Discriminative Block-Diagonal Representation Learning for Image Recognition

Zheng Zhang, Yong Xu, *Senior Member, IEEE*, Ling Shao, *Senior Member, IEEE*, Jian Yang, *Member, IEEE*

***Abstract*—Existing block-diagonal representation researches mainly focuses on casting block-diagonal regularization on training data, while only little attention is dedicated to concurrently learning both block-diagonal representations of training and test data. In this paper, we propose a discriminative block-diagonal low-rank representation (BDLRR) method for recognition. In particular, the elaborate BDLRR is formulated as a joint optimization problem of shrinking the unfavorable representation from off-block-diagonal elements and strengthening the compact block-diagonal representation under the semi-supervised framework of low-rank representation. To this end, we first impose penalty constraints on the negative representation to eliminate the correlation between different classes such that the incoherence criterion of the extra-class representation is boosted. Moreover, a constructed subspace model is developed to enhance the self-expressive power of training samples and further build the representation bridge between the training and test samples, such that the coherence of the learned intra-class representation is consistently heightened. Finally, the resulting optimization problem is solved elegantly by employing an alternative optimization strategy, and a simple recognition algorithm on the learned representation is utilized for final prediction. Extensive experimental results demonstrate that the proposed method achieves superb recognition results on four face image datasets, three character datasets, and the fifteen scene multi-categories dataset. It not only shows superior potential on image recognition but also outperforms state-of-the-art methods.**

***Index Terms*—Discriminative representation, low-rank representation, sparse representation, block-diagonal structure, image recognition.**

I. Introduction

DISCRIMINATIVE and effective data representations play an indispensable role in computer vision and machine learning, because they tremendously influence the performance of various learning systems [1]. A favorable data representation can greatly uncover the underlying information of observed data and intensely facilitate the machine learning methods [2]. As a typical data representation method, sparse representation has earned its high reputation in both theoretical research and practical applications [2]–[4]. Recently, low-rank representation (LRR) has captured considerable attention [5]–[7], and has also been proved to be a powerful solution to a wide range of applications, especially in subspace segmentation [6], feature extraction [7] and image classification [8]–[10]. In this paper, we focus on learning an appropriate data representation by constructing a block-diagonal low-rank representation for image recognition.

Sparse representation has been widely studied and applied in signal processing, machine learning and computer vision [3], [11]. The key idea of sparse representation is based on the assumption that each signal can be approximately represented by a linear combination of a few atoms of an over-completed dictionary. With the successful application of sparse representation based classification (SRC) [3] in face recognition, numerous SRC based modifications have been proposed. For example, Nie et al. introduced an efficient and robust feature selection method by imposing the l_{21}-norm constraint on both loss function and regularization terms [13]. Xu et al. [14] proposed the semi-supervised sparse representation by employing a coarse-to-fine strategy, and Lu et al. [15] developed a weighted sparse representation based classifier by leveraging both data locality and linearity to sparse coding. Based on the basic theorem [4] that locality can always lead to sparsity but not necessarily vice versa, the locality-constrained linear coding (LLC) [4] method achieves the sparse target by enforcing the locality embedding of codebook. In addition, some researchers argue that sparsity is not the ultimate reason of achieving decent recognition results [12], [16]–[18]. For example, Zhang et al. [16] presented a collaborative representation based classification (CRC) method by employing the l_2-norm regularization rather than l_1-norm regularization for face recognition. It is demonstrated that CRC can achieve comparable performance but more efficient than SRC [16]. The linear regression based classification (LRC) [18] is another well-known representation based method. More specifically, LRC exploits each class of training samples to represent the test sample, and classifies it to the class leading to the minimum representation residual. A recent survey [2] comprehensively reviews most representative sparse representation based algorithms, and empirically summarizes its wide applications from both theoretical and practical perspectives.

Recently, low-rank representation has gained increasing interest from different research fields. It is noted that the sparsity constraint can only dominate the local structure of each data vector, whereas the low-rank constraint can directly control the global structure of data [19]. Furthermore, low-rank representation can greatly capture the underlying correlation behind the observed data [19], [20]. The most representative

Manuscript received May 25, 2016; revised *** **, 2016 and *** **, 2017; accepted June 4th, 2017. This work was partially supported by the National Natural Science Foundation of China under Grant 61233011, and Guangdong Province high-level personnel of special support program (No. 2016TX03X164). (*Corresponding author: Yong Xu.*)

Z. Zhang and Y. Xu are with the Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, P. R. China (e-mail:darrenzz219@gmail.com; yongxu@ymail.com).

L. Shao is with the School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, U.K. (e-mail:E-mail: ling.shao@ieee.org.)

Jian Yang is with the College of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, P. R. China (e-mail: csjyang@njust.edu.cn).

low-rank method, robust principal component analysis (RPCA) [21], was proposed to recover the clean data with the low-rank constraint from corrupted observations. In particular, RPCA first assumes that the observations lie in a single subspace such that they can be decomposed into two separate components, i.e. the low-rank and sparse noise parts. However, RPCA cannot handle the situation where corrupted or outlying data are drawn from a union of multiple subspaces. To this end, Liu et al. [6] proposed to perform matrix recovery by exploiting low-rank representation for subspace segmentation. The latent low-rank representation (LatLRR) [7] was then developed for joint subspace segmentation and feature extraction by discovering the hidden information underlying the observations. Moreover, lots of low-rank representation based dictionary learning methods were proposed for robust image classification. For example, Zhang et al. [10] constructed a structured low-rank representation (SLRR) by regularizing all training images of the same class to have the same representation code. However, the ideal structure in SLRR is questionable, because, though data from the same class usually lie in the same subspace, it does not mean that images belonging to the same class should have the same data representation. Wei et al. [22] developed a low-rank matrix approximation method by learning sub-dictionaries independently for each class and meanwhile enforcing the structural incoherence between different classes. Li et al. [23] explored a class-wise block-diagonal structure (CBDS) dictionary learning method, which learned discriminative low-rank representation by imposing the class-wise structure constraint. In addition, some variations of low-rank representation based methods were proposed to solve different problems. For example, Li et al. [24] designed a cross-view low-rank analysis framework to address the multi-view outlier detection problem. Zhuang et al. [25] presented a nonnegative low-rank sparse graph construction method for semi-supervised learning. As a result, it is widely agreed that the low-rank criterion indeed can disclose the potential data structures of different classes or tasks' correlation patterns, such that the effectiveness of the learned data representation can be enhanced.

There is a well-attested fact that the sparse representation in [3] is a discriminative representation, whereas it only considers the data representation of each input signal independently, which does not take advantage of the global structural information in the set. In addition, existing research [22] [23], [24] has demonstrated that imposing specific structures on the low-rank representation matrix is beneficial to improve the discriminative capability of data representation. However, the performance of these methods is still far from being satisfactory. The main reason may be that these methods cannot perfectly transfer the original data features to the discriminative feature representations. Based on the well-explored self-expression property [26], the ideal block-diagonal representation can capture the underlying data information of samples by embedding the global semantic structure information and discriminative identification capability [10]. Consequently, promising results can be achieved if the discriminative data representation with the block-diagonal structure is exploited for recognition. In this paper, a novel block-diagonal low-rank representation (BDLRR) method is proposed to learn discriminative data representations which can simultaneously shrink the off-block-diagonal components and highlight the block-diagonal representation under the framework of low-rank representation. More specifically, BDLRR first eliminates the negative representation and boosts the incoherence of the extra-class representation by minimizing the off-block-diagonal representation, such that it can remove the noisy representation and transfer the positive representation to the block-diagonal components. Furthermore, BDLRR constructs a subspace model to enhance the self-expressive power of training samples and simultaneously bridge the representation gap between the training and test samples in a semi-supervised manner, such that the coherence of the intra-class representation is further improved and the learned representations are consistent. Finally, we introduce an effective iterative algorithm to solve the resulting optimization problem, and our method is evaluated to verify its adaptive capabilities for different recognition tasks. In summary, our key contributions are summarized as follows:

(1) A discriminative block-diagonal data representation structure is designed to boost the incoherent power of the extra-class representation by jointly removing the negative representation from the off-block-diagonal components and conveying the positive representation to the block-diagonal structure, such that better discriminative data representations are obtained for recognition tasks.

(2) A constructed subspace structure is developed to enhance the coherence of the intra-class representation by simultaneously improving the self-expressive capabilities of training samples and further narrowing the representation gap between training and test samples. Moreover, a low-rank criterion is enforced to capture the underlying feature structures of different classes or tasks' correlation patterns such that more competent representation results are achieved.

(3) By virtue of the semi-supervised learning superiority, the well-defined representation learning framework simultaneously learns both of discriminative training and test representations to keep consistence of the learned representations for recognition. To accommodate our method for large-scale problems, the out-of-sample extension is further explored to deal with new data instances.

(4) An effective optimization strategy based on the alternating direction method of multipliers (ADMM) is developed to solve the resulting optimization problem, and the convergence analysis of the designed optimization problem is presented from both theoretical and experimental perspectives.

The rest of this paper is organized as follows. We briefly review the related work on the low-rank theory in Section II. Then, we elaborate the proposed BDLRR method in Section III, and the solution to the optimization problem of the proposed BDLRR method is presented in Section IV. Section V reports extensive experimental results, as well as convergence and parameter sensitiveness analysis. Finally, the conclusion remarks are given in Section VI.

II. Related Work

In this section, we give a brief review of two typical low-rank criterion based methods, i.e. robust principal component analysis (RPCA) [21] and low-rank representation (LRR) [6].

Let us first introduce our notations used in this paper. Matrices are represented with bold uppercase letters, e.g. $\boldsymbol{X}$, and column vectors are denoted by bold lower letters, e.g. $\boldsymbol{x}$. In particular, $\boldsymbol{I}$ denotes an identity matrix, and the entries of a matrix or vector are denoted by using $[\cdot]$ with subscripts. The i-th row and j-th column element of matrix $\boldsymbol{X}$ is denoted as $\boldsymbol{x}_{ij}$, and the block-diagonal matrix composed of a collection of matrices $[\boldsymbol{X}_1, \ldots, \boldsymbol{X}_C]$ is denoted by

$$diag(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_C) = \begin{bmatrix} \boldsymbol{X}_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \boldsymbol{X}_C \end{bmatrix}.$$

A matrix's l_0, l_1 and l_{21} norms are denoted as $\|\boldsymbol{X}\|_0 = \sharp\{(i,j) : \boldsymbol{x}_{ij} \neq 0\}$, $\|\boldsymbol{X}\|_1 = \sum_{ij} |\boldsymbol{x}_{ij}|$, and $\|\boldsymbol{X}\|_{21} = \sum_j ||\boldsymbol{X}_{\cdot j}||$, respectively. The norm induced by the l_∞-norm on the matrix is denoted as $||\boldsymbol{X}||_\infty = \max_i \sum_j |\boldsymbol{x}_{ij}|$. The matrix Frobenius norm designates $\|\boldsymbol{X}\|_F^2 = tr(\boldsymbol{X}^T\boldsymbol{X}) = tr(\boldsymbol{X}\boldsymbol{X}^T) = \sum_{ij} \boldsymbol{x}_{ij}^2$, where $tr(\bullet)$ is the trace operator. $\|\boldsymbol{X}\|_*$ is the trace or nuclear norm of matrix $\boldsymbol{X}$, i.e. $\|\boldsymbol{X}\|_* = \sum_i |\sigma_i|$, where σ_i is the i-th singular value of matrix $\boldsymbol{X}$. $\boldsymbol{X}^T$ denotes the transposed matrix of $\boldsymbol{X}$. $\boldsymbol{0}_{mn}$ denotes an all-zero matrix with the size of $m \times n$, and the all-one vector $\boldsymbol{1}_N = [\underbrace{1, \cdots, 1}_{N}]^T$.

A. Robust principal component analysis (RPCA)

Suppose that $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n] \in \Re^{d \times n}$ is the observed data matrix and composed of n samples, where each column is a sample vector and usually has been highly corrupted. The main objective of RPCA is to determine a low-rank matrix $\boldsymbol{X}_0$ from the corrupted observations $\boldsymbol{X}$, and meanwhile filter out the sparse noise components $\boldsymbol{E}$, i.e. $\boldsymbol{X} = \boldsymbol{X}_0 + \boldsymbol{E}$. Consequently, the objective function of RPCA can be easily formulated as

$$\min_{\boldsymbol{X}_0, \boldsymbol{E}} \ rank(\boldsymbol{X}_0) + \lambda ||\boldsymbol{E}||_0 \ \ s.t. \ \ \boldsymbol{X} = \boldsymbol{X}_0 + \boldsymbol{E}, \quad (1)$$

where the $rank(\bullet)$ operator denotes the rank of matrix $\boldsymbol{X}_0$, λ is the balance parameter, and $\|\cdot\|_0$ means the l_0 pseudo-norm. Given an appropriate value of λ, RPCA can recover the clean data by $\boldsymbol{X}_0$. Due to the discrete properties of the rank function and the l_0-norm minimization, both of them are NP-hard problems and even difficult to approximate. An advisable choice [21] is to replace the rank constraint and l_0-norm regularization by the nuclear norm and l_1-norm regularization, respectively. As a result, problem (1) can be reformulated as

$$\min_{\boldsymbol{X}_0, \boldsymbol{E}} \ ||\boldsymbol{X}_0||_* + \lambda ||\boldsymbol{E}||_1 \ \ s.t. \ \ \boldsymbol{X} = \boldsymbol{X}_0 + \boldsymbol{E}, \quad (2)$$

where $||\cdot||_*$ and $||\cdot||_1$ are the nuclear norm and l_1-norm, respectively. It is known that problem (2) can be efficiently solved by Augmented Lagrange Multiplier (ALM) method [27].

B. Low-rank representation (LRR) based method

It is noted that RPCA is essentially based on the priori hypothesis that the observed data is approximately drawn from a low-rank subspace, that is, data can be described by a single subspace [6]. However, this assumption is very difficult to be satisfied for real-world datasets, where multiple subspaces are more reasonable. To this end, LRR [6] assumes that each data can be approximately represented by a union of several linear low-rank subspaces. The objective function of LRR is formulated as

$$\min_{\boldsymbol{Z}, \boldsymbol{E}} \ rank(\boldsymbol{Z}) + \lambda ||\boldsymbol{E}||_l \ \ s.t. \ \ \boldsymbol{X} = \boldsymbol{D}\boldsymbol{Z} + \boldsymbol{E}, \quad (3)$$

where $\boldsymbol{D}$ and λ are the dictionary and balance parameter, respectively. $||\cdot||_l$ indicates the constraint of different norms, and imposing different norms tends to remove specific noise as illustrated in [6]. For example, the matrix Frobenius norm can effectively capture Gaussian noise, while the l_1-norm can better process the random noise or corruptions. Similar to RPCA, problem (3) can be approximately reformulated as

$$\min_{\boldsymbol{Z}, \boldsymbol{E}} \ ||\boldsymbol{Z}||_* + \lambda ||\boldsymbol{E}||_l \ \ s.t. \ \ \boldsymbol{X} = \boldsymbol{D}\boldsymbol{Z} + \boldsymbol{E}, \quad (4)$$

which can also be effectively solved by the ALM algorithm [6], [27].

III. The proposed Block-Diagonal Low-Rank Representation

In this section, we introduce a novel block-diagonal low-rank representation (BDLRR) method, which collaboratively learns appropriate block-diagonal representations of training and test samples by jointly enforcing the incoherence of extra-class data representations and enhancing the coherence of intra-class data representations.

Assumption 1: *Let $\boldsymbol{X} = [\boldsymbol{X}_1, \cdots, \boldsymbol{X}_C] \in \Re^{d \times n}$ denote N training samples with a dimension of d from C classes, where each column is a sample vector. Suppose that all the samples are rearranged based on the class labels, and each class of training samples are stacked together to form a sub-matrix $\boldsymbol{X}_i \in \Re^{d \times n_i}$, which denotes n_i samples from the i-th class (j=1,2, .., C).*

Definition 1 (self-expressiveness property) [26]: Each data instance from a union of multiple subspaces can be effectively represented by a linear combination of other data instances, which is referred as the *self-expressiveness property*.

Definition 2: Suppose that a data point $\boldsymbol{y} \in \Re^d$ is from the i-th class. $\boldsymbol{z}$ is a solution of the linear equation $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{z}$, where the sub- vectors $\boldsymbol{z}_j$ (j=1,2, .., C) of $\boldsymbol{z}$ respectively corresponding to the j-th class. Based on the self-expressiveness property, the sub-matrix $\boldsymbol{X}_i$ should be able to well represent $\boldsymbol{y}$, and there is $\boldsymbol{y} \approx \boldsymbol{X}_i\boldsymbol{z}_i$. We define $\boldsymbol{z}_i$ as the *intra-class representation*, otherwise the coding coefficients $\boldsymbol{z}_j (j \neq i)$ are called the *extra-class representation*.

It is worth noting that the self-expressiveness property has already been successfully utilized in the context of classification [3], [22] and low-rank matrix approximation [6] and clustering [26]. Typically, SSC and LRR are the most representative methods, and the explicit self-expressiveness formulation, $\boldsymbol{X} = \boldsymbol{X}\boldsymbol{Z}$, is easily satisfied, where $\boldsymbol{Z}$ is

data representation. Furthermore, in the presence of the self-expressiveness property, the key underlying observation of SSC and LRR is disclosed that each data point in a dataset can be ideally represented by a linear combination of a few points from its own subspace. Based on this observation and Assumption 1, the desired self-expressive representation should be block-diagonal and the obtained data representation is sufficiently discriminative. So, the ideal block-diagonal structure based representation is

$$\boldsymbol{X} = \boldsymbol{X}\hat{\boldsymbol{Z}} \;\; s.t. \;\; \hat{\boldsymbol{Z}} = diag(\boldsymbol{Z}), \tag{5}$$

where $\boldsymbol{Z} = [\boldsymbol{Z}_{11}, \cdots, \boldsymbol{Z}_{CC}]$, and $\boldsymbol{Z}_{ij}$ is the representation coefficient of $\boldsymbol{X}_i$ corresponding to $\boldsymbol{X}_j$. However, the absolute block-diagonal structure is not easy to learn. To this end, it is natural to assume that the off-block-diagonal components are as small as possible to enhance the incoherent extra-class representation, which means that $\boldsymbol{Z}_{ij}$ tends to a zero sub-matrix for $i \neq j$. In addition, the coherent intra-class representation at the same time is further boosted. We formulate the following structured representation as

$$\min_{\boldsymbol{Z}} \lambda_1 ||\boldsymbol{A} \odot \boldsymbol{Z}||_F^2 + \lambda_2 ||\boldsymbol{D} \odot \boldsymbol{Z}||_0 \;\; s.t. \;\; \boldsymbol{X} = \boldsymbol{X}\boldsymbol{Z}, \tag{6}$$

where λ_1 and λ_2 are positive constants to weigh corresponding terms, $\odot$ indicates the Hadamard product (i.e. element-wise product), and $\boldsymbol{X} \in \Re^{d \times n}$. More specifically, the first term attempts to minimize the off-block-diagonal entries, and $\boldsymbol{A} = \mathbf{1}_n \mathbf{1}_n^T - \boldsymbol{Y}$ where $\boldsymbol{Y} = \begin{bmatrix} \mathbf{1}_{n_1}\mathbf{1}_{n_1}^T & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{1}_{n_C}\mathbf{1}_{n_C}^T \end{bmatrix}$. The second term is the constructed subspace measure to improve the coherent representation of intra-class representation. $\boldsymbol{d_{ij}}$ is a distance metric between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ such that similar samples have high probabilities to be similar data representations. There are many distance metric methods. In this work, we simply define the distance between two samples as the square of the Euclidean distance, i.e. $||\boldsymbol{x}_i - \boldsymbol{x}_j||_2^2$. Because solving the l_0-norm minimization problem is an NP-hard problem, a relaxed counterpart of the second term is formulated as $||\boldsymbol{D} \odot \boldsymbol{Z}||_1$. Thus, problem (6) can be reformulated as

$$\min_{\boldsymbol{Z}} \lambda_1 ||\boldsymbol{A} \odot \boldsymbol{Z}||_F^2 + \lambda_2 ||\boldsymbol{D} \odot \boldsymbol{Z}||_1 \;\; s.t. \;\; \boldsymbol{X} = \boldsymbol{X}\boldsymbol{Z}. \tag{7}$$

In general, a low-rank criterion can further capture the underlying classes' correlation patterns such that the performance of resulting models can be improved [6], [10], [20]. By integrating problems (7) and (4), we propose the following objective function for the semi-supervised BDLRR:

$$\begin{aligned} \min_{\boldsymbol{Z},\boldsymbol{E}} ||\boldsymbol{Z}||_* + \lambda_1 ||\tilde{\boldsymbol{A}} \odot \boldsymbol{Z}||_F^2 + \lambda_2 ||\boldsymbol{D} \odot \boldsymbol{Z}||_1 \\ +\lambda_3 ||\boldsymbol{E}||_{21} \;\; s.t. \;\; \boldsymbol{X} = \boldsymbol{X}_{tr}\boldsymbol{Z} + \boldsymbol{E}, \end{aligned} \tag{8}$$

where λ_1, λ_2 and λ_3 are positive scalars that weigh the corresponding terms in (8). $\boldsymbol{X}_{tr} \in \Re^{d \times n}$ is the training data matrix and $\boldsymbol{X} \in \Re^{d \times N}$ includes both training and test samples, i.e. $\boldsymbol{X} = [\boldsymbol{X_{tr}}, \boldsymbol{X_{tt}}]$. For the second term, $\tilde{\boldsymbol{A}} = [\boldsymbol{A}, \mathbf{1}_n \mathbf{1}_{N-n}^T]$ where $\boldsymbol{A}$ is the same as (6), and data representation $\boldsymbol{Z} = [\boldsymbol{Z}_{tr}, \boldsymbol{Z}_{tt}]$ such that an implicit $||\boldsymbol{Z}_{tt}||_F^2$ term is imposed to avoid overfitting. For the third term, $\boldsymbol{D} \in \Re^{n \times N}$ is denoted as the distance between training samples $\boldsymbol{X}_{tr}$ and all samples $\boldsymbol{X}$ such that the coherent representation of both $\boldsymbol{X}_{tr}$ and $\boldsymbol{X}_{tt}$ corresponding to $\boldsymbol{X}_{tr}$ can be enhanced simultaneously. $\boldsymbol{E}$ denotes the noise term with the l_{21}-norm regularization to capture sample-specific noise information [6]. Moreover, data representation $\boldsymbol{Z}$ of training and test samples is incorporated into a unified optimization problem such that $\boldsymbol{Z}_{tr}$ and $\boldsymbol{Z}_{tt}$ are both optimal and discriminative.

IV. Optimization and algorithm analysis

To solve the optimization problem of BDLRR in (8), we propose to utilize an alternating direction method, and separate the problem into several subproblems, which have close-form solutions.

A. Optimization Algorithm

To solve optimization problem (8), we first make an equivalent transformation by introducing two auxiliary variables to make the problem separable, and then problem (8) can be rewritten as

$$\begin{aligned} \min_{\boldsymbol{P},\boldsymbol{Z},\boldsymbol{Q},\boldsymbol{E}} ||\boldsymbol{P}||_* + \frac{\lambda_1}{2}||\tilde{\boldsymbol{A}} \odot \boldsymbol{Z}||_F^2 + \lambda_2 ||\boldsymbol{D} \odot \boldsymbol{Q}||_1 + \lambda_3 ||\boldsymbol{E}||_{21} \\ s.t. \;\; \boldsymbol{X} = \boldsymbol{X}_{tr}\boldsymbol{Z} + \boldsymbol{E}, \;\; \boldsymbol{P} = \boldsymbol{Z}, \;\; \boldsymbol{Q} = \boldsymbol{Z}. \end{aligned} \tag{9}$$

Then, we can get the following objective function of the problem by the augmented Lagrangian multiplier method. Here the augmented Lagrangian function of problem (9) is

$$\begin{aligned} \mathcal{L}(\boldsymbol{P},\boldsymbol{Z},\boldsymbol{Q},\boldsymbol{E},\boldsymbol{C_1},\boldsymbol{C_2},\boldsymbol{C_3}) = ||\boldsymbol{P}||_* + \frac{\lambda_1}{2}||\tilde{\boldsymbol{A}} \odot \boldsymbol{Z}||_F^2 + \\ \lambda_2 ||\boldsymbol{D} \odot \boldsymbol{Q}||_1 + \lambda_3 ||\boldsymbol{E}||_{21} + \langle \boldsymbol{C}_1, \boldsymbol{X} - \boldsymbol{X}_{tr}\boldsymbol{Z} - \boldsymbol{E} \rangle \\ + \langle \boldsymbol{C}_2, \boldsymbol{P} - \boldsymbol{Z} \rangle + \langle \boldsymbol{C}_3, \boldsymbol{Q} - \boldsymbol{Z} \rangle + \frac{\mu}{2}(||\boldsymbol{P} - \boldsymbol{Z}||_F^2 + \\ ||\boldsymbol{X} - \boldsymbol{X}_{tr}\boldsymbol{Z} - \boldsymbol{E}||_F^2 + ||\boldsymbol{Q} - \boldsymbol{Z}||_F^2), \end{aligned} \tag{10}$$

where $\langle \boldsymbol{P}, \boldsymbol{Q} \rangle = tr(\boldsymbol{P}^T\boldsymbol{Q})$. $\boldsymbol{C}_1$, $\boldsymbol{C}_2$ and $\boldsymbol{C}_3$ are the Lagrangian multipliers, and $\mu > 0$ is a penalty parameter. The augmented Lagrangian is minimized along one coordinate direction at each iteration, i.e. minimizing the loss with respect to one variable with the remaining variables fixed. We introduce the detailed procedures as follows.

Updating $\boldsymbol{Z}$: Fix the other variables and update $\boldsymbol{Z}$ by solving the following problem

$$\begin{aligned} \mathcal{L} &= \min_{\boldsymbol{Z}} \frac{\lambda_1}{2}||\tilde{\boldsymbol{A}} \odot \boldsymbol{Z}||_F^2 + \langle \boldsymbol{C}_1^t, \boldsymbol{X} - \boldsymbol{X}_{tr}\boldsymbol{Z} - \boldsymbol{E}^t \rangle \\ &\quad + \langle \boldsymbol{C}_2^t, \boldsymbol{P}^{t+1} - \boldsymbol{Z} \rangle + \langle \boldsymbol{C}_3^t, \boldsymbol{Q}^t - \boldsymbol{Z} \rangle + \frac{\mu^t}{2}(||\boldsymbol{P}^{t+1} - \boldsymbol{Z}||_F^2 \\ &\quad + ||\boldsymbol{X} - \boldsymbol{X}_{tr}\boldsymbol{Z} - \boldsymbol{E}^t||_F^2 + ||\boldsymbol{Q}^t - \boldsymbol{Z}||_F^2) \\ &= \frac{\lambda_1}{2}||\tilde{\boldsymbol{A}} \odot \boldsymbol{Z}||_F^2 + \frac{\mu^t}{2}(||\boldsymbol{X} - \boldsymbol{X}_{tr}\boldsymbol{Z} - \boldsymbol{E}^t + \frac{\boldsymbol{C}_1^t}{\mu^t}||_F^2 \\ &\quad + ||\boldsymbol{P}^{t+1} - \boldsymbol{Z} + \frac{\boldsymbol{C}_2^t}{\mu^t}||_F^2 + ||\boldsymbol{Q}^t - \boldsymbol{Z} + \frac{\boldsymbol{C}_3^t}{\mu^t}||_F^2), \end{aligned} \tag{11}$$

Algorithm 1. Solving Problem (8) by ADMM

Require: All feature matrix $\boldsymbol{X} = [\boldsymbol{X}_{tr}, \boldsymbol{X}_{tt}]$; Parameters λ_1, λ_2, λ_3; Distance measure matrix $\boldsymbol{D}$.
Initialization: $\boldsymbol{P} = \boldsymbol{0}$, $\boldsymbol{Z} = \boldsymbol{0}$, $\boldsymbol{Q} = \boldsymbol{0}$, $\boldsymbol{E} = \boldsymbol{0}$, λ_1, λ_2, $\lambda_3 > 0$, $\boldsymbol{C}_1 = \boldsymbol{0}$, $\boldsymbol{C}_2 = \boldsymbol{0}$, $\boldsymbol{C}_3 = \boldsymbol{0}$, $\mu_{max} = 10^8$, $tol = 10^{-6}$, $\rho = 1.15$.
While not converged **do**
1). Update $\boldsymbol{Z}$ by using (13);
2). Update $\boldsymbol{P}$ by using (15);
3). Update $\boldsymbol{Q}$ by using (18);
4). Update $\boldsymbol{E}$ by using (21);
5). Update Lagrange multipliers $\boldsymbol{C_1}$, $\boldsymbol{C_2}$ and $\boldsymbol{C_3}$:
$$\begin{cases} \boldsymbol{C}_1^{t+1} = \boldsymbol{C}_1^t + \mu^t(\boldsymbol{X} - \boldsymbol{X}_{tr}\boldsymbol{Z}^{t+1} - \boldsymbol{E}^{t+1}) \\ \boldsymbol{C}_2^{t+1} = \boldsymbol{C}_2^t + \mu^t(\boldsymbol{P}^{t+1} - \boldsymbol{Z}^{t+1}) \\ \boldsymbol{C}_3^{t+1} = \boldsymbol{C}_3^t + \mu^t(\boldsymbol{Q}^{t+1} - \boldsymbol{Z}^{t+1}). \end{cases}$$
6). Update μ:
$\mu^{t+1} = \min(\mu_{max}, \rho\mu^t)$
7). Check convergence: if
$$\max\begin{pmatrix} ||\boldsymbol{X} - \boldsymbol{X}_{tr}\boldsymbol{Z}^{t+1} - \boldsymbol{E}^{t+1}||_\infty, \\ ||\boldsymbol{P}^{t+1} - \boldsymbol{Z}^{t+1}||_\infty, ||\boldsymbol{Q}^{t+1} - \boldsymbol{Z}^{t+1}||_\infty \end{pmatrix} \leq tol,$$
and then stop.
End While

which is equivalent to

$$\begin{aligned}\mathcal{L} = \min_{\boldsymbol{Z}} \frac{\lambda_1}{2}||\boldsymbol{Z} - \boldsymbol{R}||_F^2 + \frac{\mu^t}{2}(||\boldsymbol{X} - \boldsymbol{X}_{tr}\boldsymbol{Z} - \boldsymbol{E}^t + \frac{\boldsymbol{C}_1^t}{\mu^t}||_F^2 \\ + ||\boldsymbol{P}^{t+1} - \boldsymbol{Z} + \frac{\boldsymbol{C}_2^t}{\mu^t}||_F^2 + ||\boldsymbol{Q}^t - \boldsymbol{Z} + \frac{\boldsymbol{C}_3^t}{\mu^t}||_F^2),\end{aligned} \tag{12}$$

where $\boldsymbol{R} = [\boldsymbol{Y}, \boldsymbol{0}_{n(N-n)}] \odot \boldsymbol{Z}^t$. By setting the derivation $\frac{\partial \mathcal{L}}{\partial \boldsymbol{Z}} = \boldsymbol{0}$, we can easily infer the optimal solution of $\boldsymbol{Z}$, and the closed-form solution to problem (12) is given by the following form,

$$\boldsymbol{Z}^{t+1} = \left[(2 + \frac{\lambda_1}{\mu^t})\boldsymbol{I} + \boldsymbol{X}_{tr}^T\boldsymbol{X}_{tr}\right]^{-1}(\frac{\lambda_1}{\mu^t}\boldsymbol{R} + \boldsymbol{X}_{tr}^T\boldsymbol{S}_1 + \boldsymbol{S}_2 + \boldsymbol{S}_3), \tag{13}$$

where $\boldsymbol{S}_1 = \boldsymbol{X} - \boldsymbol{E}^t + \frac{\boldsymbol{C}_1^t}{\mu^t}$, $\boldsymbol{S}_2 = \boldsymbol{P}^{t+1} + \frac{\boldsymbol{C}_2^t}{\mu^t}$, and $\boldsymbol{S}_3 = \boldsymbol{Q}^t + \frac{\boldsymbol{C}_3^t}{\mu^t}$.

Updating $\boldsymbol{P}$: When fixing the other variables, the objective function of (10) is degenerated into a function with respect to $\boldsymbol{P}$, i.e.

$$\begin{aligned}\boldsymbol{P}^{t+1} &= \arg\min_{\boldsymbol{P}} ||\boldsymbol{P}||_* + \langle \boldsymbol{C}_2^t, \boldsymbol{P} - \boldsymbol{Z}^t\rangle + \frac{\mu^t}{2}||\boldsymbol{P} - \boldsymbol{Z}^t||_F^2 \\ &= ||\boldsymbol{P}||_* + \frac{\mu^t}{2}||\boldsymbol{P} - (\boldsymbol{Z}^t - \frac{\boldsymbol{C}_2^t}{\mu^t})||_F^2.\end{aligned} \tag{14}$$

This problem has a closed-form solution by using the singular value thresholding (SVT) operator [27] [28], i.e.

$$\boldsymbol{P}^{t+1} = \mathcal{T}_{\frac{1}{\mu^t}}(\boldsymbol{Z}^t - \frac{\boldsymbol{C}_2^t}{\mu^t}) = \boldsymbol{U}\mathcal{S}_{\frac{1}{\mu^t}}(\boldsymbol{\Sigma})\boldsymbol{V}^T, \tag{15}$$

where $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ is the singular value decomposition of $(\boldsymbol{Z}^t - \frac{\boldsymbol{C}_2^t}{\mu^t})$, and $\mathcal{S}_{\frac{1}{\mu^t}}(\cdot)$ is the soft-thresholding operator [2] [27], which is defined as

$$\mathcal{S}_\lambda(\boldsymbol{x}) = \begin{cases} \boldsymbol{x} - \lambda, & if \quad \boldsymbol{x} > \lambda \\ \boldsymbol{x} + \lambda, & if \quad \boldsymbol{x} < -\lambda \\ 0, & otherwise. \end{cases} \tag{16}$$

Updating $\boldsymbol{Q}$: When the other variables are fixed, the objective optimization problem (10) with respective to $\boldsymbol{Q}$ is degenerated to the following problem

Algorithm 2. BDLRR model for recognition

Input: Training feature set $\boldsymbol{X}_{tr}$ with label matrix $\boldsymbol{Y}$, test sample set $\boldsymbol{X}_{tt}$.
Output: Predicted label matrix $\boldsymbol{L}$ for test samples.
1). Normalize all the samples of both training and test samples to unit-norm by using $\boldsymbol{x}_i = \boldsymbol{x}_i/\|\boldsymbol{x}_i\|_2$.
2). Exploit **Algorithm 1** to solve problem (8), and a discriminative representation matrix $\boldsymbol{Z} = [\boldsymbol{Z}_{tr}, \boldsymbol{Z}_{tt}]$ is obtained.
3). Employ Eqn. (23) to learn an optimal linear classifier $\hat{\boldsymbol{W}}$.
4). Predict the label matrix $\boldsymbol{L}$ of test samples $\boldsymbol{X}_{tt}$ by utilizing Eqn. (24) one by one.

$$\begin{aligned}\mathcal{L} &= \min_{\boldsymbol{Q}} \lambda_2||\boldsymbol{D} \odot \boldsymbol{Q}||_1 + \langle \boldsymbol{C}_3^t, \boldsymbol{Q} - \boldsymbol{Z}^{t+1}\rangle \\ &\quad + \frac{\mu^t}{2}||\boldsymbol{Q} - \boldsymbol{Z}^{t+1}||_F^2 \\ &= \lambda_2||\boldsymbol{D} \odot \boldsymbol{Q}||_1 + \frac{\mu^t}{2}||\boldsymbol{Q} - (\boldsymbol{Z}^{t+1} - \frac{\boldsymbol{C}_3^t}{\mu^t})||_F^2,\end{aligned} \tag{17}$$

which can be updated by the element-wise strategy. Obviously, problem (17) can be equivalently decoupled into $n \times N$ subproblems. For the i-th row and j-th column element $\boldsymbol{Q}_{ij}$, the optimal solution of problem (17) is

$$\begin{aligned}\boldsymbol{Q}_{ij}^{t+1} &= \arg\min_{\boldsymbol{Q}_{ij}} \ \lambda_2\boldsymbol{D}_{ij}|\boldsymbol{Q}_{ij}| + \frac{\mu^t}{2}(\boldsymbol{Q}_{ij} - \boldsymbol{M}_{ij})^2 \\ &= \mathcal{S}_{\frac{\lambda_2\boldsymbol{D}_{ij}}{\mu^t}}(\boldsymbol{M}_{ij}),\end{aligned} \tag{18}$$

where $\boldsymbol{M}_{ij} = \boldsymbol{Z}_{ij}^{t+1} - \frac{(\boldsymbol{C}_3^t)_{ij}}{\mu^t}$.

Updating $\boldsymbol{E}$: When other variables are fixed, problem (10) can be converted into the following problem

$$\begin{aligned}\min_{\boldsymbol{E}} \lambda_3||\boldsymbol{E}||_{21} + \langle \boldsymbol{C}_1^t, \boldsymbol{X} - \boldsymbol{X}_{tr}\boldsymbol{Z}^{t+1} - \boldsymbol{E}\rangle \\ + \frac{\mu^t}{2}||\boldsymbol{X} - \boldsymbol{X}_{tr}\boldsymbol{Z}^{t+1} - \boldsymbol{E}||_F^2,\end{aligned} \tag{19}$$

which is equivalent to

$$\min_{\boldsymbol{E}} \lambda_3||\boldsymbol{E}||_{21} + \frac{\mu^t}{2}||\boldsymbol{E} - (\boldsymbol{X} - \boldsymbol{X}_{tr}\boldsymbol{Z}^{t+1} + \frac{\boldsymbol{C}_1^t}{\mu^t})||_F^2. \tag{20}$$

The solution to problem (20) is demonstrated in [13]. In particular, let $\boldsymbol{\Gamma} = \boldsymbol{X} - \boldsymbol{X}_{tr}\boldsymbol{Z}^{t+1} + \frac{\boldsymbol{C}_1^t}{\mu^t}$, the i-th row of $\boldsymbol{E}^{t+1}$ is

$$\boldsymbol{E}^{t+1}(i,:) = \begin{cases} \frac{||\boldsymbol{\Gamma}^i||_2 - \frac{\lambda_3}{\mu^t}}{||\boldsymbol{\Gamma}^i||_2}\boldsymbol{\Gamma}^i, & if \quad ||\boldsymbol{\Gamma}^i||_2 > \frac{\lambda_3}{\mu^t} \\ 0, & if \quad ||\boldsymbol{\Gamma}^i||_2 \leq \frac{\lambda_3}{\mu^t}, \end{cases} \tag{21}$$

where $\boldsymbol{\Gamma}^i$ is the i-th row of matrix $\boldsymbol{\Gamma}$. Here we denote the solution of $\boldsymbol{E}$ as $\mathcal{H}_{\frac{\lambda_3}{\mu^t}}(\boldsymbol{\Gamma})$ for convenience.

After we optimize variables $\boldsymbol{P}$, $\boldsymbol{Z}$, $\boldsymbol{Q}$ and $\boldsymbol{E}$, the ADMM algorithm also needs to update the Lagrange multipliers $\boldsymbol{C}_1$, $\boldsymbol{C}_2$, $\boldsymbol{C}_3$, as well as parameter μ, for faster convergence. The detailed procedures of solving the proposed optimization problem (8) are described in Algorithm 1.

B. Recognition Method

When problem (8) is optimized by exploiting Algorithm 1, the discriminant data representation $\boldsymbol{Z} = [\boldsymbol{Z}_{tr}, \boldsymbol{Z}_{tt}]$ is obtained. We directly employ a simple linear classifier to perform

final recognition [10]. A linear classifier $\boldsymbol{W}$ is learned based on the training data representation $\boldsymbol{Z}_{tr}$ and its corresponding label matrix $\boldsymbol{L} \in \Re^{C\times n}$ of training samples. The following optimization problem is considered

$$\hat{\boldsymbol{W}} = \arg\min_{\boldsymbol{W}} ||\boldsymbol{L} - \boldsymbol{W}\boldsymbol{Z}_{tr}||_F^2 + \gamma||\boldsymbol{W}||_F^2, \tag{22}$$

where γ is a positive regularization parameter. It is easy to verify that problem (22) has the close-form solution, i.e.

$$\hat{\boldsymbol{W}} = \boldsymbol{L}\boldsymbol{Z}_{tr}^T(\boldsymbol{Z}_{tr}\boldsymbol{Z}_{tr}^T + \gamma\boldsymbol{I})^{-1}. \tag{23}$$

The identity of test sample $\boldsymbol{y}$, the i-th sample from the test dataset, is determined by judging

$$label(\boldsymbol{y}) = \arg\max_j(\hat{\boldsymbol{W}}\boldsymbol{z}^i), \tag{24}$$

where $\boldsymbol{z}^i$ is the i-th column of matrix $\boldsymbol{Z}_{tt}$. The complete procedures of our BDLRR model for recognition are summarized in Algorithm 2.

C. Convergence Analysis

To solve the proposed formulation (8), an iterative update scheme, the ADMM algorithm, is developed as shown in Section IV-A. This section presents a theoretical convergence proof of the proposed Algorithm 1.

Proposition 1: *Algorithm 1 is equivalent to a two-block ADMM.*

The classical ADMM is intended to solve problems in the form

$$\min_{\boldsymbol{z}\in\Re^n,\boldsymbol{w}\in\Re^m} f(\boldsymbol{z}) + h(\boldsymbol{w})\ s.t.\ \boldsymbol{R}\boldsymbol{z} + \boldsymbol{T}\boldsymbol{w} = \boldsymbol{u}, \tag{25}$$

where $\boldsymbol{R} \in \Re^{p\times n}$, $\boldsymbol{T} \in \Re^{p\times m}$, $\boldsymbol{u} \in \Re^p$ and f and h are convex functions. It is apparent that ADMM for problem (25) can be directly extended to solve the matrix optimization problem as follows:

$$\min_{\boldsymbol{Z}\in\Re^{n\times N},\boldsymbol{W}\in\Re^{m\times N}} f(\boldsymbol{Z})+h(\boldsymbol{W})\ s.t.\ \boldsymbol{R}\boldsymbol{Z}+\boldsymbol{T}\boldsymbol{W} = \boldsymbol{U}, \tag{26}$$

where $\boldsymbol{U} \in \Re^{p\times N}$. The augmented Lagrangian of problem (26), in the method of multipliers, is formulated as

$$\begin{aligned}\mathcal{L}_\mu(\boldsymbol{Z},\boldsymbol{W},\boldsymbol{C}) &= f(\boldsymbol{Z})+h(\boldsymbol{W})\\ &+\frac{\mu}{2}\|\boldsymbol{R}\boldsymbol{Z}+\boldsymbol{T}\boldsymbol{W}-\boldsymbol{U}\|_F^2 + \langle\boldsymbol{C},\boldsymbol{R}\boldsymbol{Z}+\boldsymbol{T}\boldsymbol{W}-\boldsymbol{U}\rangle,\end{aligned} \tag{27}$$

where $\boldsymbol{C} \in \Re^{p\times N}$ is the Lagrangian multiplier, and μ is a penalty coefficient.

It should be noted that problem (9) is a special case of problem (26). Specifically, it can be verified that the constraints in (9) can be transformed into the form of $\boldsymbol{R}\boldsymbol{Z} + \boldsymbol{T}\boldsymbol{W} = \boldsymbol{U}$, where $\boldsymbol{R} = \begin{pmatrix} -\boldsymbol{I}_n \\ -\boldsymbol{I}_n \\ \boldsymbol{X}_{tr}\end{pmatrix}$, $\boldsymbol{T} = \begin{bmatrix} \boldsymbol{I}_n & & \\ & \boldsymbol{I}_n & \\ & & \boldsymbol{I}_d\end{bmatrix}$, $\boldsymbol{W} = \begin{pmatrix}\boldsymbol{P}\\ \boldsymbol{Q}\\ \boldsymbol{E}\end{pmatrix}$, $\boldsymbol{U} = \begin{pmatrix}\boldsymbol{0}\\ \boldsymbol{0}\\ \boldsymbol{X}\end{pmatrix}$, and $\boldsymbol{I}_n$ is an $n\times n$ identity matrix. In this way, problem (9) is reformulated as problem (26). Moreover, ADMM updates two primal variables in an alternating fashion, and iteratively solves problem (27) as follows:

$$\boldsymbol{Z}^{t+1} = \arg\min_{\boldsymbol{Z}\in\Re^{n\times N}} \mathcal{L}_\mu(\boldsymbol{Z},\boldsymbol{W}^t,\boldsymbol{C}^t), \tag{28a}$$

$$\boldsymbol{W}^{t+1} = \arg\min_{\boldsymbol{W}\in\Re^{m\times N}} \mathcal{L}_\mu(\boldsymbol{Z}^{t+1},\boldsymbol{W},\boldsymbol{C}^t), \tag{28b}$$

$$\boldsymbol{C}^{t+1} = \boldsymbol{C}^t + \mu(\boldsymbol{R}\boldsymbol{Z}^{t+1} + \boldsymbol{T}\boldsymbol{W}^{t+1} - \boldsymbol{U}), \tag{28c}$$

which have the same updating procedures as Algorithm 1 in subsection IV-A. In fact, we can see that optimization of $\boldsymbol{Z}$ in (28a) is equivalent to optimize $\boldsymbol{Z}$ in (11). Furthermore, it is very important that when fixing $\boldsymbol{Z}$, solutions of $\boldsymbol{P}$ in (15), $\boldsymbol{Q}$ in (18), and $\boldsymbol{E}$ in (21), are independent on one another, for instance, computation of $\boldsymbol{E}^{t+1}$ only depends on $\boldsymbol{Z}^{k+1}$ and $\boldsymbol{C}^{k+1}$ rather than $\boldsymbol{P}^{k+1}$ or $\boldsymbol{Q}^{k+1}$. Hence, optimizations of $\boldsymbol{P}$, $\boldsymbol{Q}$ and $\boldsymbol{E}$ can be accumulated in $\boldsymbol{W}$ by using Eqn. (28b), updating of which is the same as fashion of Jacobian iterative method. In this way, problem (9) is a special case of classical ADMM problem (26), and the proposed optimization algorithm shown in Algorithm 1 has the same optimization style of classical ADMM (28). Therefore, the proposed optimization algorithm shown in Algorithm 1 is equivalent to a two-block ADMM, the global convergence of which is theoretically guaranteed [29]–[31]. The convergence nature of the proposed optimization algorithm is given by the following theorem.

Theorem 1. *([30], [31]) Consider the problem (25) where $f(\boldsymbol{Z})$ and $h(\boldsymbol{W})$ are closed proper convex functions, $\boldsymbol{R}$ has full column rank and $h(\boldsymbol{W}) + \|\boldsymbol{T}\boldsymbol{W}\|_F^2$ is strictly convex. Let $\boldsymbol{C}^0$ and $\boldsymbol{W}^0$ be arbitrary matrix and $\mu > 0$. Assume that we have the sequences $\{\gamma_t\}$ and $\{\nu_t\}$ such that $\gamma_t \geq 0$ and $\nu_t \geq 0$, $\sum_{t=0}^{\infty}\gamma_t < \infty$ and $\sum_{t=0}^{\infty}\nu_t < \infty$. Suppose that*

$$\begin{aligned}\|\boldsymbol{Z}^{t+1} - \min_{\boldsymbol{Z}} f(\boldsymbol{Z}) + \frac{\mu}{2}\|\boldsymbol{R}\boldsymbol{Z}+\boldsymbol{T}\boldsymbol{W}^t-\boldsymbol{U}\|_F^2\\ +\langle\boldsymbol{C}^t,\boldsymbol{R}\boldsymbol{Z}\rangle\|_F^2 \leq \gamma_t.\end{aligned} \tag{29}$$

$$\begin{aligned}\|\boldsymbol{W}^{t+1} - \min_{\boldsymbol{W}} h(\boldsymbol{W}) + \frac{\mu}{2}\|\boldsymbol{R}\boldsymbol{Z}^{t+1}+\boldsymbol{T}\boldsymbol{W}-\boldsymbol{U}\|_F^2\\ +\langle\boldsymbol{C}^t,\boldsymbol{T}\boldsymbol{W}\rangle\|_F^2 \leq \nu_t.\end{aligned} \tag{30}$$

$$\boldsymbol{C}^{t+1} = \boldsymbol{C}^t + \mu(\boldsymbol{R}\boldsymbol{Z}^{t+1} + \boldsymbol{T}\boldsymbol{W}^{t+1} - \boldsymbol{U}). \tag{31}$$

If there exists a saddle point of $\mathcal{L}_\mu(\boldsymbol{Z},\boldsymbol{W},\boldsymbol{C})$ (27), then $\boldsymbol{Z}^k \to \boldsymbol{Z}^*$, $\boldsymbol{W}^k \to \boldsymbol{W}^*$ and $\boldsymbol{C}^k \to \boldsymbol{C}^*$, where $(\boldsymbol{Z}^*,\boldsymbol{W}^*,\boldsymbol{C}^*)$ is such a saddle point. On the other hand, if no such saddle point exists, then at least one of the sequences $\{\gamma_t\}$ or $\{\nu_t\}$ must be unbounded.

Clearly, the optimization results shown in subsection IV-A indicate that the proposed method exists an optimal solution according to the *Proposition* 1.1.5 in [32], and the values sequences $\{\gamma_t\}$ and $\{\nu_t\}$ are directly set to zeros in Algorithm 1. Therefore, the convergence nature of our optimization method is demonstrated. Moreover, we empirically show in Section V-F that the experimental convergence of the resulting ADMM is well preserved.

D. Computational Complexity Analysis

In this section, the computational complexity for Algorithm 1 is presented, and it is easy to see that the recognition process

of Algorithm 2 is very efficient, which is linear with the sample number. More specifically, the major computation cost of Algorithm 1 is in steps 1-4, which require computing the singular value decomposition (SVD) and matrix computation operation. Thus, they will be time consuming when the number of training samples n and the total number of samples N are very large. In particular, computing SVD decomposition of matrix $\boldsymbol{P} \in \Re^{n\times N}$ needs the complexity of $\mathcal{O}(n^2N)$ $(N > n)$. Note that due to the matrix inverse calculation, calculating $\boldsymbol{Z}$ will scale in about $\mathcal{O}(n^2d+n^2N)$ where d is the dimensionality of the samples. The computational complexity of step 3 is $\mathcal{O}(nN)$, and computing $\boldsymbol{E}$ in step 4 costs $\mathcal{O}(dN)$. Therefore, the total computational complexity of BDLRR is $\mathcal{O}_\kappa(2n^2N + n^2d + dN + nN)$, where κ is the number of iterations.

In comparison, the computation burden of the sparse representation based classification methods such as SRC, LRSI and LatLRR are $\mathcal{O}(n^2(N-n)d)$ by solving $(N-n)$ independent l_1-norm minimization problems in an iterative optimization manner [3], [16], [22], which is slower than that of our method. The computation complexities of regression methods such as LRLR and LRRR are $\mathcal{O}(dn + n^2d)$, which is a little faster than our method. The low-rank and sparse representation based methods such as NNLRS, SRRS, CBDS, and our BDLRR need to simultaneously compute SVD of feature matrix and solve a simple soft-thresholding problem, and a linear classification algorithm is used to predict final labels of test data. Generally, the overall computation burden of our BDLRR is the same as those of the low-rank sparse representation learning methods.

E. Out-of-sample Extension

It is worth noting that the low-rank representation based methods have been extensively studied, but how to address the out-of-sample problem, the capability of dealing with new data instances, is much less well-solved. The stage of BDLRR mentioned above only obtains the discriminative representations of the available samples $\boldsymbol{X} \in \Re^{d\times N}$. However, given unseen instances outside the training and test data, it would be unrealistic and time-consuming to reimplement the whole model to produce the representations of novel images. In this subsection, we will show that the proposed BDLRR method can naturally cope with the out-of-sample examples to learn discriminative visual representations.

Suppose we have obtained the optimal block-diagonal representation $\boldsymbol{Z} \in \Re^{n\times N}$ from the available samples $\boldsymbol{X}$ over $\boldsymbol{X}_{tr} \in \Re^{d\times n}$ using the proposed model (8). Now, we extend the proposed BDLRR method to learn preferable representation of a novel image $\boldsymbol{b} \in \Re^{d\times 1}$ in the original observed space. Specifically, we aim at learning the discriminative representation $\boldsymbol{z}$ for $\boldsymbol{b}$ over $\boldsymbol{X}_{tr}$, while fixing the previously learned representation $\boldsymbol{Z}$. Therefore, adding terms for a novel data point $\boldsymbol{b}$ in model (8) and keeping the already learned variables, the objective function of the augmented BDLRR is formulated as

$$\begin{aligned}\min_{\boldsymbol{z},\boldsymbol{e}} &|||[\boldsymbol{Z},\boldsymbol{z}]||_* + \lambda_1||\hat{\boldsymbol{A}} \odot [\boldsymbol{Z},\boldsymbol{z}]||_F^2 + \lambda_2||\hat{\boldsymbol{D}} \odot [\boldsymbol{Z},\boldsymbol{z}]||_1 \\ &+\lambda_3||[\boldsymbol{E},\boldsymbol{e}]||_{21} \quad s.t. \quad [\boldsymbol{X},\boldsymbol{b}] = \boldsymbol{X}_{tr}[\boldsymbol{Z},\boldsymbol{z}] + [\boldsymbol{E},\boldsymbol{e}],\end{aligned} \tag{32}$$

where $\hat{\boldsymbol{A}} = [\boldsymbol{A}, \boldsymbol{1}_n\boldsymbol{1}_{N+1-n}^T]$, $\boldsymbol{A}$ is defined as in Eqn. (8), $\hat{\boldsymbol{D}} \in \Re^{n\times(N+1)}$ is the distance metric between the training samples $\boldsymbol{X}_{tr}$ and all samples $[\boldsymbol{X},\boldsymbol{b}]$, and $\boldsymbol{e}$ is the representation error of $\boldsymbol{b}$ over $\boldsymbol{X}_{tr}$. We argue that $\|[\boldsymbol{Z},\boldsymbol{z}]\|_* = \|\boldsymbol{Z}\|_*$. Particularly, for the learned representation $\boldsymbol{Z} \in \Re^{n\times N}(n < N)$, it is easy to find that the linear problem for $\boldsymbol{\alpha}$ in $\boldsymbol{z} = \boldsymbol{Z}\boldsymbol{\alpha}$ is an underdetermined system for practical data. Generally speaking, $\boldsymbol{z} = \boldsymbol{Z}\boldsymbol{\alpha}$ has infinitely many solutions in practice [33]. Provided $n \ll N$, the matrix $\boldsymbol{Z}$ is row full rank and $\boldsymbol{z} = \boldsymbol{Z}\boldsymbol{\alpha}$ has solution. In this way, the singular values of matrix $\boldsymbol{Z}$ coincide with those of $[\boldsymbol{Z},\boldsymbol{z}]$, which means $rank([\boldsymbol{Z},\boldsymbol{z}]) = rank(\boldsymbol{Z})$. Therefore, $\|[\boldsymbol{Z},\boldsymbol{z}]\|_* = \|\boldsymbol{Z}\|_*$ and it does not change for practical data in Eqn. (32). By removing the irrelevant terms with respective to the variables $\boldsymbol{z}$ and $\boldsymbol{e}$, it is easy to check that problem (32) will be degenerated to the following formulation:

$$\begin{aligned}\min_{\boldsymbol{z},\boldsymbol{e}} &\lambda_1\|\boldsymbol{z}\|_2^2 + \lambda_2\|\boldsymbol{d} \odot \boldsymbol{z}\|_1 + \lambda_3\|\boldsymbol{e}\|_2 \\ &s.t. \quad \boldsymbol{b} = \boldsymbol{X}_{tr}\boldsymbol{z} + \boldsymbol{e},\end{aligned} \tag{33}$$

which can be equivalently reformulated as

$$\begin{aligned}\min_{\boldsymbol{z},\boldsymbol{e}} &\lambda_1\|\boldsymbol{z}\|_2^2 + \lambda_2\|\boldsymbol{d} \odot \boldsymbol{z}\|_1 + \lambda_3\|\boldsymbol{e}\|_2^2 \\ &s.t. \quad \boldsymbol{b} = \boldsymbol{X}_{tr}\boldsymbol{z} + \boldsymbol{e},\end{aligned} \tag{34}$$

where $\boldsymbol{d}_i$ is the distance between $\boldsymbol{x}_i$ and $\boldsymbol{b}$. To make problem (34) more compact, it can be rewritten as

$$\min_{\boldsymbol{z}} \frac{1}{2}\|\boldsymbol{b} - \boldsymbol{X}_{tr}\boldsymbol{z}\|_2^2 + \frac{\beta_1}{2}\|\boldsymbol{z}\|_2^2 + \beta_2\|\boldsymbol{d} \odot \boldsymbol{z}\|_1, \tag{35}$$

where $\beta_1 = \lambda_1/\lambda_3$ and $\beta_2 = \lambda_2/2\lambda_3$. Apparently, problem (35) is an elastic-net regularized regression problem. For convenient interpretation, we denote $g(\boldsymbol{z}) = \frac{1}{2}\|\boldsymbol{b} - \boldsymbol{X}_{tr}\boldsymbol{z}\|_2^2 + \frac{\beta_1}{2}\|\boldsymbol{z}\|_2^2$. With some algebra, problem (35) can be approximately transformed to the following optimization problem:

$$\begin{aligned}\boldsymbol{z}^{k+1} &= \arg\min_{\boldsymbol{z}} \beta_2\|\boldsymbol{d} \odot \boldsymbol{z}\|_1 + \langle\nabla_{\boldsymbol{z}} g(\boldsymbol{z}^k), \boldsymbol{z} - \boldsymbol{z}^k\rangle + \frac{\eta}{2}\|\boldsymbol{z} - \boldsymbol{z}^k\|_2^2 \\ &= \arg\min_{\boldsymbol{z}} \beta_2\|\boldsymbol{d} \odot \boldsymbol{z}\|_1 + \frac{\eta}{2}\|\boldsymbol{z} - \boldsymbol{z}^k + \nabla_{\boldsymbol{z}} g(\boldsymbol{z}^k)/\eta\|_2^2 + const,\end{aligned} \tag{36}$$

where $\boldsymbol{z}^k$ is the k-th iteration of $\boldsymbol{z}$, and $\eta = \|\boldsymbol{X}_{tr}\|_F^2$ is a fixed step size in our paper. Similar to problem (18), the optimal solution of the i-th entry of $\boldsymbol{z}$ is calculated by using $\boldsymbol{z}_i^{k+1} = \mathcal{S}_{\frac{\beta_2\boldsymbol{d}_i}{\eta}}([\boldsymbol{z}^k - \nabla_{\boldsymbol{z}} g(\boldsymbol{z}^k)/\eta]_i)$. After obtaining the optimal solution $\hat{\boldsymbol{z}}$, we identify the new data instance $\boldsymbol{b}$ by employing Eqn. (24), i.e. $label(\boldsymbol{b}) = \arg\max_j(\hat{\boldsymbol{W}}\boldsymbol{z})$. The promising recognition results can be guaranteed based on the observation that the discriminative block-diagonal training representations are learned in the training stage. Therefore, based on the proposed BDLRR model, the problem of recognizing new instances outside the training and test samples is well addressed.

F. Discussion

As we know, BDLRR simultaneously takes advantages of supervised information, i.e. label information, and semi-supervised learning superiority, i.e. learning training and test representations in one formulation. Moreover, our method intrinsically inherits the superiorities of sparse, low-rank, structured and elastic-net representation learning techniques. This characteristic naturally differentiates it from previous

works, yielding superior recognition results. In this section, we establish the relationships between the proposed BDLRR method and some related discriminative low-rank representation methods, such as the nonnegative low-rank representation sparse (NNLRS) method [25], the structured sparse and low-rank representation (SSLR) method [10], and the very recently proposed supervised regularization based robust subspace (SRRS) method [24].

1) Connection to the NNLRS method: The NNLRS method focuses on constructing the informative graph by jointly considering the low-rank and sparse representation to capture the global and local structures of data, respectively. Specifically, the objective function of NNLRS is formulated as

$$\begin{aligned} &\min_{\boldsymbol{Z},\boldsymbol{E}} \|\boldsymbol{Z}\|_* + \lambda_2\|\boldsymbol{Z}\|_1 + \lambda_3\|\boldsymbol{E}\|_{21} \\ &s.t.\ \boldsymbol{X_{tr}} = \boldsymbol{X_{tr}}\boldsymbol{Z} + \boldsymbol{E}, \boldsymbol{Z} \geq \boldsymbol{0}. \end{aligned} \tag{37}$$

The rationale of NNLRS is under the guidance of the observation that the sparse constraint ensures each sample connected to only few other samples resulting in sparse representation, while the low-rank constraint enforces the learned representation from the same class with high correlations. In other words, NNLRS is designed to capture the global structure of the training data using the low-rank property, and the locality information of each data vector is interpolated into NNLRS by introducing the sparse term. The following proposition shows the close relationship between the proposed BDLRR method and the LRR and NNLRS methods.

Proposition 2: *The proposed BDLRR method is a generalized but discriminative low-rank representation learning model, and both of LRR and NNLRS are the special cases of the proposed BDLRR method.*

Proof. From the objective function of BDLRR, i.e. Eqn. (8) in the main paper, if we set both balance parameters $\lambda_1 = 0$ and $\lambda_2 = 0$, it is easy to find that BDLRR will degenerate to LRR along with learning the training and test representations in a semi-supervised manner. Moreover, if the penalty parameter $\lambda_1 = 0$ and $\boldsymbol{X} = \boldsymbol{X}_{tr}$, BDLRR will be reformulated as a low-rank and weighted sparse representation learning model. As a result, BDLRR will degenerate to weighted NNLRS without considering the nonnegative constraint. Therefore, both of LRR and NNLRS are the special cases of the proposed BDLRR method.

More importantly, our BDLRR method jointly considers suppressing the unfavorable representations from off-block-diagonal components and highlighting the compact block-diagonal representations under the framework of the semi-supervised low-rank representation learning such that the margins between different classes are greatly enlarged and the intra-class compactness is also enhanced simultaneously. In this way, BDLRR takes the intra-class and inter-class visual correlations into consideration to concurrently learn both discriminative representations of training and test data in one unified learning paradigm. As a result, our method can be viewed as a generalized discriminative representation learning framework.

Therefore, our BDLRR method not only intrinsically generalizes the previous LRR and NNLRS models, but also extends the existing low-rank representation models to more robust and discriminative cases. □

2) Comparison with the SSLR method: SSLR first learns a structured low-rank sparse dictionary by imposing an ideal representation regularization term, and then a structured low-rank representation is achieved based on the learned dictionary. The objective function of SSLR is

$$\begin{aligned} \min_{\boldsymbol{Z},\boldsymbol{E},\boldsymbol{\Xi}} &\|\boldsymbol{Z}\|_* + \lambda_1\|\boldsymbol{Z}\|_1 + \lambda_2\|\boldsymbol{E}\|_1 + \lambda_3\|\boldsymbol{Z} - \boldsymbol{Q}\|_F^2 \\ &s.t.\ \boldsymbol{X_{tr}} = \boldsymbol{\Xi}\boldsymbol{Z} + \boldsymbol{E}, \end{aligned} \tag{38}$$

where $\boldsymbol{\Xi}$ is the learned dictionary. $\boldsymbol{Q}$ is the ideal data representation of training samples, i.e. $\begin{bmatrix} \mathbf{1}_{s_1}\mathbf{1}_{s_1}^T & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \mathbf{1}_{s_C}\mathbf{1}_{s_C}^T \end{bmatrix}$, where s_i is the number of the i-th class of $\boldsymbol{\Xi}$. By solving the optimization problem (38), the learned dictionary $\boldsymbol{\Xi}$ is obtained, and then representations of the training and test data are respectively achieved by directly removing the ideal representation term from (38), resulting in the following optimization problem:

$$\min_{\boldsymbol{Z},\boldsymbol{E}} \|\boldsymbol{Z}\|_* + \lambda_1\|\boldsymbol{Z}\|_1 + \lambda_2\|\boldsymbol{E}\|_1 \ s.t.\ \boldsymbol{B} = \boldsymbol{\Xi}\boldsymbol{Z} + \boldsymbol{E}, \tag{39}$$

where $\boldsymbol{B}$ is the observations, i.e. $\boldsymbol{X}_{tr}$ or $\boldsymbol{X}_{tt}$. Although the experimental results reported in [10] are good, we hold the view that enforcing the representation approximal to the ideal representation matrix $\boldsymbol{Q}$ is questionable because it is impossible to regularize all the training samples of the same class to have the same representation codes. Moreover, the solution of learning $\boldsymbol{\Xi}$ by solving problem (38) sensitively depends on the initialization because of the nonconvex optimization. Furthermore, learning representations of the training and test data are divided into two separate stages, and there are respectively three and two parameters in (38) and (39), which are very difficult to tune.

In contrast, our method is reasonable and discriminative. BDLRR first shrinks the off-block-diagonal elements to eliminate the unfavorable representations resulting in marginalized inter-class representations and highlights the block-diagonal elements yielding compact intra-class representations. In this way, the discriminative constraints in BDLRR simultaneously separates the common visual representations from different classes, and effectively prevents zero entities from appearing in the class-specific representations. Moreover, we believe that it is significant for recognition that the learned representations of training and testing samples should be consistent. To this end, BDLRR builds the representation bridge between the training and test samples by imposing the low-rank and locality coherence property. Thus, the proposed BDLRR method unifies the discriminative representations of training and test data into one robust learning framework such that better recognition results are achieved.

3) Comparison with the SRRS method: The main objective of SRRS is dedicated to learning a discriminative subspace from the clean data recovered by using the low-rank representation constraint. The main idea of SRRS is to remove noise from contaminated data depending on the denoising capability

of the low-rank representation, and then the discriminative subspace is learned based on the recovered 'clean' data. The objective function of SRRS is

$$\begin{aligned}\min_{\boldsymbol{Z},\boldsymbol{E},\boldsymbol{P}} & \|\boldsymbol{Z}\|_* + \lambda_2\|\boldsymbol{E}\|_{21} + \eta\|\boldsymbol{P}^T\boldsymbol{X}\boldsymbol{Z}\|_F^2 \\ & + \lambda_1[tr(S_b(\boldsymbol{P}^T\boldsymbol{X}\boldsymbol{Z})) - tr(S_w(\boldsymbol{P}^T\boldsymbol{X}\boldsymbol{Z}))] \\ s.t.\ & \boldsymbol{X}_{tr} = \boldsymbol{X}_{tr}\boldsymbol{Z} + \boldsymbol{E}, \boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{I},\end{aligned} \tag{40}$$

where $S_b(\bullet)$ and $S_w(\bullet)$ are respectively the between-class and within-class scatter matrices, and η is a balance parameter.

Apparently, our method is different from SRRS. First, SRRS directly utilizes the 'clean' data $\boldsymbol{X}\boldsymbol{Z}$ to perform discriminant analysis, which means that the performance of SRRS is greatly subject to the denoising ability of LRR. However, BDLRR aims at directly learning discriminative representations from data by imposing the discriminant constraints, which are not confined to any other conditions. Moreover, SRRS is a subspace learning method, and then the dimension selection of the final representations is very important for recognition. However, BDLRR directly learns discriminative representations from data, and recognition is performed on the optimal representations without bearing the burden of dimension selection. In addition, our BDLRR method jointly learns the representations of training and test data, whereas the test representations of SRRS is achieved by using $\boldsymbol{P}\boldsymbol{X}_{tt}$, which can not capture the component connections between the learned representations of the training and test data. Therefore, the proposed BDLRR method is more robust and discriminative than SRRS, which is also verified by the subsequent experimental results.

V. Experimental Validation

In this section, the performance of the proposed BDLRR method is evaluated for different recognition tasks. Extensive experiments are performed on different types of datasets to demonstrate the effectiveness and superiority of the proposed method in comparison with state-of-the-art recognition methods. Subsequently, the algorithmic convergence and the selection of parameters are well analyzed.

A. Experimental Setup

We test our method on eight benchmark datasets for three basic recognition tasks. Moreover, we compare with some state-of-the-art recognition methods, including representation based methods (such as LRC [18], CRC [16], SRC [3] and LLC [4]), low-rank criterion based methods (such as RPCA [21], LatLRR [7], Low-rank linear regression (LRLR) [20], Low-rank robust regression (LRRR) [20], CBDS [23], LRSI [22], NNLRS [25], SRRS [24]), and conventional classification methods such as support vector machine (SVM) [34] with Gaussian kernel. We randomly select several images per class to construct the training dataset, and the rest of images are regarded as the test set. All the selection processes are repeated 10 times, and the average recognition accuracies are reported for all the methods.

For fair comparison in all experiments, we use the Matlab codes from the corresponding authors with the default or optimal parameter settings, or directly cite the experimental results from their original papers. More specifically, for RPCA [21], we first use the original RPCA algorithm on both training and test datasets to eliminate some noise and corrupted terms, and then exploit SRC [3] for recognition. For LatLRR [7], the learned salient features are used for recognition. For SVM, the LibSVM software [34] is used for multi-class recognition, where the important regularization parameter C in SVM is selected by cross-validation from the candidate set $\{0.01, 0.1, 1.0, 10.0, 100.0, 1000.0\}$. The parameters of our method, i.e. λ_1, λ_2 and λ_3, are tuned to achieve the best performance via 5-fold cross validations from [0.1, 0.5, 1, 5, 10, 15, 20, 25]. To guarantee the same experimental settings between all the compared methods and our method on each benchmark, we re-implemented all the algorithms using respective optimal parameters via the cross-validation strategy, and the training and test samples were randomly selected from each dataset ten times. Since the scene character recognition datasets have the standard splits of the training and test data, we directly employ the full training and test data for recognition, and the compared experimental results are cited from the original papers. Similarly, for scene recognition, experiments are performed with the same experiment protocols as that of the LC-KSVD method [39], and we directly cite some experimental results from the original papers. For the compared methods that are not included in [39], we rerun them following the same experimental settings. Therefore, all the methods presented in our paper are performed on the same testbed for each dataset such that our experimental results are convincing and reliable. All algorithms are implemented with Matlab 2013a, and *the Matlab code of the proposed method has been released at http://www.yongxu.org/lunwen.html*.

TABLE I: Recognition accuracies (mean±std %) of different methods with different numbers of training samples on the Extended YaleB database.

Alg.	20	25	30	35
LRC	92.15±0.95	93.55±0.65	94.55±0.68	95.49±0.55
CRC	94.36±1.17	95.89±0.91	97.14±0.75	97.93±0.55
SRC	93.73±0.70	95.58±0.26	96.37±0.45	97.13±0.42
LLC	91.60±0.50	94.20±0.49	95.29±0.38	96.05±0.51
SVM	92.81±0.68	95.20±0.44	96.11±0.41	96.70±0.69
RPCA	93.58±0.61	95.51±0.36	96.70±0.46	96.96±0.49
LatLRR	93.05±0.95	93.91±0.68	95.03±0.83	97.14±0.36
LRLR	83.91±1.53	85.15±1.50	85.49±1.05	85.95±1.47
LRRR	83.95±0.82	85.66±0.93	86.21±0.99	86.55±0.81
CBDS	95.99±1.11	96.56±0.85	97.61±0.82	98.13±0.55
LRSI	94.19±0.44	96.28±0.61	96.99±0.57	97.72±0.48
NNLRS	94.35±0.79	96.06±0.63	97.02±0.61	97.62±0.42
SRRS	93.74 ±0.86	96.05±0.95	96.89±0.84	97.15±0.58
BDLRR	**96.89±0.67**	**97.96±0.42**	**98.70±0.46**	**99.46±0.29**

B. Experiments for Face Recognition

In this section, we perform experiments on four face image datasets, including the Extended YaleB [35], CMU PIE [36], AR [37] and LFW [38] datasets.

The Extended YaleB Database: The extended YaleB database is composed of 2414 face images of 38 subjects, where each person has 59-64 near frontal images under different illumination conditions. All the images for our experiments

TABLE II: Recognition accuracies (mean±std %) of different methods with different numbers of training samples on the CMU PIE database.

Alg.	20	25	30	35
LRC	90.07±0.52	92.65±0.38	94.11±0.26	94.88±0.17
CRC	92.52±0.33	93.84±0.39	94.31±0.16	95.52±0.16
SRC	92.14±0.29	93.65±0.38	94.51±0.28	95.86±0.24
LLC	91.90±0.25	93.27±0.56	94.66±0.41	95.26±0.49
SVM	90.69±0.73	92.78±0.68	93.19±0.51	94.10±0.30
RPCA	88.34±0.32	91.56±0.18	92.96±0.22	93.81±0.36
LatLRR	88.84±0.32	91.96±0.82	93.26±0.22	94.41±0.38
LRLR	85.83±0.56	86.90±0.45	87.82±0.49	88.23±0.42
LRRR	85.98±0.61	86.89±0.58	88.50±0.87	89.06±0.42
CBDS	91.81±0.62	93.50±0.73	94.45±0.77	94.90±0.68
LRSI	90.68±0.56	93.55±0.69	94.62±0.54	95.12±0.26
NNLRS	91.72±0.43	92.04±0.53	93.55±0.40	94.38±0.39
SRRS	90.87±0.61	93.16±0.45	94.41±0.35	95.15±0.27
BDLRR	**94.67±0.31**	**95.79±0.29**	**96.46±0.15**	**96.81±0.14**

TABLE III: Recognition accuracies (mean±std %) of different methods with different numbers of training samples on the AR database.

Alg.	11	14	17	20
LRC	76.97±1.33	85.51±1.20	90.99±0.97	94.22±0.76
CRC	91.76±0.77	94.36±0.97	95.84±0.76	96.63±0.87
SRC	89.62±0.74	92.35±1.29	95.24±0.67	96.19±0.75
LLC	60.89±0.97	66.98±1.13	71.58±1.32	73.53±2.15
SVM	86.30±1.33	92.03±0.77	95.19±0.88	96.43±1.26
RPCA	84.53±1.43	88.92±0.95	92.62±0.77	94.90±0.78
LatLRR	92.83±1.06	95.96±0.70	97.13±0.85	97.78±0.56
LRLR	88.93±0.86	93.33±0.73	94.92±0.68	96.37±0.88
LRRR	93.82±0.70	95.42±0.48	96.47±0.70	96.88 ±0.61
CBDS	92.99±0.59	95.57±0.60	96.83±0.63	97.49±0.82
LRSI	86.93±1.00	90.02±0.76	93.27±0.97	94.82±0.99
NNLRS	92.11±0.70	95.24±0.49	96.69±0.56	97.40±0.65
SRRS	87.53±1.00	93.33±1.04	96.22±1.03	97.17±0.54
BDLRR	**96.69±0.41**	**97.92± 0.30**	**98.72±0.42**	**99.03±0.38**

TABLE IV: Recognition accuracies (mean±std %) of different methods with different numbers of training samples on the LFW database.

Alg.	5	6	7	8
LRC	29.48±1.48	33.63±1.76	35.57±1.89	37.63±1.99
CRC	29.64±1.22	31.79±1.52	32.96±1.32	33.86±1.55
SRC	29.13±1.27	32.25±1.55	33.46±2.10	36.51±2.24
LLC	27.63±1.62	29.58±1.39	31.16±1.28	31.94±0.88
SVM	30.72±1.57	33.36±1.70	36.46±1.42	37.73±1.45
RPCA	31.55±1.27	34.17±1.65	36.68±1.88	37.99±1.36
LatLRR	30.00±1.11	33.09±1.95	35.33±1.91	37.28±1.68
LRLR	29.68±1.05	30.18±1.01	34.55±1.82	35.39±2.17
LRRR	30.98±1.28	32.93±1.70	34.86±1.03	36.59±1.87
CBDS	34.77±1.46	36.54±1.81	37.50±1.56	38.53±1.79
LRSI	31.57±2.10	34.42±1.25	37.18±0.92	39.25±1.58
NNSLR	34.59±0.92	35.51±1.49	36.83±0.93	39.96±1.53
SRRS	31.67±1.54	34.29±1.74	38.06±1.59	39.43±1.65
BDLRR	**37.83±1.00**	**40.94±1.78**	**43.11±1.45**	**44.51±1.15**

on this database have been resized to 32×32 pixels. For all the compared methods, the suggested parameters from the corresponding papers are used for recognition. For the LLC [4] method, we directly treat the training samples as the bases, and the coding coefficients are obtained using the approximated LLC strategy. The number of neighbors of LLC is set to fifteen for this dataset, which can achieve the highest recognition accuracies. In the experiments, we randomly select 20, 25, 30, 35 images per subject for training and the rest for testing. The recognition accuracies of different methods on this database are shown in Table I. Note that the mean classification accuracies and the corresponding standard deviations (acc±std) are reported, and the bold numbers suggest the highest recognition accuracies. From Table I, it is easy to find that our method can consistently achieve the highest recognition results, and outperforms the other eleven competing methods significantly, even when using a small number of training samples. Moreover, the experimental results also validate that our method has an outstanding capability on overcoming the challenges of illumination and expression variations.

The CMU PIE Database: The CMU PIE face database contains more than 40,000 face images of 68 individuals in total. In our experiments, we utilize the images under five near frontal poses (C05, C07, C09, C27 and C29), and then about 170 image samples are obtained for each individual. We randomly select 20, 25, 30, 35 images from each subject as training samples and the remaining images are regarded as test samples. Each image is cropped and resized to be only 32×32 pixels. The detailed comparison results obtained using different methods are summarized in Table II. We can see that, with different numbers of training samples per class, our results are always better than those of all the other state-of-the-art methods, which demonstrates the effectiveness of our method.

The AR Database: The AR face database contains about 4,000 color face images of 126 subjects. For each subject, there are 26 images taken in two separate sessions under different conditions. In our experiments, we randomly choose a subset including 2600 images of 50 female and 50 male subjects.Random face images of the AR face database[1] are employed in our experiments. Following the implementation in [39], each image is projected onto a 540-dimensional feature vector with a randomly generated matrix with a zero-mean normal distribution. We randomly select 11, 14, 17, 20 images of each subject as training samples and treat the remaining images as test samples. The experimental results obtained using different recognition methods are shown in Table III. From the results shown in Table III, we know that our method still achieves the best recognition results, which also verifies the fact that the proposed method has particular potential for image recognition. It is notable that even when using smaller number of training samples, the performance gain of our method is still obvious in comparison with other methods.

The LFW Database: The Labeled Faces in the Wild (LFW) face database is designed for the study of unconstrained identity verification and face recognition. It contains more than 13,000 face images from 1680 subjects pictured under the unconstrained conditions. In our experiments, we employ a subset including 1251 images from 86 people, and each subject has only 10-20 images [40] with an imbalanced number of samples. Each image was manually cropped and

[1]This dataset is publicly available from http://www.umiacs.umd.edu/~zhuolin/projectlcksvd.html.

TABLE V: Recognition accuracies (mean±std %) of different methods with different numbers of training samples on the USPS database.

Alg.	30	60	90	120
LRC	89.53±0.40	92.68±0.29	94.17±0.22	94.94±0.13
CRC	89.53±0.63	90.79±0.30	91.47±0.32	91.71±0.23
SRC	90.06±0.61	93.46±0.22	94.87±0.25	95.38±0.28
LLC	91.30±0.46	93.72±0.23	94.78±0.22	95.42±0.28
SVM	90.77±0.70	92.67±0.33	93.59±0.25	94.01±0.24
RPCA	90.07±0.29	93.54±0.34	94.72±0.12	95.38±0.20
LatLRR	88.75±0.70	90.26±0.55	91.08±0.34	91.56±0.33
LRLR	84.71±2.02	87.91±0.82	88.17±0.76	88.66±0.47
LRRR	86.02±2.16	88.22±0.84	88.38±0.69	88.77±0.45
CBDS	87.80±0.69	89.46±0.52	90.46±0.24	91.54±0.19
LRSI	90.62±0.41	93.51±0.31	94.54±0.17	95.39±0.18
NNSLR	90.54±0.57	93.00±0.35	94.03±0.22	94.88±0.33
SRRS	91.13±0.20	92.93±0.36	93.94±0.21	94.44±0.20
BDLRR	**92.90±0.32**	**95.08±0.27**	**95.91±0.25**	**96.41±0.25**

resized to 32 × 32 pixels. In our experiments, we randomly select 5, 6, 7 and 8 images of each subject as training samples and the remaining face images are treated as test samples. The experimental results of different recognition methods on this dataset are presented in Table IV. We can see that the best recognition results are still achieved by our BDLRR method. Especially, the performance of our method has exceedingly advantages for this dataset in comparison with the rest of methods.

C. Experiments for Character Recognition

In this section, we evaluate the performance of our method for character recognition. More specifically, three character image datasets are employed for our experiments, including one handwriting dataset (i.e. the USPS [41] dataset) and two scene character recognition datasets (i.e. the Char74K [42] and SVT [43] datasets). It is worth noting that this work for the first time learns discriminative data representations for scene character recognition.

1) Handwriting image recognition: **The USPS Database**[2] refers to numeric data images cropped from the scanning of handwritten digits from envelopes. It consists of 9,298 handwritten digits ('0'-'9').All the images are resized into 16 × 16 pixels with 8-bit grayscale images. Each digit has about 1,100 images. In the experiments, we randomly choose 30, 60, 90 and 120 images of each digit as training samples, and regard the rest of images as test samples. The experimental results of different methods with varying numbers of training samples are shown in Table V. The proposed method performs consistently better than all the compared methods, which further confirms that the proposed method has apparent advantages on recognizing handwriting digit images.

2) Scene character image recognition: : Two scene character image datasets are utilized for measuring the effectiveness of our method. As we know, natural scene character recognition is a typical yet challenging pattern recognition task due to the cluttered background, which is very difficult to separate from text. We evaluate the performance of our method in comparison with the state-of-the-art methods experimented on both datasets, including CoHOG [45], ConvCoHOG [46], PHOG [47], MLFP [48], RTPD [49], GHOG [50], LHOG [50], HOG+NN [43], SBSTR [51] and GB [42] (GB+SVM, GB+NN). All the images in the experiments are first resized into 32 × 32 pixels, and gray scale images are used in all the experiments. To make fair comparisons, we directly employ the standard partitions of training and test samples for each dataset as in [44]–[46], and the state-of-the-art algorithms evaluated on respective datasets are directly cited from their original papers. For the features[3] used in our experiments, we exploit the method in a recent paper [44] for feature extraction. Specifically, we first use RPCA [21] to jointly remove noisy pixels and recover clean character images from the blurred or corrupted images, and then the well-known HOG method is applied to extract gradient features from the recovered images. The obtained HOG features are utilized for recognition.

The Char74K Database was collected for the study of recognizing characters in images of natural scenes. An annotated database of images including English and Kannada characters were obtained from images captured in Bangalore and India. We mainly focus on the recognition of English characters and digits (i.e. '0'-'9', 'A'-'Z','a'-'Z') with 62 classes in total. In our experiments, a small subset is used in our experiments, i.e. Char74K-15, which contains 15 training samples and 15 test samples per class. Table VI presents the recognition results of our method and several recently proposed character recognition methods. From Table VI, we can see that our method can continually achieve the highest recognition results in comparison with the state-of-the-art methods. Specifically, it is easy to see that our method outperforms the second best algorithm by a large margin of three percent.

The Street View Text (SVT) Database was collected from Google Street View of road-side scenes. All the images are very difficult and challenging to recognize due to the large variations in illumination, character image size, and font size and style. The SVT character dataset, which was annotated in [43], is utilized for evaluating different scene character recognition methods. About 3,796 character samples from 52 categories (no digit images) are annotated for recognition. Moreover, the SVT character dataset is more difficult to recognize than the Char74K dataset. The experimental results of using different methods on the SVT dataset are summarized in Table VI. For this dataset, the proposed method significantly outperforms all the other state-of-the-art methods. We can see that the proposed method (BDLRR) achieves 79% accuracy, which improves the accuracy by 4% in comparison with the second best competitors such as SRC used in [44], CoHOG [45] and PHOG [47].

D. Experiments for Scene Recognition

The performance of the proposed method for scene recognition is evaluated on the fifteen scene categories database [52]. It contains 4485 scene images falling into 15 categories

[2]In this study, the publicly available set is from http://cs.nyu.edu/~roweis/data.html is used.

[3]The features of both datasets are publicly available at http://www.yongxu.org/databases.html.

TABLE VI: Recognition accuracies (%) of different methods on the scene character database.

Alg.	Testing datasets Accuracy	
	Char74K-15	*SVT*
BDLRR	**70**	**79**
RPCA+HOG+SRC [44]	67	75
RPCA+HOG+Linear SVM [44]	63	73
RPCA+HOG+SVM(RBF) [44]	63	74
ConvHOG [46]	-	75
CoHOG [45]	-	73
PHOG (Chi-Square Kernel) [47]	-	75
MLFP [48]	64	-
RTPD [49]	-	67
GHOG+SVM [50]	62	-
LHOG+SVM [50]	58	-
SBSTR [51]	60	74
HOG+NN [43]	58	68
GB+SVM [42]	53	-
GB+NN [42]	47	-

TABLE VII: Recognition accuracies (mean $\pm$ std %) of different methods on the fifteen scene categories database.

Alg.	Accuracy		Alg.	Accuracy
LLC	79.4		SVM	93.6
LLC*	89.2		LRSI	92.4
LRC	91.9		CBDS	95.7
CRC	92.3		LRRC [10]	90.1
SRC	91.8		SLRRR [10]	91.3
LRLR	94.4		SRRS	95.9
LRRR	87.2		Lazebnik [52]	81.4
RPCA	92.1		Lian [53]	86.4
NNLRS	96.4		Yang [54]	80.3
LatLRR	91.5		Boureau [55]	84.3
LC_KSVD1 [39]	90.4		Gao [56]	89.7
LC_KSVD2 [39]	92.9		**BDLRR**	**98.9 $\pm$ 0.19**

including livingroom, bedroom, mountain, outdoor street, suburb, industrial, kitchen, opencountry, coast, forest, highway, insidecity, tallbuilding, office and store. The features[4] of fifteen scene categories provided in [39] is employed for recognition. More specifically, the obtained features are processed as the following steps. First, the spatial pyramid feature with a four-level spatial pyramid [52] is computed on a SIFT-descriptor codebook with a size of 200, and then the spatial pyramid features are reduced to 3,000 by exploiting PCA to make feature dimension reduction. Following the same experimental setting of [39] [52], we randomly select 100 images per category as training data, and regard the remaining samples as test samples. For LLC, the numbers of local bases of LLC* and LLC are set to 30 and 70 respectively, which are the same parameters used in [4] [39]. Similar to above experiments, we also report the mean recognition results (mean$\pm$std) of our method over 10 times run. For fair comparison, we directly cite the results reported in LC-KSVD [39] for performance evaluation. The experimental results are summarized in Table VII. There is no doubt that our approach maintains the highest recognition accuracies and outperforms all the competing methods. Specifically, at least three percent improvements are achieved when comparing with the other methods.

[4]In this experiment, the features used are publicly available at http://www.umiacs.umd.edu/~zhuolin/projectlcksvd.html.

E. Experimental Analysis

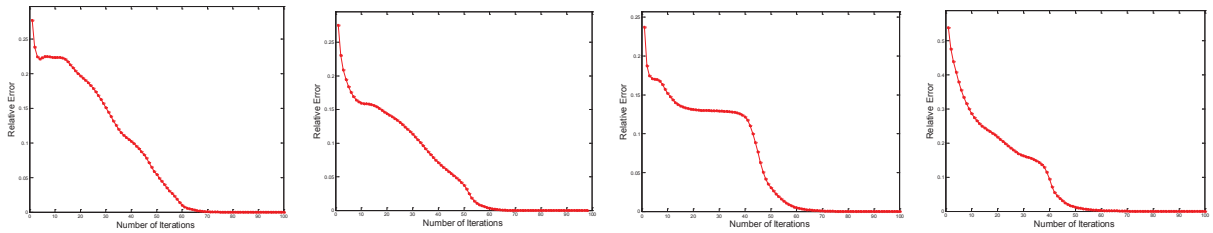
Based on the numerical experimental results shown in Table I-VII, the following observations are reached.

First, the proposed BDLRR method gains the best performances in comparison with all of the compared state-of-the-art methods for recognition tasks on eight data sets. This demonstrates that the proposed method enables to effectively learn a discriminative and robust representation from data. Moreover, we can conclude that it is beneficial to image recognition when transferring the original image features to the discriminative BDLRR based on the pivot features, i.e. training features, in a semi-supervised manner.

Second, the proposed BDLRR is significantly superior to some related methods, i.e., RPCA, LRSI, LatLRR, LRLR, LRRR, and CBDS, which demonstrates the benefit and necessity of imposing the discriminative structure on LRR and leveraging the l_{21}-norm to overcome noise and outliers. With the purpose of constructing the discriminative structure, the margin between block-diagonal and off-block-diagonal components is enlarged such that the incoherent data representation is boosted and the coherent data representation is enhanced simultaneously. Furthermore, it is also revealed that jointly learning the training and test representations can greatly improve the performance of recognition tasks.

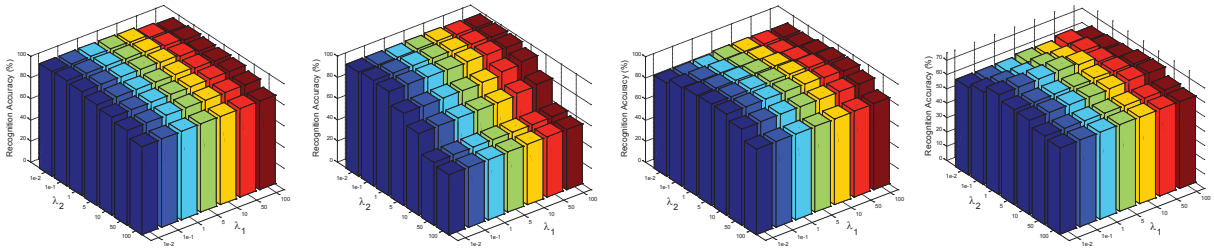
Third, an interesting scenario in the experimental results is that there does not exist the absolute best algorithm among all the compared methodologies on eight datasets, because the performance relative to each other is mixed and inconsistent for different recognition applications. However, our BDLRR method outperforms all other methods on these low-resolution, limited training sample experiments. The main reason may be that our method intrinsically inherit the superiorities of sparse, low-rank, structured and elastic-net representation learning techniques. Specifically, the low-rank regularization, on the one hand, can effectively mine the underlying structure of data correlation, and the global latent structure of the data matrix is uncovered. One the other hand, the sparsity characteristic mainly focuses on finding the nearest subspace of data. However, they neglect the fact that constructing block-diagonal representation is the most straightforward fashion to explore the intrinsic structure of data and elucidate the nearest subspace of data points. For instance, we use the first 10 classes of test samples from the Extended YaleB dataset to visually present the representation results of SRC and BDLRR, which are shown in Fig. 1. Images of the first ten subjects from the Extended YaleB dataset are used for experiments. We randomly select 35 images per subject as training samples and treat the rest of images as test samples. All images are rearranged by Assumption 1. From Fig. 1, we can see that our method can more clearly illustrate the nearest subspace (block-diagonal structure) of test samples, leading to better recognition results.

Fourth, our BDLRR method consistently outperforms CBDS and LatLRR on all datasets. For CBDS, it locally enforces the class-wise diagonal structure on the low-rank criterion, whereas our BDLRR method globally imposes the block-diagonal constraint on the low-rank criterion by directly



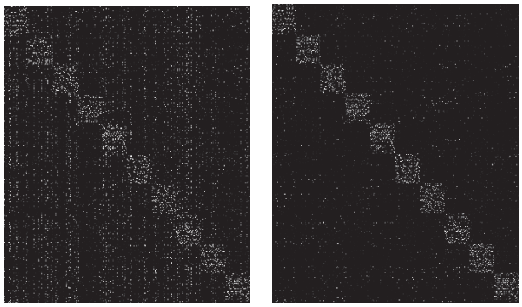
(a) Extended YaleB ($\#Tr$ 35) (b) AR ($\#Tr$ 20) (c) USPS ($\#Tr$ 90) (d) Char74K-15 ($\#Tr$ 15)

Fig. 2: Convergence curves of the proposed method on different databases. (a)-(d) are the convergence curves on the Extended YaleB, AR, USPS and Char74K-15 datasets, respectively.



(a) Extended YaleB ($\#Tr$ 35) (b) AR ($\#Tr$ 20) (c) USPS ($\#Tr$ 90) (d) Char74K-15 ($\#Tr$ 15)

Fig. 3: The performance evaluation (%) of BDLRR versus parameters λ_1 and λ_2 on (a) Extended YaleB (b) AR (c) USPS and (d) Char74K-15 datasets.



(a) SRC (b) BDLRR

Fig. 1: Data representation comparisons on the Extended YaleB dataset. (a) and (b) are data representations of the test set obtained using SRC and BDLRR, respectively. The data representation values are multiplied by 5.

minimizing the off-block-diagonal components. Moreover, our method further enhances the within block-diagonal structure to be more compact by increasing the coherent intra-class representation. For LatLRR, it extracts salient features from observations for recognition, and the unfavorable performance may result from the certain truth that too many image details are lost.

Finally, we can see that the proposed method can overcome the difficulty of noise-induced data uncertainty in face recognition, such as occlusion, disguise, severe illumination changes and expression variations. Moreover, our method works well on the challenging natural scene character recognition task, which further indicates that BDLRR is robust to the obstacles and difficulties of scene text images, such as complex background, low-resolution, occlusion, blurring, and the changes of text size or font.

F. Convergence and Parameter Sensitiveness Analysis

In this section, the convergence property of BDLRR and the influence of parameter selection are empirically studied on four data sets, i.e. the extended YaleB, AR, USPS and Char74K-15 datasets.

1) Convergence study: The theoretical convergence proof of the proposed optimization method is analyzed in Section IV-C. It is demonstrated that BDLRR can converge to a stationary point under mild conditions. Now we experimentally validate its convergence on different datasets to demonstrate its efficient convergence. The convergence curves on four datasets are presented in Fig. 2, where $\#Tr$ denotes the number of training samples per subject selected for experiments. Similar to [8], the relative error (i.e. $||X - X_{tr}Z - E||_F/||X||_F$) is employed to show its convergence. We can see that the relative error generally decreases with the increasing number of iterations. More specifically, although the relative error exists a little vibration at the first fifteen iterations on the Extended YaleB data set, the overall values of the relative error change only slightly after 60 iterations for these four datasets shown in Fig. 2, which demonstrates that the proposed optimization algorithm holds the convergent nature.

2) Parameter Sensitiveness: In the proposed optimization problem (8), there are three parameters to be tuned. In our experiments, it is observed that the performance of BDLRR is not sensitive to λ_3 when it is in the range of [10,25], which is also an empirical setting. To test how the remaining parameters λ_1 and λ_2 influence the performance of BDLRR, we perform extensive experiments to validate their robustness. Similar to convergence validations, we still use the Extended YaleB, AR, USPS and Char74K-15 datasets for evaluation. Fig. 3 presents the performance variations with respect to parameters λ_1 and λ_2. We can see that the performance of our BDLRR method

is generally insensitive to varying values of λ_1 and λ_2. More specifically, the performance is promising when parameter λ_1 is not too large or small, which indicates the necessity of boosting the extra-class data incoherent representation. Moreover, for parameter λ_2, it is easy to see that it should be small, and the best results are usually achieved when the value is smaller than 1, yet bigger than 0.01. The possible reason of a smaller λ_2 may be that the Euclidean distance metric used in our experiments is too simple to perfectly measure the similarity of samples. However, we have achieved very impressive experimental results, even with a simple distance metric. In a word, our BDLRR method is robust to parameter changes in most cases.

G. Limitation

From the objective function of our BDLRR method, i.e. Eqn. (8), we can see that the proposed model is a semi-supervised representation learning model and concurrently learns both block-diagonal representations of training and test samples, which indicates that the test samples and the label of training samples are both given in the learning process. However, in some cases we cannot get access of test data at the training stage, which may limit the generalization of our model. To this end, we extend our BDLRR method to address the out-of-sample problem in section IV-E to circumvent this problem. In this way, our results are somewhat subject to the learning capability of algorithms in handling the out-of-sample cases. Fortunately, these methods have been examined to effectively formulate favorable representations of new instances. Moreover, the learned data representations of training samples are reasonably block-diagonal in the training stage, which in turn guarantees the satisfactory recognition results.

VI. Conclusion

In this paper, we have proposed a novel discriminative block-diagonal representation learning model, i.e. BDLRR, for robust image recognition. BDLRR focuses on learning a discriminative data representation by imposing an effective structure in a low-rank representation framework, where the extra-class incoherent representation and intra-class coherent representation are simultaneously enhanced. The proposed method incorporates the learned BDLRR into the semi-supervised model to collaboratively optimize the training data representation and test data representation, and then an efficient linear classifier is obtained to perform final robust image recognition. Moreover, an effective optimization algorithm is developed to solve the resulting optimization problem. Last but not least, the proposed method was evaluated on eight publicly available benchmark datasets for three different recognition tasks. Extensive experimental results have demonstrated that the proposed BDLRR method is superior to state-of-the-art methods.

VII. Acknowledgement

We would like to thank Prof. Chenglin Liu and Dr. Xuyao Zhang for many inspiring discussions and constructive suggestions. Moreover, we also thank the editor, an associate editor, and referees for helpful comments and suggestions which greatly improved this paper.

References

[1] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798-1828, 2013.

[2] Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang, "A survey of sparse representation: algorithms and applications", *IEEE Access*, vol. 3, pp. 490-530, 2015.

[3] 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, 2009.

[4] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification", *in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3360-3367, 2010.

[5] S. Li and Y. Fu, "Learning Balanced and Unbalanced Graphs via Low-Rank Coding", *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 5, pp. 1274-1287, 2015.

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

[7] G. Liu and S. Yan, "Latent low-rank representation for subspace segmentation and feature extraction", *in Proceeding of IEEE International Conference on Computer Vision*, pp. 1615-1622, 2011.

[8] S. Li and Y. Fu, "Learning Robust and Discriminative Subspace With Low-Rank Constraints", *IEEE Trans. Neural Netw. Learn. Sys.*, 2016, DOI:10.1109/TNNLS.2015.2464090.

[9] S. Xiao, M. Tan, D. Xu, Z. Dong, "Robust Kernel Low-Rank Representation", *IEEE Trans. Neural Netw. Learn. Sys.*, vol. 27, no.11, pp. 2268-2281, 2016.

[10] Y. Zhang, Z. Jiang, and L. Davis, "Learning structured low-rank representations for image classification", *in Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 676-683, 2013.

[11] Z. Li, Z. Lai, Y. Xu, et al. A Locality-Constrained and Label Embedding Dictionary Learning Algorithm for Image Classification, *IEEE Trans. Neural Netw. Learn. Sys.*, vol. 28, no. 2, pp. 278-293, 2017.

[12] Y. Xu, Z. Zhong, J. Yang, J. You, D. Zhang, "A New Discriminative Sparse Representation Method for Robust Face Recognition via L2 Regularization," *IEEE Trans. Neural Netw. Learn. Sys.*, DOI:10.1109/TNNLS.2016.2580572, 2017.

[13] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint l_{21}-norms minimization", *in Proceedings Advances in Neural Information Processing Systems*, pp. 1813-1821, 2010.

[14] 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 Sys. Video Technol.*, vol. 21, no. 9, pp. 1255-1262, 2011.

[15] C. Lu, H. Min, J. Gui, L. Zhu, and Y. Lei, "Face recognition via weighted sparse representation", *J. Vis. Commun. Image Represent.*, vol. 24, no. 2, pp. 111-116, 2013.

[16] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" *in Proceedings of IEEE International Conference on Computer Vision*, pp. 471-478, 2011.

[17] Z. Zhang, L. Wang, et al. "Noise modeling and representation based classification methods for face recognition," Neurocomputing, vol. 148, pp. 420-429, 2015.

[18] I. Naseem, R. Togneri, and M. Bennamoun, "Linear regression for face recognition", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 2106-2112, 2010.

[19] Z. Zhang, Z. Lai, Y. Xu, L. Shao, J. Wu, G. Xie, Discriminative Elastic-Net Regularized Linear Regression, *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1466-1481, 2017.

[20] X. Cai, C. Ding, F. Nie, and H. Huang, "On the equivalent of low-rank linear regressions and linear discriminant analysis based regressions", *in Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1124-1132, 2013.

[21] E. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, pp. 11, 2011.

[22] C. Wei, C. Chen, and Y. Wang, "Robust Face Recognition With Structurally Incoherent Low-Rank Matrix Decomposition", *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3294-3307, 2014.

[23] Y. Li, J. Liu, Z. Li, Y. Zhang, H. Lu, and S. ma, "Learning low-rank representations with classwise block-diagonal structure for robust face recognition", *Twenty-Eighth AAAI Conference on Artificial Intelligence*, pp. 2810-2816, 2014.

[24] S. Li, M. Shao, and Y. Fu. "Multi-view low-rank analysis for outlier detection", *SIAM International Conference on Data Mining*, pp. 748-756, 2015.

[25] L. Zhuang, S. Gao, J. Tang, et al. "Constructing a nonnegative low-rank and sparse graph with data-adaptive features," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3717-3728, 2015.

[26] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765-2781, 2013.

[27] Z. Lin, M. Chen, and Y. Ma. "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices", arXiv preprint arXiv:1009.5055, 2010.

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

[29] R. Glowinski and P. Le Tallec, "Augmented Lagrangian and Operator-splitting Methods in Nonlinear Mechanics," *SIAM*, 1989.

[30] E. Esser, "Applications of Lagrangian-based alternating direction methods and connections to split Bregman," *CAM report*, vol. 9, pp. 31, 2009.

[31] J. Eckstein and D. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming*, vol. 55, 1992.

[32] Dimitri P. Bertsekas, "Convex Optimization Theory," Belmont: Athena Scientific, 2009.

[33] L. N. Trefethen, D. Bau, "Numerical linear algebra," SIAM, 1997.

[34] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Sys. Technol.*, vol. 2, no. 3, pp. 27, 2011.

[35] A. Georghiades, P. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Machin. Intell.*, vol. 23, no. 6, pp. 643-660, 2001.

[36] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression (PIE) database", *Proceedings of the fifth IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 46-51, 2002.

[37] A. M. Martinez and R. Benavente. "The AR Face Database," *CVC Technical Report*, no. 24, June 1998.

[38] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments", *Univ. Massachusetts*, Amherst, MA, USA, Tech. Rep. 07-49, 2007.

[39] Z. Jiang, Z. Lin, and L. Davis. "Label consistent K-SVD: Learning a discriminative dictionary for recognition", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2651-2664, 2013.

[40] S. J. Wang, J. Yang, M. F. Sun, X.-J. Peng, M.-M. Sun, and C.-G. Zhou, "Sparse tensor discriminant color space for face verification", *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, pp. 876-888, 2012.

[41] J. Hull, "A database for handwritten text recognition research", *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 16, no. 5, pp. 550-554, 1994.

[42] T. E. de Campos, B. R. Babu, and M. Varma. "Character Recognition in Natural Images", *In Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP)*, Lisboa, Portugal, vol. 2, pp. 273-280, 2009.

[43] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition", *in Proceedings of IEEE International Conference on Computer Vision*, Barcelona, Spain, pp. 1457-1464, 2011.

[44] Z. Zhang, Y. Xu, and C.-L. Liu, "Natural Scene Character Recognition Using Robust PCA and Sparse Representation", *in Proceedings of the IAPR International Workshop on Document Analysis Systems (DAS)*, 2016.

[45] S. Tian, S. Lu, B. Su, and C.L. Tan, "Scene text recognition using co-occurrence of histogram of oriented gradients", *in Proceeding of 12th International Conference on Document Analysis and Recognition*, Washington, D.C., USA, pp. 912-916, 2013.

[46] B. Su, S. Lu, S. Tian, and C.L. Tan, "Character Recognition in Natural Scenes using Convolutional Co-occurrence HOG", *in Proceeding of 22nd International Conference of Pattern Recognition*, Istanbul, Turkey, pp. 2926-2931, 2014.

[47] Z. R. Tan, S. Tian, and C.L. Tan, "Using pyramid of histogram of oriented gradients on natural scene text recognition", *in Proceeding of 2014 International Conference of Image Processing*, Paris, France, pp. 2629-2633, 2014.

[48] C. Y. Lee, A. Bhardwaj, W. Di, V. Jagadeesh, and R. Piramuthu, "Region-based discriminative feature pooling for scene text recognition", *in Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, pp. 4050-4057, 2014.

[49] T. Q. Phan, P. Shivakumara, S. Tian, and C. L. Tan. "Recognizing text with perspective distortion in natural scenes", *in Proceedings of 14th IEEE International Conference on Computer Vision*, Sydney, Australia, pp. 569-576, 2013.

[50] C. Yi, X. Yang, and Y. Tian, "Feature representations for scene text character recognition: A comparative study", *in Proceeding of 12th International Conference on Document Analysis and Recognition*, Washington, DC, USA, pp. 907-911, 2013.

[51] C. Yi and Y. Tian, "Scene text recognition in mobile applications by character descriptor and structure configuration", *IEEE Trans. Image Processing*, vol. 23(7), pp. 2972-2982, 2014.

[52] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories", *in proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2169-2178, 2006.

[53] X. Lian, Z. Li, B. Lu, and L. Zhang, "Max-margin dictionary learning for multiclass image categorization", *in proceddings of European Conference on Computer Vision*, pp. 157-170, 2010.

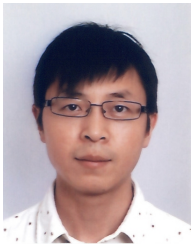
[54] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification", *in proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1794-1801, 2009.

[55] Y. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition", *in proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2559-2566, 2010.

[56] S. Gao, I. Tsang, L. Chia, and P. Zhao, "Local features are not lonelyCLaplacian sparse coding for image classification," *in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3555-3561, 2010.

Zheng Zhang received the B.S degree from Henan University of Science and Technology and M.S degree from Shenzhen Graduate School, Harbin Institute of Technology (HIT) in 2012 and 2014, respectively. Currently, he is pursuing the Ph.D. degree in computer science and technology at Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China. His current research interests include pattern recognition, machine learning and computer vision.

Yong Xu (M'06-SM'15) was born in Sichuan, China, in 1972. He received his B.S. degree and M.S. degree at Air Force Institute of Meteorology (China) in 1994 and 1997, respectively. He received the Ph.D. degree in Pattern recognition and Intelligence System at the Nanjing University of Science and Technology (NUST) in 2005. Now, he works at Shenzhen Graduate School, Harbin Institute of Technology. His current interests include pattern recognition, biometrics, machine learning and video analysis.



Ling Shao (M'09-SM'10) is a professor with the School of Computing Sciences at the University of East Anglia, Norwich, UK. Previously, he was a professor (2014-2016) with Northumbria University, a senior lecturer (2009-2014) with the University of Sheffield and a senior scientist (2005- 2009) with Philips Research, The Netherlands. His research interests include computer vision, image/video processing and machine learning. He is an associate editor of IEEE Transactions on Image Processing, IEEE Transactions on Neural Networks and Learning Systems and several other journals. He is a Fellow of the British Computer Society and the Institution of Engineering and Technology. He is a senior member of the IEEE.

Jian Yang received the B.S. degree in mathematics from the Xuzhou Normal University in 1995. He received the M.S. degree in applied mathematics from the Changsha Railway University in 1998 and the Ph.D. degree from the Nanjing University of Science and Technology (NUST), on the subject of pattern recognition and intelligence systems in 2002. In 2003, he was a postdoctoral researcher at the University of Zaragoza. From 2004 to 2006, he was a Postdoctoral Fellow at Biometrics Centre of Hong Kong Polytechnic University. From 2006 to 2007, he was a Postdoctoral Fellow at Department of Computer Science of New Jersey Institute of Technology. Now, he is a professor in the School of Computer Science and Technology of NUST. He is the author of more than 80 scientific papers in pattern recognition and computer vision. His journal papers have been cited more than 1600 times in the ISI Web of Science, and 2800 times in the Web of Scholar Google. His research interests include pattern recognition, computer vision and machine learning. Currently, he is an associate editor of Pattern Recognition Letters and IEEE TRANSACTION ON NEURAL NETWORKS AND LEARNING SYSTEMS, respectively.