Matrix Regression

Different from the 1-D linear representation methods that measure the regression error in vector, NMR directly estimates the 2-D error matrix [30]. Given a set of training image matrices, $\mathbf{B} = \{\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_c\}$ from c classes and a query image matrix $\mathbf{Y} \in \mathbf{R}^{u \times v}$, where $\mathbf{B}_i$ denotes the training samples of the ith class. $\mathbf{B}_{ij} \in \mathbf{R}^{u \times v}$ is the jth image matrix of class i. The ith class contains n_i training samples, and the total number of training samples is $n = \sum_{i=1}^{c} n_i$. NMR also assumes that the query sample $\mathbf{Y}$ can be linearly represented by all training samples, i.e.,

$$\mathbf{Y} = \mathbf{B}_1(\mathbf{x}_1) + \mathbf{B}_2(\mathbf{x}_2) + \cdots + \mathbf{B}_c(\mathbf{x}_c) + \mathbf{E} \tag{3}$$

where $\mathbf{E} \in \mathbf{R}^{u \times v}$ is an error matrix, $\mathbf{B}_i(\mathbf{x}_i)$ is the representation component of the ith class, i.e.,

$$\mathbf{B}_i(\mathbf{x}_i) = x_{i1}\mathbf{B}_{i1} + x_{i2}\mathbf{B}_{i2} + \cdots + x_{in_i}\mathbf{B}_{in_i}. \tag{4}$$

Let $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_c]^T$ denote the representation vector with $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in_i})$. Equation (3) can be compactly expressed as:

$$\mathbf{Y} = \mathbf{B}(\mathbf{x}) + \mathbf{E}. \tag{5}$$

Focusing on the low-rank information of error matrix $\mathbf{E}$, NMR solves the following rank function minimization problem [30]:

$$\min_{\mathbf{x}} \ \mathrm{rank}(\mathbf{E}) \quad \text{s.t. } \mathbf{E} = \mathbf{Y} - \mathbf{B}(\mathbf{x}). \tag{6}$$

Since problem (6) is NP-hard, it is relaxed to the following tractable formulation:

$$\min_{\mathbf{x}} \ \|\mathbf{E}\|_* + \alpha\|\mathbf{x}\|_p^p \quad \text{s.t. } \mathbf{E} = \mathbf{Y} - \mathbf{B}(\mathbf{x}). \tag{7}$$

In problem (7), nuclear norm is used to estimate the regression error to capture its low-rank characteristics. For the convenience of derivation, let $\mathbf{D} = [\mathrm{Vec}(\mathbf{B}_{11}), \mathrm{Vec}(\mathbf{B}_{12}), \ldots, \mathrm{Vec}(\mathbf{B}_{cn_c})] \in \mathbb{R}^{(uv) \times n}$, where $\mathrm{Vec}(\cdot)$ is an operator that stretches a matrix to vector, and $\mathbf{y} = \mathrm{Vec}(\mathbf{Y}) \in \mathbb{R}^{(uv) \times 1}$. The representation result $\mathbf{B}(\mathbf{x})$ can be transformed into vector form $\mathbf{D}\mathbf{x}$, i.e., $\mathbf{D}\mathbf{x} = \mathrm{Vec}(\mathbf{B}(\mathbf{x}))$. NMR model (7) can be unified as follows:

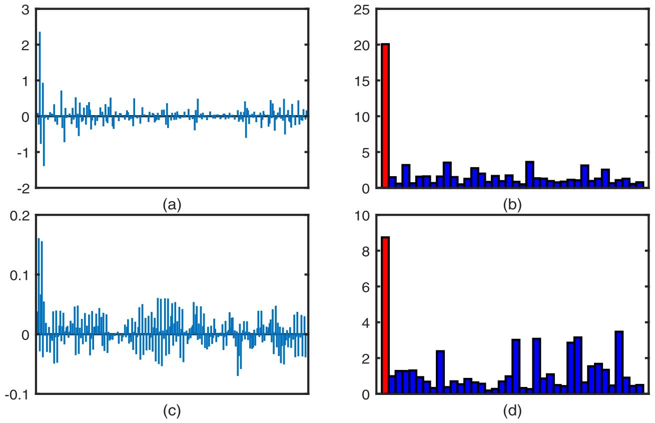


Fig. 1. (a) Representation coefficients and (b) representation components of SRC. (c) Representation coefficients and (d) representation components of CRC. The red marker in (b) and (d) correspond to the correct class.

$$\min_{\mathbf{x}} \ \|\mathrm{Mat}(\mathbf{y} - \mathbf{D}\mathbf{x})\|_* + \alpha\|\mathbf{x}\|_p^p \tag{8}$$

where $\mathrm{Mat}(\cdot)$ is an operator that converts a vector into matrix, i.e., the inverse operator of $\mathrm{Vec}(\cdot)$. In problem (8), $p = 1$ and $p = 2$ correspond to NMR_L1 [31] and NMR [30], respectively.

III. Proposed Method

In this section, we present the formulation, optimization, and discussions about convergence and complexity issues of our proposed LDMR method in detail.

A. Formulation

Although those regression methods mentioned earlier are various in representation learning, they all use class-specific representation losses, i.e., $\{\mathrm{Loss}(\mathbf{y} - \mathbf{D}_i\mathbf{x}_i)\}_{i=1}^{c}$, where $\mathbf{D}_i$ and $\mathbf{x}_i$ denote the training matrix and coefficients of the ith class respectively, to predict the label of test sample $\mathbf{y}$. From the view of classification, it is expected that $\mathrm{Loss}(\mathbf{y} - \mathbf{D}_k\mathbf{x}_k)$ can be small, whereas $\{\mathrm{Loss}(\mathbf{y} - \mathbf{D}_i\mathbf{x}_i)\}_{i=1, i\neq k}^{c}$ can be as large as possible if the ground-truth label of $\mathbf{y}$ is k. In linear representation methods, the test sample $\mathbf{y}$ is collaboratively represented as $\mathbf{y} \approx \mathbf{D}\mathbf{x}$. Most existing approaches focus on the coefficients $\mathbf{x}$, and various regularization terms are designed to learn discriminative representation. Actually, compared with representation coefficients $\mathbf{x}$, the representation components $\{\mathbf{D}_i\mathbf{x}_i\}_{i=1}^{c}$ have a direct connection with final identification process. For example, Fig. 1 shows the representation coefficients and components of SRC and CRC when a given test sample is identified correctly. ℓ^2-norm is used to measure the representation components. It can be clearly seen that the correct class takes most responsibility for final representation result, whereas other irrelevant classes make little contribution. Thus, it is reasonable to penalize the representation components of irrelevant classes (i.e., $\{\mathbf{D}_j\mathbf{x}_j\}_{j\neq k}$). However, the class information k is unknown. Under this observation, all the representation components are penalized here. The preliminary objective function of LDMR can be described as follows:

$$\min_{\mathbf{x}} \ \|\mathrm{Mat}(\mathbf{y}-\mathbf{D}\mathbf{x})\|_* + \frac{\alpha}{2}\sum_{i=1}^{c}\|\mathbf{D}_i\mathbf{x}_i\|_2^2. \tag{9}$$

In (9), we adopt nuclear norm to measure representation losses and penalize all c representation components, which directly bridges representation and classification.

Problem (9) treats all representation components equally, which ignores the locality structure of data. It is known that the samples of multiclasses lie in multiple subspaces [49]. Generally, the probability that a test sample comes from a close subspace is higher than those far subspaces. In TPTSR [28], linear regression is performed in advance to select the NNs. Inspired by this issue, we first solve the following problem:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{y}-\mathbf{D}\mathbf{x}\|_2^2 = (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{y}. \tag{10}$$

If $\mathbf{D}^T\mathbf{D}$ is singular, we can solve $\mathbf{x}$ by $(\mathbf{D}^T\mathbf{D}+\lambda\mathbf{I})^{-1}\mathbf{D}^T\mathbf{y}$, where λ is a small positive constant and $\mathbf{I}$ is an identity matrix. To reveal the local structure of test sample, we can use the classwise representation errors to characterize the distances between $\mathbf{y}$ and subspaces S_i. The representation error of the ith class can be computed by

$$r_i = \|\mathbf{y}-\mathbf{D}_i\hat{\mathbf{x}}_i\|_2. \tag{11}$$

Then, we define the weight w_i of the ith class by

$$w_i = \exp\left(\frac{\mathrm{dist}(\mathbf{y}, S_i)}{\delta}\right), \quad \mathrm{dist}(\mathbf{y}, S_i) = \frac{r_i - r_{\min}}{r_{\max} - r_{\min}} \tag{12}$$

where δ is a bandwidth parameter and $r_{\min}$ and $r_{\max}$ denote the minimum and maximum of errors $\{r_i\}_{i=1}^{c}$, respectively. $\mathrm{dist}(\mathbf{y}, S_i)$ characterizes the distance from $\mathbf{y}$ to the subspace S_i spanned by training samples $\mathbf{D}_i$. It can be easily obtained that $\mathrm{dist}(\mathbf{y}, S_i)$ lies in the interval [0, 1].

By incorporating the class weights (12), the LDMR model becomes

$$\min_{\mathbf{x}} \ \|\mathrm{Mat}(\mathbf{y}-\mathbf{D}\mathbf{x})\|_* + \frac{\alpha}{2}\sum_{i=1}^{c} w_i\|\mathbf{D}_i\mathbf{x}_i\|_2^2. \tag{13}$$

In problem (13), different classes are assigned with different prior weights. The classes that are far from test sample $\mathbf{y}$ are imposed more penalization in representation. On the other hand, problem (13) pays more attention to interclass difference since the c representation components are directly constrained rather than representation coefficients. The difference between training samples in the same class is not much concerned, which alleviates the influence of outliers. To further reduce the correlation of different classes and improve the discrimination of representation, we integrate the pairwise competition relationships among all classes into LDMR. The final LDMR model can be written as follows:

$$\begin{aligned}\min_{\mathbf{x}} \|\mathrm{Mat}(\mathbf{y}-\mathbf{D}\mathbf{x})\|_* &+ \frac{\alpha}{2}\sum_{i=1}^{c} w_i\|\mathbf{D}_i\mathbf{x}_i\|_2^2 \\ &+ \frac{\beta}{2}\sum_{i=1}^{c}\sum_{j=1}^{c}(w_i\mathbf{D}_i\mathbf{x}_i)^T(w_j\mathbf{D}_j\mathbf{x}_j). \end{aligned} \tag{14}$$

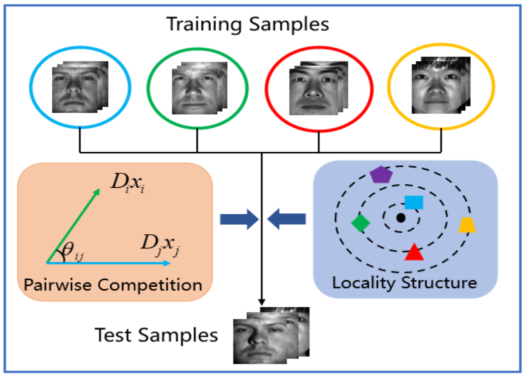


Fig. 2. Framework of LDMR. LDMR directly constrains the representation component of each class, and locality structure and pairwise class competition are integrated to enhance the discrimination of representation.

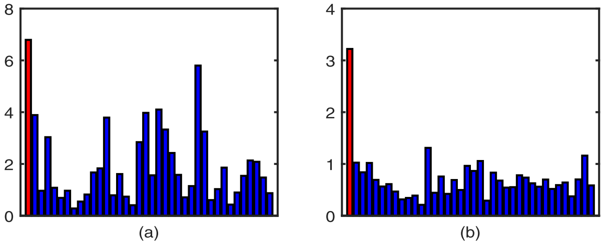


Fig. 3. Representation components of all classes of (a) NMR and (b) LDMR. The red marker corresponds to the correct class.

The LDMR model (14) takes the label information, locality structure, and pairwise class competition into consideration simultaneously. Compared with (13), the third term in (14) impacts all pairs of classes, which enhances the competition among them and further improves the discrimination of representation. Fig. 2 shows the overall framework of LDMR, which seamlessly integrate the class competitions and locality structure into representation learning. Since NMR is closely related to LDMR, we make a comparison between the two methods. We select a face image from subset 4 of Extended Yale B database for testing and subset 1 for training. Fig. 3 shows the representation components of all classes produced by NMR and LDMR. Though both NMR and LDMR give correct identification results, LDMR significantly depresses the irrelevant classes and achieves more discriminative representation, which is beneficial for identification.

B. Optimization

For solving problem (14), we first introduce an auxiliary variable $\mathbf{e}$ and rewrite the original problem as follows:

$$\begin{aligned}\min_{\mathbf{e},\mathbf{x}} \ &\|\mathrm{Mat}(\mathbf{e})\|_* + \frac{\alpha}{2}\sum_{i=1}^{c} w_i\|\mathbf{D}_i\mathbf{x}_i\|_2^2 \\ &+ \frac{\beta}{2}\sum_{i=1}^{c}\sum_{j=1}^{c}(w_i\mathbf{D}_i\mathbf{x}_i)^T(w_j\mathbf{D}_j\mathbf{x}_j) \\ \text{s.t. } &\mathbf{e} = \mathbf{y}-\mathbf{D}\mathbf{x}. \end{aligned} \tag{15}$$

Solving (15) is equivalent to solve the augmented Lagrange function $\mathcal{L}_\rho$ that is defined as

$$\begin{aligned}\mathcal{L}_\rho = \|\mathrm{Mat}(\mathbf{e})\|_* &+ \frac{\alpha}{2}\sum_{i=1}^{c} w_i\|\mathbf{D}_i\mathbf{x}_i\|_2^2\\ &+\frac{\beta}{2}\sum_{i=1}^{c}\sum_{j=1}^{c}(w_i\mathbf{D}_i\mathbf{x}_i)^T(w_j\mathbf{D}_j\mathbf{x}_j)\\ &+\mathbf{z}^T(\mathbf{y}-\mathbf{D}\mathbf{x}-\mathbf{e})+\frac{\rho}{2}\|\mathbf{y}-\mathbf{D}\mathbf{x}-\mathbf{e}\|_2^2\end{aligned}\tag{16}$$

where $\mathbf{z}$ is the Lagrange multiplier vector and $\rho > 0$ is a penalty factor. The augmented Lagrange function can be minimized by solving the subproblems with respect to each unknown variable iteratively.

Step 1 (Update $\mathbf{e}$*):* With $\mathbf{x}$ and $\mathbf{z}$ fixed, $\mathbf{e}$ can be updated by solving the following problem:

$$\min_{\mathbf{e}} \|\mathrm{Mat}(\mathbf{e})\|_* + \frac{\rho}{2}\left\|\mathbf{y}-\mathbf{D}\mathbf{x}-\mathbf{e}+\frac{1}{\rho}\mathbf{z}\right\|_2^2 \tag{17}$$

which is equivalent to

$$\min_{\mathbf{e}} \ \|\mathrm{Mat}(\mathbf{e})\|_* + \frac{\rho}{2}\left\|\mathrm{Mat}(\mathbf{e})-\mathrm{Mat}(\mathbf{y}-\mathbf{D}\mathbf{x}+\frac{1}{\rho}\mathbf{z})\right\|_F^2 \tag{18}$$

where $\|\cdot\|_F$ is the Frobenius norm. Problem (18) has a closed-form solution, i.e.,

$$\mathrm{Mat}(\tilde{\mathbf{e}}) = \Psi_{1/\rho}(\mathrm{Mat}(\mathbf{y}-\mathbf{D}\mathbf{x}+\mathbf{z}/\rho)) \tag{19}$$

where $\Psi_{1/\rho}(\cdot)$ is the singular value shrinkage operator [50]. Thus, the optimum of problem (17) is

$$\tilde{\mathbf{e}} = \mathrm{Vec}(\Psi_{1/\rho}(\mathbf{G})) \tag{20}$$

where $\mathbf{G} = \mathrm{Mat}(\mathbf{y}-\mathbf{D}\mathbf{x}+\mathbf{z}/\rho)$.

Step 2 (Update $\mathbf{x}$*):* With $\mathbf{e}$ and $\mathbf{z}$ fixed, $\mathbf{x}$ can be updated by solving the following problem:

$$\begin{aligned}\min_{\mathbf{x}}\ \frac{\alpha}{2}\sum_{i=1}^{c} w_i\|\mathbf{D}_i\mathbf{x}_i\|_2^2 &+ \frac{\rho}{2}\|\mathbf{y}-\mathbf{D}\mathbf{x}-\mathbf{e}+\frac{1}{\rho}\mathbf{z}\|_2^2\\ &+\frac{\beta}{2}\sum_{i=1}^{c}\sum_{j=1}^{c}(w_i\mathbf{D}_i\mathbf{x}_i)^T(w_j\mathbf{D}_j\mathbf{x}_j).\end{aligned}\tag{21}$$

It can be easily obtained that the above objective function is convex and smooth, and therefore, we can get the optimum by setting its derivative with respect to $\mathbf{x}$ to zero.

First, for $f_1(\mathbf{x}) = (\rho/2)\|\mathbf{y}-\mathbf{D}\mathbf{x}-\mathbf{e}+\mathbf{z}/\rho\|_2^2$, its derivative over $\mathbf{x}$ can be easily computed by

$$\partial f_1(\mathbf{x})/\partial\mathbf{x} = -\rho\mathbf{D}^T(\mathbf{y}-\mathbf{D}\mathbf{x}-\mathbf{e}+\mathbf{z}/\rho). \tag{22}$$

Then, we need to determine the derivative of the other two terms in (21) with respect to $\mathbf{x}$.

For $f_2(\mathbf{x}) = (\alpha/2)\sum_{i=1}^{c} w_i\|\mathbf{D}_i\mathbf{x}_i\|_2^2$, since it does not explicitly contain $\mathbf{x}$, we construct a diagonal matrix $\mathbf{H}_i = \mathrm{diag}(0,\ldots,0,1,\ldots,1,0,\ldots,0) \in \mathbb{R}^{n\times n}$ in which the elements corresponding to the ith class are 1 and others are 0. With the help of $\mathbf{H}_i$, it has $\|\mathbf{D}_i\mathbf{x}_i\|_2^2 = \|\mathbf{D}\mathbf{H}_i\mathbf{x}\|_2^2$, and we can obtain the derivative of $f_2(\mathbf{x})$ as follows:

$$\begin{aligned}\frac{\partial}{\partial\mathbf{x}}\left(\frac{\alpha}{2}\sum_{i=1}^{c} w_i\|\mathbf{D}_i\mathbf{x}_i\|_2^2\right) &= \frac{\partial}{\partial\mathbf{x}}\left(\frac{\alpha}{2}\sum_{i=1}^{c} w_i\|\mathbf{D}\mathbf{H}_i\mathbf{x}\|_2^2\right)\\ &= \alpha\sum_{i=1}^{c} w_i\mathbf{H}_i\mathbf{D}^T\mathbf{D}\mathbf{H}_i\mathbf{x}.\end{aligned}\tag{23}$$

For $f_3(\mathbf{x}) = (\beta/2)\sum_{i=1}^{c}\sum_{j=1}^{c}(w_i\mathbf{D}_i\mathbf{x}_i)^T(w_j\mathbf{D}_j\mathbf{x}_j)$, we first obtain the partial derivatives $\partial f_3/\partial\mathbf{x}_k$ and then exploit all $\partial f_3/\partial\mathbf{x}_k$ $(k = 1, 2, \ldots, c)$ to achieve $\partial f_3/\partial\mathbf{x}$

$$\begin{aligned}(w_i\mathbf{D}_i\mathbf{x}_i)^T(w_j\mathbf{D}_j\mathbf{x}_j) = \frac{1}{2}\big(&\|w_i\mathbf{D}_i\mathbf{x}_i + w_j\mathbf{D}_j\mathbf{x}_j\|_2^2\\ &- \|w_i\mathbf{D}_i\mathbf{x}_i\|_2^2 - \|w_j\mathbf{D}_j\mathbf{x}_j\|_2^2\big).\end{aligned}\tag{24}$$

It is noted that $w_i\mathbf{D}_i$ can be merged by transforming original data $\mathbf{D}_i$ to weighted data $\mathbf{D}_i^w$, i.e., $\mathbf{D}_i^w = w_i\mathbf{D}_i$ and $\mathbf{D}_j^w = w_j\mathbf{D}_j$. For clarity, we give the following definition on weighted data.

Definition 1: Given a matrix $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n] \in \mathbb{R}^{m\times n}$ whose each column represents a training sample, the weighted matrix $\mathbf{A}^w$ is defined as $\mathbf{A}^w = [\mathbf{a}_1^w, \mathbf{a}_2^w, \ldots, \mathbf{a}_n^w]$ with

$$\mathbf{a}_i^w = w_i\mathbf{a}_i, \quad i = 1, 2, \ldots, n$$

where w_i is the weight of the class to which $\mathbf{a}_i$ belongs.

Based on (24) and Definition 1, we rewrite $f_3(\mathbf{x})$ as

$$\begin{aligned}&f_3(\mathbf{x})\\ &= \frac{\beta}{2}\sum_{i=1}^{c}\sum_{j=1}^{c}(w_i\mathbf{D}_i\mathbf{x}_i)^T(w_j\mathbf{D}_j\mathbf{x}_j)\\ &= \frac{\beta}{4}\Bigg[\sum_{\substack{i=1\\ i\neq k}}^{c}\left(\|\mathbf{D}_i^w\mathbf{x}_i + \mathbf{D}_k^w\mathbf{x}_k\|_2^2 - \|\mathbf{D}_i^w\mathbf{x}_i\|_2^2 - \|\mathbf{D}_k^w\mathbf{x}_k\|_2^2\right)\\ &\qquad + \sum_{\substack{j=1\\ j\neq k}}^{c}\left(\|\mathbf{D}_k^w\mathbf{x}_k + \mathbf{D}_j^w\mathbf{x}_j\|_2^2 - \|\mathbf{D}_k^w\mathbf{x}_k\|_2^2 - \|\mathbf{D}_j^w\mathbf{x}_j\|_2^2\right)\\ &\qquad + \sum_{\substack{i=1\\ i\neq k}}^{c}\sum_{\substack{j=1\\ j\neq k}}^{c}\left(\|\mathbf{D}_i^w\mathbf{x}_i + \mathbf{D}_j^w\mathbf{x}_j\|_2^2 - \|\mathbf{D}_i^w\mathbf{x}_i\|_2^2 - \|\mathbf{D}_j^w\mathbf{x}_j\|_2^2\right)\Bigg]\\ &= \frac{\beta}{2}\sum_{\substack{i=1\\ i\neq k}}^{c}\left(\|\mathbf{D}_i^w\mathbf{x}_i + \mathbf{D}_k^w\mathbf{x}_k\|_2^2 - \|\mathbf{D}_i^w\mathbf{x}_i\|_2^2 - \|\mathbf{D}_k^w\mathbf{x}_k\|_2^2\right)\\ &\quad + \frac{\beta}{4}\sum_{\substack{i=1\\ i\neq k}}^{c}\sum_{\substack{j=1\\ j\neq k}}^{c}\left(\|\mathbf{D}_i^w\mathbf{x}_i + \mathbf{D}_j^w\mathbf{x}_j\|_2^2 - \|\mathbf{D}_i^w\mathbf{x}_i\|_2^2 - \|\mathbf{D}_j^w\mathbf{x}_j\|_2^2\right).\end{aligned}$$

Since we consider the partial derivatives $\partial f_3/\partial\mathbf{x}_k$, based on the above equation, we can obtain

$$\begin{aligned}\frac{\partial f_3(\mathbf{x})}{\partial\mathbf{x}_k} &= \frac{\partial}{\partial\mathbf{x}_k}\left(\frac{\beta}{2}\sum_{\substack{i=1\\ i\neq k}}^{c}\left(\|\mathbf{D}_i^w\mathbf{x}_i + \mathbf{D}_k^w\mathbf{x}_k\|_2^2 - \|\mathbf{D}_k^w\mathbf{x}_k\|_2^2\right)\right)\\ &= \beta\sum_{\substack{i=1\\ i\neq k}}^{c}\left(\left(\mathbf{D}_k^w\right)^T\left(\mathbf{D}_i^w\mathbf{x}_i + \mathbf{D}_k^w\mathbf{x}_k\right) - \left(\mathbf{D}_k^w\right)^T\mathbf{D}_k^w\mathbf{x}_k\right)\end{aligned}$$

$$= \beta\left[\left(\sum_{i=1}^{c}\left(\mathbf{D}_k^w\right)^T\mathbf{D}_i^w\mathbf{x}_i\right) - \left(\mathbf{D}_k^w\right)^T\mathbf{D}_k^w\mathbf{x}_k\right]$$

$$= \beta\left(\mathbf{D}_k^w\right)^T\mathbf{D}^w\mathbf{x} - \beta\left(\mathbf{D}_k^w\right)^T\mathbf{D}_k^w\mathbf{x}_k.$$

Thus, the derivative of $f_3(\mathbf{x})$ over $\mathbf{x}$ is

$$\frac{\partial f_3(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_3}{\partial \mathbf{x}_1} \\ \vdots \\ \frac{\partial f_3}{\partial \mathbf{x}_c} \end{bmatrix} = \begin{bmatrix} \beta\left(\mathbf{D}_1^w\right)^T\mathbf{D}^w\mathbf{x} - \beta\left(\mathbf{D}_1^w\right)^T\mathbf{D}_1^w\mathbf{x}_1 \\ \vdots \\ \beta\left(\mathbf{D}_c^w\right)^T\mathbf{D}^w\mathbf{x} - \beta\left(\mathbf{D}_c^w\right)^T\mathbf{D}_c^w\mathbf{x}_c \end{bmatrix}$$
$$= \beta(\mathbf{D}^w)^T\mathbf{D}^w\mathbf{x} - \beta\mathbf{M}\mathbf{x} \qquad (25)$$

where $\mathbf{M} = \begin{pmatrix} (\mathbf{D}_1^w)^T\mathbf{D}_1^w & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & (\mathbf{D}_c^w)^T\mathbf{D}_c^w \end{pmatrix}$.

Now, we have calculated all the derivatives of the three terms, i.e., $f_1(\mathbf{x})$, $f_2(\mathbf{x})$, and $f_3(\mathbf{x})$, in objective function (21) over $\mathbf{x}$. Combining Eqs. (22), (23), and (25), the derivative over $\mathbf{x}$ of objective function (21) [denoted as $f(\cdot)$] is

$$\frac{\partial f}{\partial \mathbf{x}} = -\rho\mathbf{D}^T(\mathbf{y} - \mathbf{D}\mathbf{x} - \mathbf{e} + \mathbf{z}/\rho) + \alpha\sum_{i=1}^{c} w_i\mathbf{H}_i\mathbf{D}^T\mathbf{D}\mathbf{H}_i\mathbf{x}$$
$$+ \beta(\mathbf{D}^w)^T\mathbf{D}^w\mathbf{x} - \beta\mathbf{M}\mathbf{x}. \qquad (26)$$

If $\partial f/\partial \mathbf{x} = 0$, we can obtain the optimal solution $\tilde{\mathbf{x}}$ for objective function (21) as follows:

$$\tilde{\mathbf{x}} = \mathbf{P}(\mathbf{y} - \mathbf{e} + \mathbf{z}/\rho) \qquad (27)$$

where

$$\mathbf{P} = \left[\mathbf{D}^T\mathbf{D} + \frac{\alpha}{\rho}\left(\sum_{i=1}^{c} w_i\mathbf{H}_i\mathbf{D}^T\mathbf{D}\mathbf{H}_i\right)\right.$$
$$\left. + \frac{\beta}{\rho}((\mathbf{D}^w)^T\mathbf{D}^w - \mathbf{M})\right]^{-1}\mathbf{D}^T. \qquad (28)$$

Step 3 (Update $\mathbf{z}$*):* Fix $\mathbf{e}$ and $\mathbf{x}$, and we can update the Lagrange multiplier vector $\mathbf{z}$ by

$$\mathbf{z} = \mathbf{z} + \rho(\mathbf{y} - \mathbf{D}\mathbf{x} - \mathbf{e}). \qquad (29)$$

Stopping Criteria: ADMM is an iterative algorithm and it is necessary to adopt appropriate stopping criteria. Following the direction of [30] and [51], the convergence criteria of LDMR can be described as follows:

$$\|\mathbf{r}^k\|_2 \le \epsilon_{\text{pri}} \text{ and } \|\mathbf{s}^{\text{k}}\|_2 \le \epsilon_{\text{dual}} \qquad (30)$$

where $\mathbf{r}^k$ and $\mathbf{s}^k$, respectively, represent the primal and dual residual defined as follows:

$$\begin{cases} \mathbf{r}^k = \mathbf{y} - \mathbf{D}\mathbf{x}^k - \mathbf{e} \\ \mathbf{s}^k = \rho\mathbf{D}^T(\mathbf{e}^k - \mathbf{e}^{k-1}) \\ \epsilon_{\text{dual}} = \sqrt{n}\epsilon_{\text{abs}} + \epsilon_{\text{rel}}\|\mathbf{D}^T\mathbf{z}^k\|_2 \\ \epsilon_{\text{pri}} = \sqrt{pq}\epsilon_{\text{abs}} + \epsilon_{\text{rel}}\max\{\|\mathbf{D}\mathbf{x}^k\|_2, \|\mathbf{y}\|_2, \|\mathbf{e}\|_2\}. \end{cases}$$

With (20), (27), and (29), we can efficiently solve the proposed LDMR model. Algorithm 1 summarizes the optimization strategy for LDMR. Finally, we present a detailed explanation to Algorithm 1. Instead of initializing $\mathbf{x}$ to $\mathbf{0}$ in some previous works [30], [31], [37], [47], we use $\hat{\mathbf{x}}$, the coefficients vector of Ridge regression [i.e., (10)], as a starting point to accelerate the convergence of Algorithm 1, which will be illustrated in Section IV. The penalty parameter ρ is fixed in our algorithm to achieve better efficiency. In (28), $\mathbf{P}$ is fixed in iterations, and we can compute and store it in advance. Once the optimal regression coefficients $\mathbf{x}^\dagger$ for a test sample $\mathbf{y}$ are obtained by Algorithm 1, we use the classwise residuals to identify $\mathbf{y}$, i.e., identity$(\mathbf{y}) = \arg\min_i(e_i)$, where $e_i = \|\text{Mat}(\mathbf{D}\mathbf{x}^\dagger - \mathbf{D}_i\mathbf{x}_i^\dagger)\|_*$.

Algorithm 1 Algorithm for Solving LDMR

```
Input: Training matrix $\mathbf{D}$ and test sample $\mathbf{y}$, parameters $\lambda$, $\delta$, $\alpha$, $\beta$ and $\rho$, stopping criteria parameters $\epsilon_{abs}$ and $\epsilon_{rel}$.
Output: Optimal coefficients vector $\mathbf{x}^k$.
Compute $\hat{\mathbf{x}}$: $\hat{\mathbf{x}} = (\mathbf{D}^T\mathbf{D} + \lambda\mathbf{I})^{-1}\mathbf{D}^T\mathbf{y}$.
Compute class weights $\{w_i\}_{i=1}^c$:
    $w_i = \exp(\text{dist}(\mathbf{y}, S_i)/\delta), \text{dist}(\mathbf{y}, S_i) \propto \|\mathbf{y} - \mathbf{D}_i\hat{\mathbf{x}}_i\|$,
    where dist$(\mathbf{y}, S_i)$ is defined in equation (12).
Compute $\mathbf{P}$:
    $\mathbf{P} = \left[\mathbf{D}^T\mathbf{D} + \mathbf{G} + \frac{\beta}{\rho}\left((\mathbf{D}^w)^T\mathbf{D}^w - \mathbf{M}\right)\right]^{-1}\mathbf{D}^T$,
    where $\mathbf{G} = (\alpha/\rho)\left(\sum_{i=1}^c w_i\mathbf{H}_i\mathbf{D}^T\mathbf{D}\mathbf{H}_i\right)$, and $\mathbf{M} = \text{diag}[(\mathbf{D}_1^w)^T\mathbf{D}_1^w, \ldots, (\mathbf{D}_c^w)^T\mathbf{D}_c^w]$.
Initialization: $\mathbf{x}^0 = \hat{\mathbf{x}}$,  $\mathbf{e}^0 = \mathbf{y} - \mathbf{D}\mathbf{x}^0$.
while not converged do
    Update e: $\mathbf{e}^{k+1} = \text{Vec}(\Psi_{\frac{1}{\rho}}(\text{Mat}(\mathbf{y} - \mathbf{D}\mathbf{x}^k + \mathbf{z}^k/\rho))$.
    Update x: $\mathbf{x}^{k+1} = \mathbf{P}(\mathbf{y} - \mathbf{e}^{k+1} + \mathbf{z}^k/\rho)$.
    Update z: $\mathbf{z}^{k+1} = \mathbf{z}^k + \rho(\mathbf{y} - \mathbf{D}\mathbf{x}^{k+1} - \mathbf{e}^{k+1})$.
    $k := k + 1$.
end while
return $\mathbf{x}^k$.
```

C. Convergence and Computational Complexity Analysis

Problem (14) is typical nonconvex, and it is difficult to guarantee a global optimal solution. However, a local optimal solution can be obtained by using the ADMM framework. The classical two-block problem can be described as follows [51], [52]:

$$\min_{\mathbf{X}\in\mathbf{\Omega}_\mathbf{X}, \mathbf{Y}\in\mathbf{\Omega}_\mathbf{Y}} f(\mathbf{X}) + g(\mathbf{Y}) \quad \text{s.t. } \mathbf{U}\mathbf{X} + \mathbf{V}\mathbf{Y} = \mathbf{L} \qquad (31)$$

where $\mathbf{\Omega}_\mathbf{X}$ and $\mathbf{\Omega}_\mathbf{Y}$ are the domains of $\mathbf{X}$ and $\mathbf{Y}$, respectively, and $f(\cdot)$ and $g(\cdot)$ are convex functions. ADMM converts the original constrained problem (31) into its augmented Lagrangian function

$$\mathcal{L} = f(\mathbf{X}) + g(\mathbf{Y}) + \frac{\rho}{2}\|\mathbf{U}\mathbf{X} + \mathbf{V}\mathbf{Y} - \mathbf{L} + \frac{1}{\rho}\mathbf{Z}\|_F^2 - \frac{1}{2\rho}\|\mathbf{Z}\|_F^2.$$

Then, ADMM iteratively updates variables as follows to minimize the objective function, i.e.,

$$\begin{cases} \mathbf{X}^{k+1} = \arg\min_{\mathbf{X}\in\mathbf{\Omega}_\mathbf{X}} \mathcal{L}_\rho(\mathbf{X}, \mathbf{Y}^k, \mathbf{Z}^k) \\ \mathbf{Y}^{k+1} = \arg\min_{\mathbf{Y}\in\mathbf{\Omega}_\mathbf{Y}} \mathcal{L}_\rho(\mathbf{X}^{k+1}, \mathbf{Y}, \mathbf{Z}^k) \\ \mathbf{Z}^{k+1} = \mathbf{Z}^k + \rho(\mathbf{U}\mathbf{X}^{k+1} + \mathbf{V}\mathbf{Y}^{k+1} - \mathbf{L}). \end{cases} \qquad (32)$$

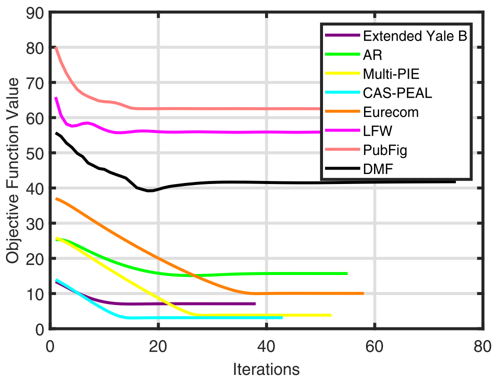


Fig. 4. Convergence curves of the proposed algorithm on different data sets.

From Algorithm 1, it is obvious that our optimization problem is consistent with classical two-block ADMM. Specifically, optimizing $\mathbf{e}$ is equivalent to optimize $\mathbf{X}$ in (32), and the optimization of $\mathbf{x}$ is equivalent to optimize $\mathbf{Y}$ in (32). For two-block ADMM, the convergence property has been theoretically proved in [51]. We experimentally illustrate the convergence property of the proposed algorithm in Fig. 4. It can be observed that the objective function value obviously decrease to a stable value, which indicates that our algorithm can converge fast, usually within several tens of iterations.

Computational complexity is another important issue when estimating the performance of an algorithm. From Algorithm 1, the major computational complexity of LDMR consists of two parts: the matrix inverse computation outside and the singular value decomposition (SVD) inside iterations. Given the image size $u \times v$ and the number of training samples n, let $m = u \times v$ denote the dimension of each sample. It takes $\mathrm{O}(mn^2)$ on the computation of $\mathbf{D}^T\mathbf{D}$ and $(\mathbf{D}^w)^T\mathbf{D}^w$. The cost for $\mathbf{H}_i^T\mathbf{D}^T\mathbf{D}\mathbf{H}_i$ and $\mathbf{M}$ in (25) is $\mathrm{O}(mt^2)$, assuming that each class has average t training samples. Thus, the total consumption of computing $\mathbf{P}$ is $\mathrm{O}(mn^2 + n^3)$, where $\mathrm{O}(n^3)$ is the cost of matrix inverse operation. For SVD in step 6, the computational complexity is $\mathrm{O}(uv^2)$ assuming that $u > v$, which is relevant to the image size. For step 7, the computational complexity is $\mathrm{O}(mn)$. Thus, the total complexity of LDMR is about $\mathrm{O}(mn^2 + n^3 + \tau(uv^2 + mn))$ if there are τ iterations.

IV. Experiments

A. Experimental Settings

Several popular face data sets are used in our experiments, including the Extended Yale B,[1] Multi-PIE,[2] AR,[3] CAS-PEAL,[4] EURECOM Kinect,[5] LFW,[6] PubFig,[7] and Disguised and Makeup Faces (DMF)[8] data sets, to evaluate the effectiveness and robustness of the proposed method against illumination changes, random occlusion, and real-world disguises.

The Extended Yale B (YaleB) data set contains 2414 frontal face images over 38 individuals in different illumination conditions. The whole data set is divided into five subsets. From subsets 1 to 5, the face images characterize slight-moderate-severe illumination changes. All face images are resized to 48 × 42 pixels [38].

The Multi-PIE data set contains face images over 337 individuals under 15 poses and 19 illumination conditions. We utilize total 4470 images of 149 individuals with different illumination and pose changes in Session 1. All the images are resized to 50 × 40 pixels [38].

The AR face data set contains over 4000 face images of 126 individuals with different illumination, expression, and occlusion (i.e., sunglasses and scarves) changes. The whole data set consists of two sessions. Total 3120 face images of 120 individuals are used in our experiments. All images are cropped and resized to 50 × 40 pixels [38].

The whole CAS-PEAL database contains over 90 000 images of 1040 individuals with varying poses, expressions, illumination, accessories, and so on. We choose 438 individuals from Accessory category for testing, and their corresponding neutral face images from Normal category for training. All the images are resized to 32 × 32 pixels [53].

The EURECOM Kinect face data set contains images of 52 individuals with different facial variations. In our experiment, we select 18 images per person, including 12 nonoccluded images and six images with sunglasses, hands, and paper occlusions. All images are resized to 50×40 pixels [30].

LFW, PubFig, and DMF are wild face data sets in which the photographs are captured in uncontrolled scenarios. We use the LFW-a data set, a revised version of LFW, which consists of 1580 images over 158 individuals. For each individual, five images are randomly selected for training and the rest for testing. All images are cropped and resized to 32×32 pixels [37]. For PubFig, we randomly select 20 images for each person and total 100 individuals are used for our experiments. Ten images for each person are randomly selected for training and the rest for testing. All the images are cropped and resized to 64 × 64 pixels [37]. The DMF data set contains 2460 face images from 410 different subjects. We randomly select five images per person for training and the rest for testing. The images are resized to 64 × 64 pixels.

Several state-of-the-art regression-based FI methods are tested as comparisons, including SRC [18], CRC [20], CSC [38], RRC [29], SLRC [54], TPTSR [28], RCRC [27], WSRC [40], WCRC [41], ProCRC [27], IRGSC [37], and NMR [30]. Furthermore, CNN-based methods are also tested for comparison, which will be illustrated in Section IV-F. The l_1-norm minimization problem in SRC and WSRC is solved by the homotopy algorithm [18]. The balance parameter of CRC is fine-tuned to report their best results. RRC adopts the l_1-norm regularizer due to its relatively better performance than the l_2-norm regularizer. The parameter settings of other methods follow the authors' suggestions. For LDMR, we set the parameter $\lambda = 0.01$, $\epsilon_{\mathrm{abs}} = 10^{-3}$, and $\epsilon_{\mathrm{rel}} = 10^{-3}$.

[1] http://vision.ucsd.edu/content/extended-yale-face-database-b-b
[2] http://www.cs.cmu.edu/afs/cs/project/PIE/MultiPie/Multi-Pie/Home.html
[3] http://www2.ece.ohio-state.edu/~aleix/ARdatabase.html
[4] http://www.jdl.ac.cn/peal/home.htm
[5] http://www.rgb-d.eurecom.fr/
[6] http://vis-www.cs.umass.edu/lfw/
[7] https://www.cs.columbia.edu/CAVE/databases/pubfig/
[8] http://www4.comp.polyu.edu.hk/~csajaykr/DMFaces.htm

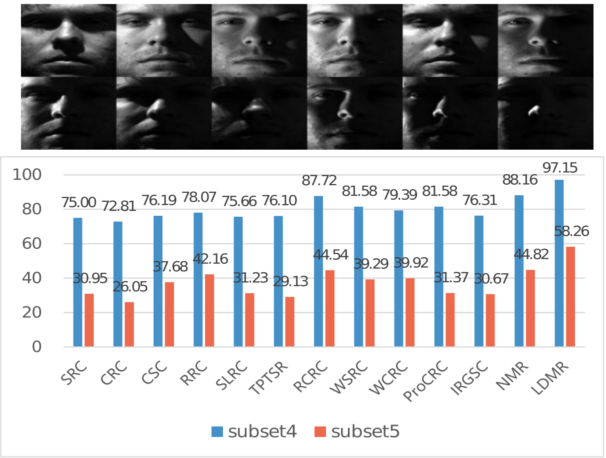


Fig. 5. Identification rates (%) of different methods on the YaleB data set.

TABLE I

AVERAGE IDENTIFICATION RATES (%) OF DIFFERENT METHODS ON THE MULTI-PIE DATA SET

Method	$T = 1$	$T = 3$	$T = 5$	$T = 8$
SRC	29.13	69.93	86.44	94.50
CRC	29.39	71.14	86.04	94.90
CSC	32.48	**76.38**	87.38	96.22
RRC	31.94	72.35	86.84	95.97
SLRC	29.48	70.10	85.10	93.04
TPTSR	29.97	71.09	86.98	95.17
RCRC	31.81	71.94	85.95	97.98
WSRC	31.58	71.56	86.84	95.86
WCRC	31.91	71.81	86.17	96.34
ProCRC	32.62	73.42	87.92	96.11
IRGSC	30.81	72.65	87.19	95.97
NMR	33.00	71.01	87.19	95.70
LDMR	**34.90**	75.30	**89.72**	**97.85**

The selection of other parameters in LDMR will be analyzed in Section IV-G.

B. FI With Illumination Changes

We first evaluate the performance of LDMR against illumination changes on the YaleB data set. We use subset 1 for training and subsets 4 and 5 for testing. Fig. 5 shows some test images and identification rates of all competing methods on two subsets. It can be clearly seen that LDMR is superior to other regression methods. Since the test images are severely contaminated by the shadows and reflections, SRC, CRC, and CSC seem not very robust to extreme illumination conditions. NMR, which ranks the second in all methods, shows robustness to illumination changes and obtains comparable performance due to its advantage in structural information exploring. However, LDMR achieves 8.99% and 13.44% improvements over NMR on subsets 4 and 5, respectively. On the other hand, the weighted methods WSRC and WCRC perform better than their nonweighted version SRC and CRC, respectively, which demonstrates locality structure and label information impact on the improvement of identification performance.

In the second experiment, we evaluate our method on the Multi-PIE data set. We randomly select T $(=1, 3, 5, 8)$ face images per person for training and the rest for testing. Table I lists the average identification rates of ten runs of different methods in all cases. It can be observed that the identification rates of all methods get improved with the increasing number of training samples. Our LDMR method outperforms most of the other compared methods except for the case of $T = 3$, in which the performance of CSC is slightly better than LDMR. Since the illumination conditions in the images of Multi-PIE data set are better than those in YaleB, RRC, RCRC, and ProCRC also achieve competitive results. These experimental results demonstrate the robustness of LDMR to extreme illumination conditions.

C. FI With Random Block Occlusion

In this section, we design four random block occlusions on the YaleB data set. We use subset 1 for training and subset 3 for testing [30]. Different types and levels of square block occlusion are randomly imposed on test images, and the locations of occlusion are unknown to algorithms. Fig. 6(a) shows some test samples.

In the first experiment, following the settings in [4], [18], [29], and [30], we randomly impose a baboon image on each test image as occlusion. Fig. 7(a) shows the identification rates of different methods on the YaleB data set with increasing level of occlusion. As is presented, LDMR is always superior to other methods and the difference becomes significant with the increment of occlusion. When the occlusion level is no more than 30%, RRC, RCRC, IRGSC, and NMR can achieve competitive accuracy. With the increase of occlusion level, the accuracies of other methods drop fast except for NMR. NMR achieves competitive results when the occlusion level is below 40%. However, the robustness of LDMR becomes outstanding when the occlusion level goes up, and the identification rate of LDMR is 6.47% and 9.15% higher than that of NMR under 50% and 60% occlusion level, respectively.

In the second experiment, each test image is occluded by a randomly selected face image from training images. The identification performance of all competitive methods is shown in Fig. 7(b). Generally, the experiment results of each method are consistent with those in Fig. 7(a). Under each level of occlusion, LDMR always achieves the best identification performance. When the occlusion level is small, the gap between other robust methods and LDMR is not significant. When the occlusion ratio goes up to 60%, the identification accuracy of LDMR is 9.91% higher than NMR. The experiment results demonstrate that LDMR is more robust and powerful to complicated contiguous occlusion compared with the other regression FI methods.

In the third experiment, two types of extreme occlusions are imposed on test images: square black block and white block with increasing level, as shown in Fig. 6(a). Tables II and III show the identification rates of all methods under two test protocols, respectively. LDMR is still superior to other methods in all levels of occlusion. When the occlusion level is 60%, LDMR can achieve an 86.10% identification rate on black block occlusion and 79.05% on white block occlusion.

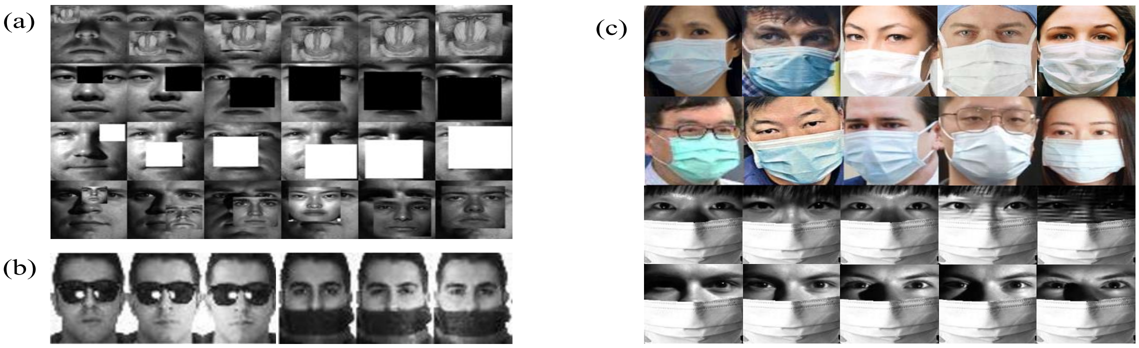
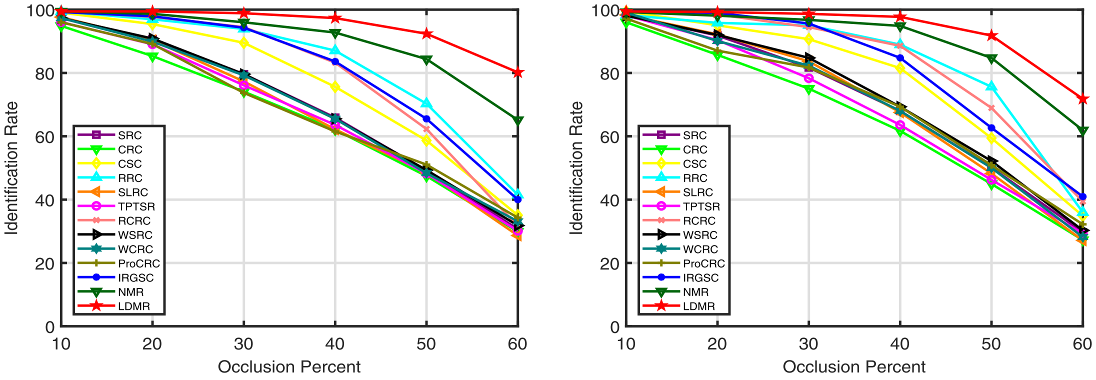


Fig. 6. Some face images used in our experiments. (a) Face images with six levels (i.e., 10%–60%) and four types (i.e., baboon, human face, black block, and white block) of occlusion from the YaleB data set. (b) Face images with sunglasses and scarves from the AR data set. (c) Top two rows contain face images with masks occlusion from the Internet, and the last two rows show some testing images in our experiments.



(a)

(b)

Fig. 7. Identification rates (%) of different methods on the YaleB data set with increasing level of occlusion. (a) Baboon image occlusion. (b) Face image occlusion.

TABLE II
IDENTIFICATION RATES (%) OF DIFFERENT METHODS ON THE YALEB DATA SET WITH DIFFERENT LEVELS OF BLACK BLOCK OCCLUSION

Method	10%	20%	30%	40%	50%	60%
SRC	97.52	83.24	64.57	48.38	36.95	28.00
CRC	95.05	79.81	64.95	54.01	40.19	28.95
CSC	98.67	90.48	68.38	43.05	27.62	15.24
RRC	99.24	98.29	97.52	93.90	80.00	43.43
SLRC	96.71	81.14	60.31	46.05	34.87	26.31
TPTSR	98.68	85.09	65.35	53.51	39.25	26.53
RCRC	99.43	93.90	74.48	54.86	38.86	18.86
WSRC	98.48	83.81	61.52	44.95	34.86	23.24
WCRC	98.29	83.24	63.04	54.29	38.86	27.05
ProCRC	98.29	90.67	79.24	68.95	60.76	44.57
IRGSC	99.43	98.67	97.52	96.38	91.81	73.71
NMR	99.43	99.05	97.90	95.43	92.38	82.10
LDMR	**99.81**	**99.43**	**98.86**	**97.71**	**94.29**	**86.10**

TABLE III
IDENTIFICATION RATES (%) OF DIFFERENT METHODS ON THE YALEB DATA SET WITH DIFFERENT LEVELS OF WHITE BLOCK OCCLUSION

Method	10%	20%	30%	40%	50%	60%
SRC	86.48	52.57	29.14	18.67	13.90	11.43
CRC	67.24	46.29	29.14	19.62	14.10	11.81
CSC	97.71	76.19	39.81	23.62	16.95	12.76
RRC	99.05	99.05	96.95	92.76	64.95	34.29
SLRC	82.23	49.56	27.19	17.32	10.96	9.21
TPTSR	74.56	50.44	31.14	18.64	14.47	10.09
RCRC	99.05	82.29	48.00	30.67	15.24	11.24
WSRC	93.52	54.10	29.33	19.24	16.00	12.57
WCRC	72.38	49.71	29.52	19.62	15.05	12.00
ProCRC	69.14	47.05	31.05	25.14	19.05	16.19
IRGSC	99.43	**99.43**	94.29	92.38	85.90	58.29
NMR	99.24	96.76	93.90	87.05	69.90	42.86
LDMR	**99.62**	**99.43**	**98.10**	**96.19**	**91.62**	**79.05**

NMR, IRGSC, and RRC also obtain good performance when the occlusion ratio is below 50%. It is interesting that IRGSC achieves comparable performance in both types of occlusion, which is much better than those in the baboon occlusion and face occlusion experiments. The main reason is that the differences between occluded and nonoccluded regions are significant, and the adaptive feature weights learning mechanism in IRGSC takes good effect and improves the performance in both cases. However, when the black block occlusion level is 60%, LDMR has an improvement of 4.00% and 12.39% over NMR and IRGSC, respectively. The performances of SRC, CRC, and CSC are not desirable in both cases. This demonstrates that our LDMR is more robust than others to various occlusions.

D. FI With Real-World Disguise

In the first experiment, AR and YaleB data sets are used to evaluate the performance of LDMR. For AR, we select

TABLE IV

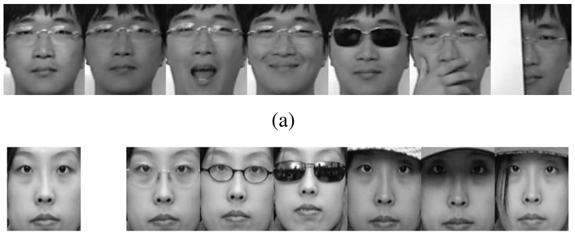
IDENTIFICATION RATES (%) OF DIFFERENT METHODS ON THE AR AND YALEB DATA SETS. AR-S1 DENOTES SESSION 1 OF THE AR DATA SET

Dataset		SRC	CRC	CSC	RRC	SLRC	TPTSR	RCRC	WSRC	WCRC	ProCRC	IRGSC	NMR	LDMR
AR-S1	sunglasses	92.50	90.00	92.50	86.94	88.89	91.67	91.39	93.06	93.06	93.61	88.61	93.06	**96.94**
	scarves	68.61	67.22	67.50	63.33	63.89	66.94	72.50	68.61	67.78	66.67	68.33	71.11	**76.39**
AR-S2	sunglasses	93.61	91.39	92.22	87.50	89.04	92.11	92.50	94.44	93.61	95.00	88.89	95.56	**97.22**
	scarves	62.22	59.44	57.78	60.00	59.87	60.09	64.44	62.78	61.39	60.56	61.11	66.11	**73.61**
YaleB	masks	30.28	30.10	38.10	76.38	27.86	31.58	57.33	32.19	35.05	50.47	84.76	80.95	**93.33**

TABLE V

IDENTIFICATION RATES (%) OF DIFFERENT METHODS ON THE LFW, PUBFIG, AND DMF FACE DATA SETS

Dataset	SRC	CRC	CSC	RRC	SLRC	TPTSR	RCRC	WSRC	WCRC	ProCRC	IRGSC	NMR	LDMR
LFW	36.46	38.99	44.81	38.29	35.19	39.24	43.67	36.71	39.24	44.30	43.52	42.05	**46.58**
PubFig	41.00	39.70	42.70	40.40	39.80	40.30	37.20	39.50	36.10	41.40	42.80	43.60	**44.50**
DMF	41.22	40.24	45.93	44.15	39.84	42.28	45.85	42.03	42.17	46.37	47.56	46.93	**48.79**



(a)

(b)

Fig. 8. Some samples from (a) EURECOM Kinect and (b) CAS-PEAL data sets.

the eight nonoccluded face images per person (i.e., the first four images in two sessions) for training and the occluded images (with sunglasses and scarf) in Sessions 1 and 2 for testing [30]. For YaleB, the basic settings are the same as in Section IV-C, and the test images are occluded by masks. Fig. 6(b) and (c) shows some face images with different disguises. In Fig. 6(c), the top two rows show face images with masks from the Internet, and the last two rows show some test images in our experiments, which simulates the real-world mask occlusion. Table IV lists the experiment results of different methods on the AR and YaleB data sets. We can clearly observe that LDMR is superior to the others in all cases. On AR, NMR and IRGSC also achieve competitive identification rates. Since the occlusion region in the test images with sunglasses is relatively small, SRC and CRC can achieve good results. For scarves occlusion, the occluded region gets larger and the performance gaps between LDMR and other methods become wide. On YaleB(mask), our LDMR also significantly outperforms other approaches.

Then, we perform our method on another two data sets with real-world occlusions: EURECOM Kinect and CAS-PEAL data sets. Fig. 8 shows some face images of the two data sets. For EURECOM, we use the 12 images without occlusions for training, and the rest six images with occlusions caused by sunglasses, hand, and paper for testing [30]. For CAS-PEAL, there is only one neutral face image per person for training and the images with accessories for testing. The identification rates of different methods on EURECOM Kinect and CAS-PEAL data sets are shown in Fig. 9. On EURECOM, LDMR is superior to other regression methods and achieves encouraging 90.38% identification accuracy, which is 3.52% and 4.16% higher than RCRC and NMR, respectively. On CAS-PEAL, LDMR still obtains the best performance among all regression methods. The experiment results on AR, YaleB(mask), EURECOM, and CAS-PEAL data sets demonstrate that LDMR is capable of recognizing faces occluded by various objects in real world.

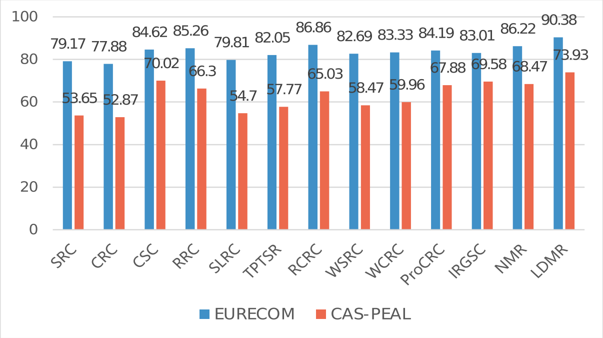


Fig. 9. Identification rates (%) of different methods on the EURECOM Kinect and CAS-PEAL data sets.

E. FI With Uncontrolled Setting

The face images tested in previous experiments are all captured in strictly controlled environment. In this experiment, we evaluate our method on three uncontrolled face data sets: LFW, PubFig, and DMF data sets. Fig. 10 shows some face images of the three data sets. Table V lists the identification rates of SRC, CRC, CSC, RRC, SLRC, TPTSR, RCRC, WSRC, WCRC, ProCRC, IRGSC, NMR, and LDMR on three data sets. For LFW, it is obvious that LDMR achieves the best performance: 46.58%. CSC, ProCRC, IRGSC, and NMR also obtain competitive results. The accuracy of LDMR is 1.77% higher than CSC, which ranks the second in all methods. For PubFig and DMF data sets, LDMR still outperforms other competing methods. It seems that our experimental results are unsatisfying since many deep learning approaches have achieved very high accuracies on these wild face data sets. However, these deep learning models use massive extra

Fig. 10. Some samples from (a) LFW, (b) PubFig, and (c) DMF data sets.

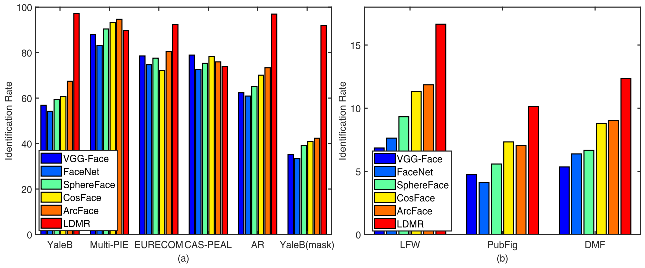


Fig. 11. Identification rates (%) of five CNN-based methods and LDMR on (a) YaleB, Multi-PIE, EURECOM, CAS-PEAL, and AR data sets. (b) LFW, PubFig, and DMF data sets with 30% random black block occlusion.

training data to achieve the excellent performance. On the other hand, our method is more robust than those deep learning approaches under contaminated conditions, which will be illustrated in Section IV-F.

F. Compared With CNN-Based Methods

Due to large-scale training data and high computational power, CNN-based methods have achieved great success in many computer vision and image analysis tasks in recent years [55], [56]. Following the directions in [57] and [58], we adopt the pretrained CNN models to extract features and classify them by NN with cosine distance metric. Five popular and publicly available deep learning models on face recognition are utilized here: VGG-Face [10], FaceNet [7], SphereFace [11], CosFace [8], and ArcFace [9], which are well-trained and evaluated on very large wild face databases. For YaleB, subset 4 is used as test set. For Multi-PIE, we randomly select five images per person for training. For AR, the occluded images with sunglasses of Session 1 are used for testing. The basic experimental settings of YaleB, Multi-PIE, AR, CAS-PEAL, EURECOM, and YaleB(mask) are the same as in Sections IV-B and IV-D. We compare LDMR with deep learning models on these data sets to investigate their robustness under contaminated conditions, and the experimental results are shown in Fig. 11(a).

TABLE VI

IDENTIFICATION RATES (%) COMPARISON OF LDMR AND ITS VARIATIONS ON THE YALEB AND AR DATA SETS

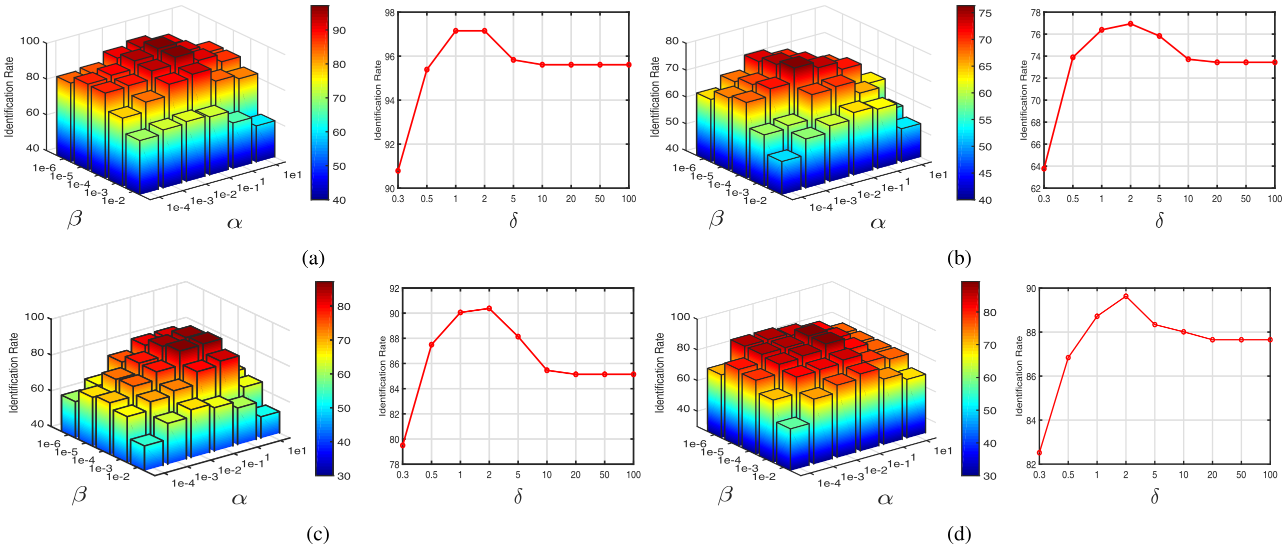
Method	YaleB		AR-S1	
	subset 4	subset 5	sunglasses	scarves
LDMR-o	95.61	53.92	94.72	73.61
LDMR-s	87.94	42.02	85.83	67.22
LDMR-r	94.52	51.54	93.06	71.94
LDMR	**97.33**	**58.26**	**96.94**	**76.39**

We can observe that LDMR outperforms CNN-based methods on the YaleB, EURECOM, and AR data sets. The main reason is that there are significant differences (i.e., severe shadows and occlusions) between training images and test images in these data sets, and the CNN models fail to work well on test sets with complex distortions unobserved in training sets [57]. The noises in the test sets of Multi-PIE and CAS-PEAL data sets are relatively small, and the CNN-based methods can achieve similar or slightly better performance than LDMR. On the YaleB(mask) data set, about 50% facial areas are occluded and deep neural networks perform poor, whereas our LDMR can obtain over 90% identification accuracy. In addition to those controlled data sets, we also conduct experiments on the LFW, PubFig, and DMF data sets. We impose 30% random black block on the test sets of these data sets, and the experimental results are shown in Fig. 11(b). We can see that CNN-based methods perform poor and are sensitive to the occlusions in test set, whereas LDMR shows robustness and obtain better performance than deep learning approaches. The possible reason is that deep learning approaches highly depend on the training data and cannot generalize well to other distortions that are scare in training sets. Compared with CNN-based methods, LDMR performs better in dealing with the new complex noises in test sets. Besides, our proposed method requires much less training data and time, and computational power.

G. Ablation Study and Parameter Analysis

As stated before, LDMR incorporates locality structure and class competitions into consideration simultaneously. To verify the effectiveness of them separately, we conduct ablation experiments. LDMR is compared with its three variations, i.e., LDMR-o, LMDR-s, and LDMR-r. LDMR-o discards the class weights of LDMR, i.e., $w_i = 1$. LDMR-s and LDMR-r discard the second term and third term in (14), respectively. The results of LDMR and its variations on the YaleB and AR data sets are shown in Table VI. From the results, we can observe that: 1) LDMR-o and LDMR-r outperform LDMR-s, which indicates that the representation components constraint can significantly improve the identification performance and 2) LDMR outperforms LDMR-o and LDMR-r, which verifies the positive effects of locality structure and class competitions.

From algorithm 1, there are three important tunable parameters α, β, and δ in LDMR. α and β make a balance between regression loss and regularization, and δ controls the strength of class weights. To analyze the parameter sensitivity, we first define candidate sets $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$, $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$, and $\{0.3, 0.5, 1, 2, 5, 10, 20, 50, 100\}$ for α, β, and δ, respectively. Then, LDMR is performed on the YaleB, AR, EURECOM, and Multi-PIE data sets with different combinations of the three parameters [47]. For YaleB and AR, subset 4 of YaleB and AR-S2 with scarves occlusion is used as test sets. The training size per subject is 5 for Multi-PIE data set. The experimental settings on the four data sets are the same as in Sections IV-B and IV-D.



(a)

(b)

(c)

(d)

Fig. 12. Identification rates (%) of LDMR versus α, β, and δ on (a) YaleB, (b) AR, (c) EURECOM, and (d) Multi-PIE data sets.

Fig. 12 shows the identification rates of LDMR versus α, β, and δ on different data sets. All three parameters impact the performance of algorithm. Generally, when α and β are selected from $[10^{-2}, 1]$ and $[10^{-5}, 10^{-3}]$, respectively, our method can achieve relative stable and satisfied identification performance. It seems that parameter β is less sensitive than α, and this is also verified in Table VI in which LDMR-s performs worse than LDMR-r. When δ is small (e.g., 0.3), the performance is undesirable due to the huge impact of prior weights, which may be not accurate enough. When δ is large (e.g., 50), the impact of prior weights is very small and the performance also degrades, which is also shown in Table VI. A value in the range of $[0.5, 2]$ is proper for δ.

However, it is still an open problem for optimal parameter selection for different data sets. In our experiments, we use the simple grid search method to determine optimal parameters. From Fig. 12, the proposed method can generally achieve satisfactory performance when α locates in the range $[10^{-2}, 1]$. Thus, we can first set α as a fixed value like 0.1 and find the optimal β and δ by grid searching in their own candidate set $[10^{-5}, 10^{-3}]$ and $[0.5, 2]$, respectively. After obtaining the optimal combination of β and δ, we can fix them with optimal values and find the optimal α. Consequently, all the optimal parameters can be achieved.

H. Running Time

Computational cost is also an important issue apart from accuracy in evaluating the performance of an algorithm. In this section, we make a comparison on the average running time of LDMR with other algorithms. We conduct experiments on the YaleB data set with subset 4 as testing set, and the basic experimental settings are the same as in Section IV-B. The experiments are performed on a personal computer with Windows 10 system and Intel Core i7-8550 CPU, 1.80 GHz, and 8.00 GB RAM. The computational platform is MATLAB R2017b.

As mentioned before, the coefficients vector $\hat{\mathbf{x}}$ of Ridge regression is used for initialization in Algorithm 1, while many research works initialize it as $\mathbf{0}$ [30], [31], [40], [42]. We first validate the advantage of our initialization method on the efficiency of algorithm. Considering the computational consumption of SVD is determined by the size of image matrix, three levels of image size are designed and tested for comparison. Table VII lists the average running time, iteration steps, and identification accuracy of LDMR under different initializations and image sizes. We can clearly see that the number of iterations and running time are greatly reduced in our proposed method, which improves the computational efficiency. This advantage becomes evident while operating on images in large size, and the identification rates keep stable in two cases of initialization.

TABLE VII

COMPARISON OF RUNNING TIME (S), ITERATIONS, AND IDENTIFICATION ACCURACY (%) UNDER DIFFERENT INITIALIZATIONS AND IMAGE SIZES

Image Size		48 × 42	96 × 84	192 × 168
Running Time	$\mathbf{x} = \hat{\mathbf{x}}$	0.2920	2.1757	20.9989
	$\mathbf{x} = \mathbf{0}$	0.4034	3.2520	26.1824
Iterations	$\mathbf{x} = \hat{\mathbf{x}}$	31	84	193
	$\mathbf{x} = \mathbf{0}$	60	118	259
Identification Rate	$\mathbf{x} = \hat{\mathbf{x}}$	97.33	97.59	98.03
	$\mathbf{x} = \mathbf{0}$	97.33	97.33	97.81

We also test the running time of different methods on the YaleB data set with different numbers of training samples per subject. The number of training samples per subject varies from 5 to 30 with an interval of 5. The average running time (base-10 log of seconds) for identifying one test sample is shown in Fig. 13. CRC and its variations, such as WCRC and ProCRC, are three fastest algorithms since these methods have closed-form solutions for representation learning. TPTSR is also efficient, which can be viewed as performing CRC twice in practice. However, these methods are not robust when face images are contaminated severely. In SRC, no dimensionality reduction algorithm is used, so SRC and its variations, such as WSRC and SLRC methods, are time-consuming. IRGSC is

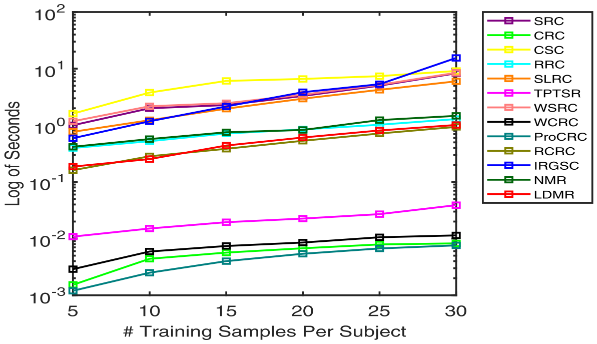


Fig. 13. Average running time (base-10 log of seconds) for identifying one test sample of different methods on the YaleB data set.

also not efficient due to its iterative sample weights and feature weights learning mechanism. Compared with other regression approaches, LDMR achieves desirable computational speed and is more robust to illumination changes and facial occlusions. Deep neural networks have achieved excellent performance on face recognition; however, they may be not robust to new distortions in test sets and need massive training data and time. For example, FaceNet spends 1000–2000 h on a CPU cluster for training [7]. Compared with deep learning methods, our method saves much training time.

V. Conclusion

In this work, we propose an LDMR method for robust FI under structural noise conditions. LDMR differs the roles of training samples at class level and directly constrains the representation components of all classes that have a closer connection to the identification process. The class weights characterized by subspace distances are integrated, and a weighted pairwise class competition constraint is designed to reduce the correlations among classes and enhance the competition among all classes. The experiments on several popular face data sets demonstrate the effectiveness and robustness of LDMR compared with other competitive regression methods. An interesting work is to incorporate dictionary learning into our proposed method, which may improve this performance for unconstrained FI. Furthermore, it is also important to reject imposters in practice, and we will work to extend our method for the open-set identification.

Acknowledgment

The authors would like to thank the editor and the anonymous reviewers for their constructive and valuable comments and suggestions.

References

[1] J. Lu, Y.-P. Tan, and G. Wang, “Discriminative multimanifold analysis for face recognition from a single training sample per person,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 39–51, Jan. 2013.

[2] Z. Lai, Y. Xu, Q. Chen, J. Yang, and D. Zhang, “Multilinear sparse principal component analysis,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 10, pp. 1942–1950, Oct. 2014.

[3] J. Wen *et al.*, “Robust sparse linear discriminant analysis,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 2, pp. 390–403, Feb. 2019.

[4] X. Jiang and J. Lai, “Sparse and dense hybrid representation via dictionary decomposition for face recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 5, pp. 1067–1079, May 2015.

[5] Y. Xu, Z. Zhong, J. Yang, J. You, and D. Zhang, “A new discriminative sparse representation method for robust face recognition via l_2 regularization,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2233–2242, Oct. 2017.

[6] M. Iliadis, H. Wang, R. Molina, and A. K. Katsaggelos, “Robust and low-rank representation for fast face identification with occlusions,” *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2203–2218, May 2017.

[7] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[8] H. Wang *et al.*, “CosFace: Large margin cosine loss for deep face recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5265–5274.

[9] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “ArcFace: Additive angular margin loss for deep face recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.

[10] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 41.1–41.12.

[11] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, “SphereFace: Deep hypersphere embedding for face recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 212–220.

[12] B. Cao, N. Wang, J. Li, and X. Gao, “Data augmentation-based joint learning for heterogeneous face recognition,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 6, pp. 1731–1743, Jun. 2019.

[13] D. Liu, X. Gao, N. Wang, J. Li, and C. Peng, “Coupled attribute learning for heterogeneous face recognition,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 11, pp. 4699–4712, Nov. 2020, doi: 10.1109/TNNLS.2019.2957285.

[14] Z. Ren, D. Dong, H. Li, and C. Chen, “Self-paced prioritized curriculum learning with coverage penalty in deep reinforcement learning,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2216–2226, Jun. 2018.

[15] B. Cao, N. Wang, X. Gao, J. Li, and Z. Li, “Multi-margin based decorrelation learning for heterogeneous face recognition,” in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 680–686.

[16] B. Cao, H. Zhang, N. Wang, X. Gao, and D. Shen, “Auto-GAN: Self-supervised collaborative learning for medical image synthesis,” in *Proc. AAAI*, 2020, pp. 10486–10493.

[17] Y. Xu, X. Zhu, Z. Li, G. Liu, Y. Lu, and H. Liu, “Using the original and ‘symmetrical face’ training samples to perform representation based two-step face recognition,” *Pattern Recognit.*, vol. 46, no. 4, pp. 1151–1158, Apr. 2013.

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

[19] C. Zhang, H. Li, Y. Qian, C. Chen, and Y. Gao, “Pairwise relations oriented discriminative regression,” *IEEE Trans. Circuits Syst. Video Technol.*, early access, Oct. 22, 2020, doi: 10.1109/TCSVT.2020.3032964.

[20] L. Zhang, M. Yang, and X. Feng, “Sparse representation or collaborative representation: Which helps face recognition?” in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 471–478.

[21] D. Liu, J. Li, N. Wang, C. Peng, and X. Gao, “Composite components-based face sketch recognition,” *Neurocomputing*, vol. 302, pp. 46–54, Aug. 2018.

[22] L. Liu, L. Chen, C. L. P. Chen, Y. Y. Tang, and C. M. Pun, “Weighted joint sparse representation for removing mixed noise in image,” *IEEE Trans. Cybern.*, vol. 47, no. 3, pp. 600–611, Mar. 2017.

[23] W. Hu, W. Li, X. Zhang, and S. Maybank, “Single and multiple object tracking using a multi-feature joint sparse representation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 816–833, Apr. 2015.

[24] X.-T. Yuan and Q. Liu, “Newton-type greedy selection methods for ℓ_0-constrained minimization,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2437–2450, Dec. 2017.

[25] G. Liu, Q. Liu, and P. Li, “Blessing of dimensionality: Recovering mixture data via dictionary pursuit,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 47–60, Jan. 2017.

[26] I. Naseem, R. Togneri, and M. Bennamoun, “Robust regression for face recognition,” *Pattern Recognit.*, vol. 45, no. 1, pp. 104–118, Jan. 2012.

[27] S. Cai, L. Zhang, W. Zuo, and X. Feng, “A probabilistic collaborative representation based approach for pattern classification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2950–2959.

[28] Y. Xu, D. Zhang, J. Yang, and J.-Y. Yang, “A two-phase test sample sparse representation method for use with face recognition,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 9, pp. 1255–1262, Sep. 2011.

[29] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Regularized robust coding for face recognition," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1753–1766, May 2013.

[30] J. Yang, L. Luo, J. Qian, Y. Tai, F. Zhang, and Y. Xu, "Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 156–171, Jan. 2017.

[31] J. Chen, J. Yang, L. Luo, J. Qian, and W. Xu, "Matrix variate distribution-induced sparse representation for robust image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 10, pp. 2291–2300, Oct. 2015.

[32] L. Luo, J. Yang, J. Qian, and Y. Tai, "Nuclear-L1 norm joint regression for face reconstruction and recognition with mixed noise," *Pattern Recognit.*, vol. 48, no. 12, pp. 3811–3824, Dec. 2015.

[33] H. Zhang, J. Yang, J. Qian, and W. Luo, "Nonconvex relaxation based matrix regression for face recognition with structural noise and mixed noise," *Neurocomputing*, vol. 269, pp. 188–198, Dec. 2017.

[34] C. Zhang, H. Li, C. Chen, Y. Qian, and X. Zhou, "Enhanced group sparse regularized nonconvex regression for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Oct. 27, 2020, doi: 10.1109/TPAMI.2020.3033994.

[35] S. Bengio, F. Pereira, Y. Singer, and D. Strelow, "Group sparse coding," in *Proc. NeurIPS*, 2009, pp. 82–89.

[36] X. Tang, G. Feng, and J. Cai, "Weighted group sparse representation for undersampled face recognition," *Neurocomputing*, vol. 145, pp. 402–415, Dec. 2014.

[37] J. Zheng, P. Yang, S. Chen, G. Shen, and W. Wang, "Iterative re-constrained group sparse face recognition with adaptive weights learning," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2408–2423, May 2017.

[38] J. Lai and X. Jiang, "Classwise sparse and collaborative patch representation for face recognition," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3261–3272, Jul. 2016.

[39] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3360–3367.

[40] Z. Fan, M. Ni, Q. Zhu, and E. Liu, "Weighted sparse representation for face recognition," *Neurocomputing*, vol. 151, pp. 304–309, Mar. 2015.

[41] R. Timofte and L. Van Gool, "Adaptive and weighted collaborative representations for image classification," *Pattern Recognit. Lett.*, vol. 43, pp. 127–135, Jul. 2014.

[42] Y. Chen and Z. Yi, "Locality-constrained least squares regression for subspace clustering," *Knowl.-Based Syst.*, vol. 163, pp. 51–56, Jan. 2019.

[43] Y. Qian, F. Li, J. Liang, B. Liu, and C. Dang, "Space structure and clustering of categorical data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 10, pp. 2047–2059, Oct. 2016.

[44] F. Li, Y. Qian, J. Wang, C. Dang, and L. Jing, "Clustering ensemble based on sample's stability," *Artif. Intell.*, vol. 273, pp. 37–55, Aug. 2019.

[45] C. Wang, Q. Hu, X. Wang, D. Chen, Y. Qian, and Z. Dong, "Feature selection based on neighborhood discrimination index," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 7, pp. 2986–2999, Jul. 2018.

[46] X. Peng, L. Zhang, Z. Yi, and K. K. Tan, "Learning locality-constrained collaborative representation for robust face recognition," *Pattern Recognit.*, vol. 47, no. 9, pp. 2794–2806, Sep. 2014.

[47] J. Wen, B. Zhang, Y. Xu, J. Yang, and N. Han, "Adaptive weighted nonnegative low-rank representation," *Pattern Recognit.*, vol. 81, pp. 326–340, Sep. 2018.

[48] J. Yang, L. Zhang, Y. Xu, and J.-Y. Yang, "Beyond sparsity: The role of L1-optimizer in pattern classification," *Pattern Recognit.*, vol. 45, no. 3, pp. 1104–1118, Mar. 2012.

[49] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.

[50] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, Jan. 2010.

[51] S. P. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2010.

[52] C. Chen, B. He, Y. Ye, and X. Yuan, "The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent," *Math. Program.*, vol. 155, nos. 1–2, pp. 57–79, Jan. 2016.

[53] Y. Gao, J. Ma, and A. L. Yuille, "Semi-supervised sparse representation based classification for face recognition with insufficient labeled samples," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2545–2560, May 2017.

[54] W. Deng, J. Hu, and J. Guo, "Face recognition via collaborative representation: Its discriminant nature and superposed representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 10, pp. 2513–2521, Oct. 2018.

[55] B. Cao, N. Wang, X. Gao, and J. Li, "Asymmetric joint learning for heterogeneous face recognition," in *Proc. AAAI*, 2018, pp. 6682–6689.

[56] C. Peng, X. Gao, N. Wang, and J. Li, "Graphical representation for heterogeneous face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 301–312, Feb. 2017.

[57] M. M. Ghazi and H. K. Ekenel, "A comprehensive analysis of deep learning based representation for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 102–109.

[58] C. Y. Wu and J. J. Ding, "Occluded face recognition using low-rank regression with generalized gradient direction," *Pattern Recognit.*, vol. 80, pp. 256–268, Aug. 2018.

Chao Zhang (Graduate Student Member, IEEE) received the B.E. degree in automation from Nanjing University, Nanjing, China, in 2018, where he is currently pursuing the M.E. degree with the Department of Control and Systems Engineering.

He is also working as a Researcher at the Research Center for Novel Technology of Intelligent Equipment, Nanjing University. His current research interests include machine learning, pattern recognition, and computer vision.

Huaxiong Li (Member, IEEE) received the M.E. degree in control theory and control engineering from Southeast University, Nanjing, China, in 2006, and the Ph.D. degree from Nanjing University, Nanjing, in 2009.

He was a Visiting Scholar with the Department of Computer Science, University of Regina, Regina, SK, Canada, from 2007 to 2008, and The University of Hong Kong, Hong Kong, in 2010. He is currently an Associate Professor with the Department of Control and Systems Engineering, Nanjing University. His current research interests include machine learning, pattern recognition, and computer vision.

Dr. Li is a Committee Member of the China Association of Artificial Intelligence (CAAI) Granular Computing and Knowledge Discovery (GCKD) Committee and the Machine Learning Committee and a Committee Member of the JiangSu Association of Artificial Intelligence (JSAI) Pattern Recognition Committee.

Yuhua Qian (Member, IEEE) received the M.S. and Ph.D. degrees in computers with applications from Shanxi University, Taiyuan, China, in 2005 and 2011, respectively.

He is currently a Professor with the Key Laboratory of Computational Intelligence and Chinese Information Processing, Ministry of Education, Shanxi University. He is best known for multigranulation rough sets in learning from categorical data and granular computing. He is involved in research on pattern recognition, feature selection, rough set theory, granular computing, and artificial intelligence. He has authored over 80 articles on these topics in international journals.

Dr. Qian served on the Editorial Board of the *International Journal of Knowledge-Based Organizations* and *Artificial Intelligence Research*. He has served as the Program Chair or the Special Issue Chair for the Conference on Rough Sets and Knowledge Technology, the Joint Rough Set Symposium, and the Conference on Industrial Instrumentation and Control, and a PC member for many machine learning, data mining, and granular computing conferences.

Chunlin Chen (Member, IEEE) received the B.E. degree in automatic control and the Ph.D. degree in control science and engineering from the University of Science and Technology of China (USTC), Hefei, China, in 2001 and 2006, respectively.

He was a Visiting Scholar with Princeton University, Princeton, NJ, USA, from 2012 to 2013. He had visiting positions at the University of New South Wales Canberra at ADFA, Campbell, ACT, Australia, and the City University of Hong Kong, Hong Kong. He is currently a Professor and the Chair of the Department of Control and Systems Engineering, Nanjing University, Nanjing, China. His recent research interests include machine learning, pattern recognition, intelligent information processing, and quantum control.

Dr. Chen is also a Committee Member of the JiangSu Association of Artificial Intelligence (JSAI) Pattern Recognition Committee and a Committee Member of the China Association of Artificial Intelligence (CAAI) Machine Learning Committee. He is also the Co-Chair of the Technical Committee on Quantum Cybernetics, IEEE Systems, Man and Cybernetics Society.

Xianzhong Zhou (Member, IEEE) received the B.S. and M.S. degrees in system engineering and the Ph.D. degree in control theory and application from the Nanjing University of Science and Technology, Nanjing, China, in 1982, 1985, and 1996, respectively.

He is currently a Professor with the Department of Control and Systems Engineering and the Director of the Research Center for Novel Technology of Intelligent Equipment, Nanjing University, Nanjing. His current research interests include eye view vision systems, intelligent information processing, and future integrated automation systems.

Dr. Zhou was among the people selected for 333 Engineering of Jiangsu Province, China, in 2002 and 2007, and the Excellent Science and Technology Worker of Jiangsu Province, China, in 2000. He is also the Executive Director of the Systems Engineering Society of China and the Honor President of the Systems Engineering Society of Jiangsu Province, China.