Block-Diagonal Sparse Representation by Learning a Linear Combination Dictionary for Recognition

Xinglin Piao, Yongli Hu, *Member, IEEE,* Yanfeng Sun, *Member, IEEE,* Junbin Gao, Baocai Yin, *Member, IEEE*

***Abstract*—In a sparse representation based recognition scheme, it is critical to learn a desired dictionary, aiming both good representational power and discriminative performance. In this paper, we propose a new dictionary learning model for recognition applications, in which three strategies are adopted to achieve these two objectives simultaneously. First, a block-diagonal constraint is introduced into the model to eliminate the correlation between classes and enhance the discriminative performance. Second, a low-rank term is adopted to model the coherence within classes for refining the sparse representation of each class. Finally, instead of using the conventional over-complete dictionary, a specific dictionary constructed from the linear combination of the training samples is proposed to enhance the representational power of the dictionary and to improve the robustness of the sparse representation model. The proposed method is tested on several public datasets. The experimental results show the method outperforms most state-of-the-art methods.**

***Index Terms*—Dictionary learning, sparse representation, recognition application.**

I. Introduction

In the past years, the sparse representation has achieved great success in many applications, such as face recognition [1], [2], [3], image classification [4], [5], [6], and human action recognition [7], [8], [9]. The main idea of the sparse representation is based on the fact that many natural signals could be represented or encoded by a few atoms of an over-complete dictionary [10]. Many dictionary learning methods have been proposed to learn an expressive dictionary for the problems at hand. Among them, an easy and direct method is to use training samples themselves as dictionary atoms. This simple strategy, based on the data self expressive property, is widely adopted by many sparse representation based recognition methods, such as the Sparse Representation Classification (SRC) method [1] and its variations like [4]. The data self expression is theoretically guaranteed by the subspace theory, which assumes that many signals constitute linear subspaces and the samples derived from a subspace can be approximately represented by other samples (in most cases, the training samples) from the same subspace [1]. Although using data self expression dictionaries shows good performance in practice, the success in applications generally depends on the quality of training data, as this type of methods is sensitive to noise and outliers. On the other hand, to obtain good representation results, the number of training samples in a dictionary should be large enough to capture variances of signals, which increases the computational complexity of sparse coding. Therefore, instead of using the training samples themselves, learning based methods construct dictionaries by optimalizing some sparse representation criteria for training samples. This type of learning algorithms includes the Method of Optimal Directions (MOD) [11] and K-SVD [12]. Following these two classic methods, many sparse representation based methods learn their specific dictionaries with many successful applications. For example, Yang *et al.* [13] proposed an SRC based MetaFace Learning (MFL) method for face recognition.

However, the unsupervised strategy in classic dictionary learning methods, such as MOD and K-SVD, does not utilize label or discriminative information of data, which is valuable for recognition applications. So many researchers try to develop supervised dictionary learning methods to improve classification or recognition performance by incorporating discriminative information of training data. For example, Zhang *et al.* [14] proposed a Discriminative K-SVD (D-KSVD) dictionary learning method, in which the discriminative information of training data was represented as a classification error term with a simple linear classifier. Jiang *et al.* [15], [16] proposed a Label Consistent K-SVD method (LC-KSVD) for recognition applications, in which a sparse recognition error term was designed to model the consistency of class labels and its recognition results for training samples. The Fisher criterion is an effective penalty to decrease the within-class scatterness and increase the between-class scatterness. Yang *et al.* [17] added the Fisher criterion constraint into the dictionary learning model to form a Fisher Discrimination Dictionary Learning method (FDDL).

Except for the label information of training samples, the structure of the sparse coefficient matrix, the incoherence between sub-dictionaries or the coherence within sub-dictionaries, the correlation within the sparse vectors of the same class and other properties are explored to improve the discriminative power of the sparse representation model for recognition [16], [18], [22]. For the sparse representation based recognition, the ideal coefficient matrix of training samples under dictionaries should be block-diagonal, i.e. the coefficients of a sample on its own sub-dictionary are nonzero and on the sub-dictionaries corresponding to different class are zero. This desired structural coefficient matrix will bring best discriminative results. In some methods, one adds some block-diagonal

Xinglin Piao, Yongli Hu and Yanfeng Sun are with Beijing Municipal Key Lab of Multimedia and Intelligent Software Technology, College of Metropolitan Transportation, Beijing University of Technology, Beijing 100124, China. e-mail: piaoxinglin1987@gmail.com,{huyongli,yfsun}@bjut.edu.cn.

Junbin Gao is with the School of Computing and Mathematics, Charles Sturt University, Bathurst, NSW 2795, Australia. e-mail: jbgao@csu.edu.au.

Baocai Yin is with the School of Software Technology at Dalian University of Technology, Dalian 116620, China; and with Beijing Municipal Key Lab of Multimedia and Intelligent Software Technology at Beijing University of Technology, Beijing 100124, China. e-mail: ybc@bjut.edu.cn.

Manuscript received March XX, 2015; revised XXX XXX, 2015.

constraint on the coefficient matrix in a dictionary learning model. For example, in LC-KSVD, one enforces a coefficient matrix to be approximately 0-1 block-diagonal [16]. For the similar purpose, in other methods one also adds constraints on the dictionaries or sub-dictionaries instead of on the coefficient matrix. For example, the Dictionary Learning method with Structure Incoherence (DLSI) [18] uses an incoherence term to decrease the correlation between sub-dictionaries. To further enhance the incoherence between different sub-dictionaries, Kong *et al.* [19] proposed a Dictionary Learning method to learn a common pattern pool (the COmmonality) and class-specific dictionaries (the PARticularity) for classification, namely DL-COPAR. In this method, a common dictionary is learned to separate the sharing information of different sub-dictionaries, and class-specific sub-dictionaries are learned to produce a clear block-diagonal sparse coefficient matrix. Along this direction, the Discriminative Group Sparse Dictionary Learning method (DGSDL) [20] assigns different weights to the common dictionary and the class-specific sub-dictionaries, respectively, to decrease the interference of the common dictionary for classification and obtain more non-zero coefficients with respect to the class-specific sub-dictionaries. Generally, there exists high correlation within the samples of same class. As a result, it is a belief that there is a correlation within their sparse vectors. So this characterization can be utilized to improve the sparse representation for recognition. An intuitive form to represent this property is the low rank constraint on the sparse coefficient matrix. Zhang *et al.* [22] proposed a joint image recognition model, which learns a low-rank and structured sparse representation for dictionary learning. Li *et al.* [23] proposed a semi-supervised model to learn Low-Rank representations with Classwise Block-Diagonal Structure (LR-CBDS) for dictionary learning. These works have shown that the low-rank constraint could not only model the coherence within each class but also reduce the noise that exists in training data.

The current research has demonstrated that enforcing certain specific structures in the sparse coefficient matrix in the dictionary learning procedure does improve the discriminative power of the dictionary and recognition performance. In many existing methods one aims to construct a block-diagonal sparse coefficient matrix by incorporating certain diagonal induced constraints on the coefficient matrix. In general, the resulting matrix is far away from the desired block-diagonal ones. So it is worth to explore a block-diagonal structure for sparse coefficient matrices in a more straight fashion for recognition applications. Whilst the block-diagonal sparse coefficient matrix is an attempt to enhance the incoherence between classes, how to represent the correlation or coherence within each class is also important. In literature, the low rank constraint is regarded as a good way to represent the correlation of the samples of one class [24]. In most of existing methods, the low rank constraint is simply imposed on the whole sparse coefficient matrix instead of individual classes [22]. In this paper, we propose a new dictionary learning method based on block-diagonal sparse representation for recognition applications, by simultaneously incorporating the block-diagonal structure for the sparse coefficient matrix and applying low rank constraint to maintain the coherence within each class.

Incorporating the block-diagonal and low rank constraints on the sparse coefficient matrix may improve the discriminative power of the sparse representation. Furthermore, in this paper, we explore the structure of dictionaries to be learned. In most of the existing dictionary learning methods, little attention has been paid to the structure of dictionaries. Either dictionary atom are freely updated only according to the reconstruction error between the training data and its sparse representation, or one uses a fixed dictionary with training data themselves based on the data self-representative property. In the first case, the learned dictionary will be dramatically influenced by the training samples. Especially, if there are outlier or noise in the training data, the dictionary will be misdirected by these samples. Therefore, instead of using the conventional dictionary in the sparse representation models, many researches try to construct specific dictionaries for different applications, such as the double-sparsity model [25], and the non-linear kernel dictionary [26]. In this paper, we take a strategy between the conventional dictionary learning and data self-representation by using a form of linear combination of the training samples. This strategy has been proved to be an optimal solution to the dictionary learning problem in the conventional sparse representation model [26]. Additionally, the linear combination dictionary will enhance the robustness of the sparse representation and the recognition performance in complex scenarios. This point is demonstrated by our experiments on several public datasets.

In summary, the contributions and novelties of this paper are four-fold,

- A strict block-diagonal sparse representation model is proposed for the dictionary learning to eliminate the correlation between classes and achieve better discriminative performance for recognition applications.
- The coherence of the sparse representation within each class is represented by a low-rank constraint, refining the sparse representation of each class for recognition.
- The linear combination of the training samples is utilized, which will enhance the representational power of the dictionary and improve the robustness of the sparse representation model.
- An efficient algorithm is constructed to solve the proposed sparse optimization model with block-diagonal and low rank constraints.

The paper is organized as follows. Section II introduces the basic sparse representation and dictionary learning method for recognition applications. In Section III, the new block-diagonal sparse representation based dictionary learning method is proposed in detail. Section IV gives the solution to the optimalization problem of the proposed sparse representation model. In Section V, recognition experiments are conducted on several public datasets to assess the performance of the proposed method. Section VI concludes the paper and discusses the future work.

II. The basic sparse representation based Dictionary learning for recognition

Given a training data set $\mathbf{Y} = [y_1, y_2, ..., y_N] \in \mathbb{R}^{n\times N}$, where n denotes the dimension of sample data and N denotes the number of training samples. We can learn a dictionary $\mathbf{D}$ and the sparse representation $\mathbf{X} = [x_1, x_2, ..., x_N]$ of $\mathbf{Y}$ by solving the following model:

$$\begin{aligned} \min_{\mathbf{D},\mathbf{X}} \quad & \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2, \\ \text{s.t.} \quad & \forall i, \quad \|x_i\|_0 \le T. \end{aligned} \tag{1}$$

where $\|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2$ denotes the error between the training data and its sparse representation on the dictionary $\mathbf{D} = [d_1, d_2, ..., d_K] \in \mathbb{R}^{n\times K}$ with $Frobenius$ norm, K the number of dictionary atoms, $\mathbf{X} \in \mathbb{R}^{K\times N}$ the sparse coefficient matrix, $\|\cdot\|_0$ the l_0 norm, and T the sparsity level.

To construct an effective dictionary $\mathbf{D}$ and obtain the sparse coefficient matrix $\mathbf{X}$, a classical algorithm namely K-SVD [12] is proposed to solve this problem by an iterating scheme. In this algorithm, the Orthogonal Matching Pursuit algorithm (OMP) [27], [28] is used for computing the coefficient matrix $\mathbf{X}$ with a fixed dictionary $\mathbf{D}$, while the dictionary $\mathbf{D}$ can be updated atom-by-atom by minimizing the energy term $\|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2$ in (1) with the fixed $\mathbf{X}$. As an efficient dictionary learning method, K-SVD is widely used in many applications, such as image reconstruction and denoising [29], [30], [31].

In general, the non-convex l_0 norm induces certain optimization difficulty, hence one usually uses the surrogate l_1 norm instead. So a new dictionary learning model with l_1 sparse norm is formulated as the following equation:

$$\min_{\mathbf{D},\mathbf{X}} \quad \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 + \tau\|\mathbf{X}\|_1. \tag{2}$$

where τ is a balance parameter trading-off between the construction error and the sparsity, and $\|\mathbf{X}\|_1 = \sum_{i,j}|x_{ij}|$ is the l_1 norm of the coefficient matrix $\mathbf{X}$. While using an alternative way to solve problem (2), the coefficient matrix $\mathbf{X}$ can be solved by a sparse coding algorithm such as the feature-sign search algorithm [33], the learned iterative shrinkage-thresholding algorithm (LISTA) [34] and the alternating direction method of multipliers (ADMM) [35], while the dictionary $\mathbf{D}$ can be solved by adopting the dictionary updating procedure of K-SVD algorithm.

Based on the above sparse representation, there is a basic framework for recognition application, known as the sparse representation based recognition. In this framework, the training dataset with label information can be rewritten as $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_C]$ according the class label, where C is the number of all classes and $\mathbf{Y}_i \in \mathbb{R}^{n\times N_i}$ denotes the subset of training dataset for the i-th class with N_i samples such that $\sum_i^C N_i = N$. Thus we can learn a dictionary for every class of training data $\mathbf{Y}_i$ by the above sparse representation model, denoted by $\mathbf{D}_i, i = 1, ..., C$, which are regarded as the sub-dictionaries and integrated to form the dictionary $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, ..., \mathbf{D}_C]$ w.r.t. the training data $\mathbf{Y}$. To this end, for a test sample $\tilde{y}$, we can calculate the sparse coefficient $\tilde{x}$ on the dictionary $\mathbf{D}$ using a sparse coding method by the following formula,

$$\tilde{x} = \arg\min_x \|\tilde{y} - \mathbf{D}x\|_2^2 + \tau\|x\|_1. \tag{3}$$

From this sparse representation, the recognition is finally realized by selecting the class with the minimal reconstruction error as follows,

$$identity(\tilde{y}) = \arg\min_i \|\tilde{y} - \mathbf{D}_i\tilde{x}_i\|_2^2. \tag{4}$$

where $\tilde{x}_i$ is the sparse coefficients of $\tilde{x}$ corresponding to the i-th sub-dictionary $\mathbf{D}_i$.

III. Block-Diagonal Sparse Representation by Learning a Linear Combination Dictionary for Recognition

From the above basic framework of the sparse representation based recognition, with the label information being available, the training data $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_C]$ can be sparse represented on the integrated dictionary $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, ..., \mathbf{D}_C]$, i.e. $\mathbf{Y} = \mathbf{D}\mathbf{X}$. However, the sparse representation does not consider the relation of different classes as the sub-dictionaries are learned respectively. So we further investigate the intrinsic structure underlying the sub-dictionaries and the sparse representations of training data to form a new dictionary learning method for recognition. Firstly, we examine the structure of the sparse coefficients matrix $\mathbf{X}$. According to the class label, the sparse coefficients matrix $\mathbf{X}$ can be further rewritten as following form

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_{11} & \mathbf{X}_{12} & \cdots & \mathbf{X}_{1C} \\ \mathbf{X}_{21} & \mathbf{X}_{22} & \cdots & \mathbf{X}_{2C} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}_{C1} & \mathbf{X}_{C2} & \cdots & \mathbf{X}_{CC} \end{bmatrix}. \tag{5}$$

where $\mathbf{X}_{ij} \in \mathbb{R}^{K_i\times N_j}, i, j \in \{1, 2, ..., C\}$ is the sub-coefficient matrix, representing the sparse coefficient of the samples in the j-th labelled class $\mathbf{Y}_j$ over the i-th sub-dictionary $\mathbf{D}_i$. Thus $\mathbf{Y}_j$ can be rewritten as:

$$\mathbf{Y}_j = \mathbf{D}_1\mathbf{X}_{1j} + \mathbf{D}_2\mathbf{X}_{2j} + ... + \mathbf{D}_C\mathbf{X}_{Cj}. \tag{6}$$

It is natural to assume that the j-th labelled class samples $\mathbf{Y}_j$ can be strictly represented by the atoms of the j-th sub-dictionary $\mathbf{D}_j$, which means that $\mathbf{X}_{ij}$ is a zero sub-matrix for $i \neq j$. In other words, the ideal coefficient matrix $\mathbf{X}$ is block-diagonal in the following form:

$$\mathbf{diag}(\mathbf{X}_{11}, \mathbf{X}_{22}, ..., \mathbf{X}_{CC}) = \begin{bmatrix} \mathbf{X}_{11} & 0 & \cdots & 0 \\ 0 & \mathbf{X}_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{X}_{CC} \end{bmatrix}. \tag{7}$$

Let $\mathbf{X} = \mathbf{diag}(\mathbf{X}_{11}, \mathbf{X}_{22}, ..., \mathbf{X}_{CC})$. With this block-diagonal parametrization of the sparse representation matrix, we can build a block-diagonal sparse representation model as following:

$$\begin{aligned} \min_{\mathbf{D},\mathbf{X}} \quad & \tau\|\mathbf{X}\|_1 + \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2, \\ \text{s.t.} \quad & \mathbf{X} = \mathbf{diag}(\mathbf{X}_{11}, \mathbf{X}_{22}, ..., \mathbf{X}_{CC}). \end{aligned} \tag{8}$$

The block-diagonal sparse representation model is proposed to eliminate correlation between different classes and enhance the discriminate power of the model. However, for each class, there exists high correlation within its training samples and their corresponding sparse coefficient vectors. Thus, to further model the coherence within each class, we adopt the low-rank constraint to capture this property within each individual class. Concretely, we add low rank constraints on each diagonal submatrix of the coefficient matrix and get the following revised sparse representation model from the model in (8).

$$\begin{aligned} \min_{\mathbf{D},\mathbf{X}} \quad & \tau\|\mathbf{X}\|_1 + \lambda\sum_{i=1}^{C}\text{rank}(\mathbf{X}_{ii}) + \|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2, \\ \text{s.t.} \quad & \mathbf{X} = \mathbf{diag}(\mathbf{X}_{11},\mathbf{X}_{22},...,\mathbf{X}_{CC}). \end{aligned} \tag{9}$$

where λ is a weight for the low-rank term.

In general, the rank minimization problem in (9) is a NP-hard problem [36]. So an alterative way to solve the problem in (9) is to replace the rank function with the so-called matrix nuclear norm $\|\cdot\|_*$, which is defined as the sum of singular values of the matrix. Thus the model in (9) can be converted into a convex optimization problem as follows:

$$\begin{aligned} \min_{\mathbf{D},\mathbf{X}} \quad & \tau\|\mathbf{X}\|_1 + \lambda\sum_{i=1}^{C}\|\mathbf{X}_{ii}\|_* + \|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2, \\ \text{s.t.} \quad & \mathbf{X} = \mathbf{diag}(\mathbf{X}_{11},\mathbf{X}_{22},...,\mathbf{X}_{CC}). \end{aligned} \tag{10}$$

To further get an efficient dictionary for recognition, instead of using the dictionary in the conventional sparse representation model, we adopt a specific dictionary constructed by the linear combination of the training samples, i.e. $\mathbf{D} = \mathbf{Y}\mathbf{W}$, where $\mathbf{W} \in \mathbb{R}^{N\times K}$ is the combination matrix. The motivation of using this linear combination dictionary lies in two folds. First, in theory, this type of dictionary has been proved an optimal solution to the dictionary learning problem in the conventional sparse representation model, see Proposition 1 in [26]. Second, the linear combination operation confines the variety of the dictionary atoms in the space spanned by the training samples. Thus the influence of few outliers or noise in the training data will be reduced, and this will enhance the robustness of the dictionary learning and the sparse representation. Under this specific dictionary, we get our revised sparse model as follows:

$$\begin{aligned} \min_{\mathbf{W},\mathbf{X}} \quad & \tau\|\mathbf{X}\|_1 + \lambda\sum_{i=1}^{C}\|\mathbf{X}_{ii}\|_* + \|\mathbf{Y}-\mathbf{Y}\mathbf{W}\mathbf{X}\|_F^2, \\ \text{s.t.} \quad & \mathbf{X} = \mathbf{diag}(\mathbf{X}_{11},\mathbf{X}_{22},...,\mathbf{X}_{CC}). \end{aligned} \tag{11}$$

This is a complicated optimization problem with l_1 norm and nuclear norm with respect to matrix variable $\mathbf{X}$. In order to get a stable feasible solution, we further add a regularizer on $\mathbf{X}$ and obtain the following sparse model,

$$\begin{aligned} \min_{\mathbf{W},\mathbf{X}} \quad & \tau\|\mathbf{X}\|_1 + \lambda\sum_{i=1}^{C}\|\mathbf{X}_{ii}\|_* + \alpha\|\mathbf{X}\|_F^2 \\ & + \|\mathbf{Y}-\mathbf{Y}\mathbf{W}\mathbf{X}\|_F^2, \\ \text{s.t.} \quad & \mathbf{X} = \mathbf{diag}(\mathbf{X}_{11},\mathbf{X}_{22},...,\mathbf{X}_{CC}). \end{aligned} \tag{12}$$

As $\mathbf{X} = \mathbf{diag}(\mathbf{X}_{11},\mathbf{X}_{22},...,\mathbf{X}_{CC})$, we have $\|\mathbf{X}\|_1 = \sum_{i=1}^{C}\|\mathbf{X}_{ii}\|_1$ and $\|\mathbf{X}\|_F^2 = \sum_{i=1}^{C}\|\mathbf{X}_{ii}\|_F^2$. So the above model is reformed as,

$$\begin{aligned} \min_{\mathbf{W},\mathbf{X}} \quad & \tau\sum_{i=1}^{C}\|\mathbf{X}_{ii}\|_1 + \lambda\sum_{i=1}^{C}\|\mathbf{X}_{ii}\|_* + \alpha\sum_{i=1}^{C}\|\mathbf{X}_{ii}\|_F^2 \\ & + \|\mathbf{Y}-\mathbf{Y}\mathbf{W}\mathbf{X}\|_F^2, \\ \text{s.t.} \quad & \mathbf{X} = \mathbf{diag}(\mathbf{X}_{11},\mathbf{X}_{22},...,\mathbf{X}_{CC}). \end{aligned} \tag{13}$$

We call this final model the Block-Diagonal Sparse Representation based Linear Combination Dictionary Learning (BDSRLCDL). Compared with the majority of existing dictionary learning methods, the most significant difference is that we keep the coefficient matrix in a block-diagonal structure. Additionally, the proposed model is characterized with the low rank constraint on each block-diagonal submatrix and the linear combination dictionary. In the next section, we will propose an efficient algorithm to solve the optimization problem.

Having learned the linear combination dictionary $\mathbf{Y}\mathbf{W}$ from the training samples, for a test sample $\tilde{y}$ we can calculate its sparse representation $\tilde{x}$ by solving the following problem:

$$\min_{\tilde{x}} \|\tilde{y}-\mathbf{Y}\mathbf{W}\tilde{x}\|_2^2 + \tau\|\tilde{x}\|_1 + \lambda\|\tilde{x}\|_2^2. \tag{14}$$

Then the sparse coefficient vector $\tilde{x}$ can be used for recognition. For convenience, we rewrite $\mathbf{W}$ as $\mathbf{W} = [\mathbf{W}_1,\mathbf{W}_2,...,\mathbf{W}_C]$, where $\mathbf{W}_i \in \mathbb{R}^{N\times K_i}, i = 1,2,...,C$ is the coefficient correspondending to i-th class. So we obtain each sub-dictionary by $\mathbf{D}_i = \mathbf{Y}\mathbf{W}_i, i = 1,2,...,C$. If we formulate the coefficient vector $\tilde{x}$ as $\tilde{x} = [\tilde{x}_1^T,\tilde{x}_2^T,...,\tilde{x}_C^T]^T$, where $\tilde{x}_i$ is the sparse coefficients corresponding to the i-th sub-dictionary $\mathbf{D}_i = \mathbf{Y}\mathbf{W}_i$, we can calculate the error on the i-th sub-dictionary as follows:

$$e_i = \|\tilde{y}-\mathbf{Y}\mathbf{W}_i\tilde{x}_i\|. \tag{15}$$

From these reconstruction errors on sub-dictionaries, the test sample $\tilde{y}$ could be identified as follows:

$$identity(\tilde{y}) = \arg\min_i\{e_i\}. \tag{16}$$

where $identity(\tilde{y})$ is the class label of the test sample $\tilde{y}$.

IV. Optimization solution to BDSRLCDL

To solve the optimization problem of BDSRLCDL in (13), we adopt an alternating direction method and separate the problem into several subproblems. We firstly introduce a set of extra variables $\{\mathbf{Z}_{ii}\}_{i=1}^{C}$ and let $\mathbf{Z}_{ii} = \mathbf{X}_{ii}$ to separate the $\|\cdot\|_1$ term and the $\|\cdot\|_*$ term involving $\mathbf{X}$ according to [35]. So the problem in (13) can be reformulated as the following problem with the introduced linear constraints,

$$\begin{aligned} \min_{\mathbf{W},\mathbf{X},\{\mathbf{Z}_{ii}\}_{i=1}^{C}} \quad & \tau\sum_{i=1}^{C}\|\mathbf{X}_{ii}\|_1 + \lambda\sum_{i=1}^{C}\|\mathbf{Z}_{ii}\|_* + \alpha\sum_{i=1}^{C}\|\mathbf{X}_{ii}\|_F^2 \\ & + \|\mathbf{Y}-\mathbf{Y}\mathbf{W}\mathbf{X}\|_F^2, \\ \text{s.t.} \quad & \mathbf{Z}_{ii} = \mathbf{X}_{ii}, i = 1,...,C \\ & and \quad \mathbf{X} = \mathbf{diag}(\mathbf{X}_{11},\mathbf{X}_{22},...,\mathbf{X}_{CC}). \end{aligned} \tag{17}$$

Then we can get the following objective function of the problem by the augmented Lagrangian multiplier method.

$$\begin{aligned}
&\mathcal{L}(\mathbf{Z},\mathbf{X},\mathbf{W},\mathbf{F},\gamma)\\
=&\tau\sum_{i=1}^{C}\|\mathbf{X}_{ii}\|_1+\lambda\sum_{i=1}^{C}\|\mathbf{Z}_{ii}\|_*+\alpha\sum_{i=1}^{C}\|\mathbf{X}_{ii}\|_F^2\\
&+\|\mathbf{Y}-\mathbf{YWX}\|_F^2\\
&+\sum_{i=1}^{C}(<\mathbf{F}_{ii},\mathbf{Z}_{ii}-\mathbf{X}_{ii}>+\frac{\gamma}{2}\|\mathbf{Z}_{ii}-\mathbf{X}_{ii}\|_F^2)\\
=&\tau\sum_{i=1}^{C}\|\mathbf{X}_{ii}\|_1+\lambda\sum_{i=1}^{C}\|\mathbf{Z}_{ii}\|_*+\alpha\sum_{i=1}^{C}\|\mathbf{X}_{ii}\|_F^2\\
&+\|\mathbf{Y}-\mathbf{YWX}\|_F^2\\
&+\sum_{i=1}^{C}(\frac{\gamma}{2}\|\mathbf{Z}_{ii}-\mathbf{X}_{ii}+\frac{\mathbf{F}_{ii}}{\gamma}\|_F^2-\frac{1}{2\gamma}\|\mathbf{F}_{ii}\|_F^2),
\end{aligned}\tag{18}$$

where $\mathbf{F}_{ii}$ is the Lagrangian multipliers and γ is an adaptive weight parameter for enforcing the condition $\mathbf{Z}_{ii}=\mathbf{X}_{ii}$, $\langle\mathbf{A},\mathbf{B}\rangle=\text{trace}(\mathbf{A}^T\mathbf{B})$. For the objective function in (18), we adopt the linearized alternating direction method in [37] to solve the optimization problem by an iteration procedure. The following steps give the detailed iterations for $\mathbf{Z}_{ii},\mathbf{X}_{ii},\mathbf{W}$ and other parameters alternately. Superscript t denotes the current iteration.

A. Calculate $\mathbf{Z}_{ii}$ while fixing $\mathbf{W}$ and $\mathbf{X}_{ii}$

When $\mathbf{W}$ and $\mathbf{X}_{ii}$ are fixed, the objective function is degenerated into a function with respect to $\mathbf{Z}_{ii}$. So we solve $\mathbf{Z}_{ii}$ by the following optimalization problem,

$$\begin{aligned}
\mathbf{Z}_{ii}^{t+1}&=\arg\min_{\mathbf{Z}_{ii}}\ \lambda\|\mathbf{Z}_{ii}\|_*+\frac{\gamma^t}{2}\|\mathbf{Z}_{ii}-\mathbf{X}_{ii}^t+\frac{\mathbf{F}_{ii}^t}{\gamma^t}\|_F^2\\
&=\arg\min_{\mathbf{Z}_{ii}}\ \frac{\lambda}{\gamma^t}\|\mathbf{Z}_{ii}\|_*+\frac{1}{2}\|\mathbf{Z}_{ii}-(\mathbf{X}_{ii}^t-\frac{\mathbf{F}_{ii}^t}{\gamma^t})\|_F^2.
\end{aligned}\tag{19}$$

This problem has closed-form solution as

$$\mathbf{Z}_{ii}^{t+1}=\mathbf{U}S_{\frac{\lambda}{\gamma^t}}[\mathbf{\Sigma}]\mathbf{V}^T\tag{20}$$

where $\mathbf{U\Sigma V}^T$ is the singular value decomposition (SVD) of $(\mathbf{X}_{ii}^t-\frac{\mathbf{F}_{ii}^t}{\gamma^t})$. $S_{\frac{\lambda}{\gamma^t}}[\cdot]$ is the soft-thresholding operator [38] with the following defination,

$$S_{\frac{\lambda}{\gamma^t}}[x]=\begin{cases}x-\frac{\lambda}{\gamma^t}, & \text{if } x>\frac{\lambda}{\gamma^t},\\ x+\frac{\lambda}{\gamma^t}, & \text{if } x<-\frac{\lambda}{\gamma^t},\\ 0, & \text{otherwise.}\end{cases}\tag{21}$$

B. Calculate $\mathbf{X}_{ii}$ while fixing $\mathbf{W}$ and $\mathbf{Z}_{ii}$

When $\mathbf{W}$ and $\mathbf{Z}_{ii}$ are fixed, we rewrite the objective function as the following form:

$$\begin{aligned}
&\mathcal{L}(\mathbf{Z},\mathbf{X},\mathbf{W},\mathbf{F},\gamma)\\
=&\tau\sum_{i=1}^{C}\|\mathbf{X}_{ii}\|_1+\lambda\sum_{i=1}^{C}\|\mathbf{Z}_{ii}\|_*+\alpha\sum_{i=1}^{C}\|\mathbf{X}_{ii}\|_F^2\\
&+\sum_{i=1}^{C}\|\mathbf{Y}_i-\mathbf{YW}_i\mathbf{X}_{ii}\|_F^2\\
&+\sum_{i=1}^{C}(\frac{\gamma}{2}\|\mathbf{Z}_{ii}-\mathbf{X}_{ii}+\frac{\mathbf{F}_{ii}}{\gamma}\|_F^2-\frac{1}{2\gamma}\|\mathbf{F}_{ii}\|_F^2),
\end{aligned}\tag{22}$$

Let

$$\begin{aligned}
&h_i(\mathbf{Z}_{ii},\mathbf{X}_{ii},\mathbf{W}_i,\mathbf{F}_{ii},\gamma)\\
=&\alpha\|\mathbf{X}_{ii}\|_F^2+\|\mathbf{Y}_i-\mathbf{YW}_i\mathbf{X}_{ii}\|_F^2+\frac{\gamma}{2}\|\mathbf{Z}_{ii}-\mathbf{X}_{ii}+\frac{\mathbf{F}_{ii}}{\gamma}\|_F^2,
\end{aligned}$$

then $\mathbf{X}_{ii}$ can be solved by the following optimalization,

$$\begin{aligned}
\mathbf{X}_{ii}^{t+1}&=\arg\min_{\mathbf{X}_{ii}}\ \tau\|\mathbf{X}_{ii}\|_1+h_i(\mathbf{Z}_{ii}^{t+1},\mathbf{X}_{ii},\mathbf{W}_i^t,\mathbf{F}_{ii}^t,\gamma^t)\\
&=\arg\min_{\mathbf{X}_{ii}}\ \tau\|\mathbf{X}_{ii}\|_1+<\bigtriangledown_{\mathbf{X}_{ii}^t}h_i,\mathbf{X}_{ii}-\mathbf{X}_{ii}^t>\\
&\qquad+\eta_i^t\|\mathbf{X}_{ii}-\mathbf{X}_{ii}^t\|_F^2\\
&=\arg\min_{\mathbf{X}_{ii}}\ \frac{\tau}{2\eta_i^t}\|\mathbf{X}_{ii}\|_1+\frac{1}{2}\|\mathbf{X}_{ii}-(\mathbf{X}_{ii}^t-\frac{\bigtriangledown_{\mathbf{X}_{ii}^t}h_i}{2\eta_i^t})\|_F^2,
\end{aligned}\tag{23}$$

where $\bigtriangledown_{\mathbf{X}_{ii}^t}h_i$ represents the partial differential at $\mathbf{X}_{ii}^t$ of h_i with respect to $\mathbf{X}_{ii}$ and has the form of $\bigtriangledown_{\mathbf{X}_{ii}^t}h_i=\gamma(\mathbf{X}_{ii}^t-\mathbf{Z}_{ii}^{t+1}-\frac{\mathbf{F}_{ii}^t}{\gamma_t})+2(\mathbf{W}_i^t)^T\mathbf{Y}^T\mathbf{YW}_i^t\mathbf{X}_{ii}^t+2\alpha\mathbf{X}_{ii}^t$, $\eta_i^t=\|\mathbf{YW}_i^t\|_2^2+\frac{\gamma^t}{2}$. From the conclusions in [39], [40], the closed-form solution to the problem is given by the following form,

$$\mathbf{X}_{ii}^{t+1}=\text{sign}(\mathbf{X}_{ii}^t-\frac{\bigtriangledown_{\mathbf{X}_{ii}^t}h_i}{2\eta_i^t})\max\{|\mathbf{X}_{ii}^t-\frac{\bigtriangledown_{\mathbf{X}_{ii}^t}h}{2\eta_i^t}|-\frac{\tau}{2\eta_i^t},0\}.\tag{24}$$

C. Calculate $\mathbf{W}$ while fixing $\mathbf{Z}_{ii}$ and $\mathbf{X}_{ii}$

When $\mathbf{Z}_{ii}$ and $\mathbf{X}_{ii}$ are fixed, the optimalization problem in (17) changes to the following problem:

$$\min_{\mathbf{W}}\quad\|\mathbf{Y}-\mathbf{YWX}^{t+1}\|_F^2,\tag{25}$$

where $\mathbf{X}^{t+1}=\mathbf{diag}(\mathbf{X}_{11}^{t+1},\mathbf{X}_{22}^{t+1},...,\mathbf{X}_{CC}^{t+1})$ is a strict block-diagonal coefficient matrix conducted from the current $\mathbf{X}_{ii}^{t+1},i=1,...,C$ in the above subsection.

To solve the problem in (25), we adopt an atom-by-atom iteration scheme to update the current $\mathbf{W}$, denoted by $\mathbf{W}^t$, similar to the algorithm in [13], [17]. For convenience, we let $\hat{\mathbf{W}}=\mathbf{W}^t$ and denote its each column by $\hat{w}_k,k=1,...,K$. Then each column $\hat{w}_k$ of $\hat{\mathbf{W}}$ can be updated by solving the following problem, while fixing the other columns.

$$\begin{aligned}
\min_{\hat{w}_k}\quad&\|\mathbf{YE}_k-\mathbf{Y}\hat{w}_kx_k^{t+1}\|_F^2\\
\text{s.t.}\quad&\|\mathbf{Y}\hat{w}_k\|_2=1.
\end{aligned}\tag{26}$$

where $x_k^{t+1}, k = 1, ..., K$ is the k-th row of $\mathbf{X}^{t+1}$, the condition $\|\mathbf{Y}\hat{w}_k\|_2 = 1$ demands the dictionary atoms having unit scalar, and $\mathbf{E}_k$ is defined as

$$\mathbf{E}_k = \mathbf{I} - \sum_{p \neq k} \hat{w}_p x_p^{t+1}. \tag{27}$$

For the problem in (26), we calculate the differential of the objective function with respect to $\hat{w}_k$ and let it be 0. Then we have a solution to the problem as follows,

$$\hat{w}_k = \mathbf{E}_k x_k^T / (x_k x_k^T). \tag{28}$$

In addition, to satisfy the normalized constraint $\|\mathbf{Y}\hat{w}_k\|_2 = 1$, we further update $\hat{w}_k$ by the following formula and obtain a final solution of $\hat{w}_k$.

$$\hat{w}_k = \hat{w}_k / \|\mathbf{Y}\hat{w}_k\|_2 \tag{29}$$

The above procedure of solving $\mathbf{W}$ is summarized in Algorithm 1.

D. Update the multiplier $\mathbf{F}_{ii}$ and parameter γ

After updating the coefficient matrix $\mathbf{Z}, \mathbf{X}$ and the dictionary combination matrix $\mathbf{W}$ at each iteration, the multiplier $\mathbf{F}_{ii}$ and parameter γ should also be updated by the following formulas:

$$\mathbf{F}_{ii}^{t+1} = \mathbf{F}_{ii}^t + \gamma(\mathbf{Z}_{ii}^{t+1} - \mathbf{X}_{ii}^{t+1}). \tag{30}$$

$$\gamma^{t+1} = \min\{\rho\gamma^t, \gamma^{max}\}, \tag{31}$$

where $\rho = 1.1$, $\gamma^{max} = 10^{10}$.

Integrating the above iterations, the solution to the proposed BDSRLCDL model is obtained and the complete algorithm is summarized in Algorithm 2. However, the convergence of the ADMM method could not be guaranteed as being discussed in [41]. However, we found that the algorithm always converges in our experiments. In our algorithm, the stopping criterion is measured by the following two conditions,

$$\max \left\{ \begin{array}{c} \|\mathbf{Z}^{t+1} - \mathbf{X}^{t+1}\|_\infty, \|\mathbf{Z}^{t+1} - \mathbf{Z}^t\|_\infty, \\ \|\mathbf{X}^{t+1} - \mathbf{X}^t\|_\infty \end{array} \right\} \leq \varepsilon_1. \tag{32}$$

$$\frac{|\mathcal{L}^{t+1} - \mathcal{L}^t|}{|\mathcal{L}^t|} \leq \varepsilon_2. \tag{33}$$

where $\|\cdot\|_\infty$ denotes the Infinite Norm, and $\mathcal{L}^t = \mathcal{L}(\mathbf{Z}^t, \mathbf{X}^t, \mathbf{W}^t, \mathbf{F}^t, \gamma^t)$ represents the value of the objective function at t iteration. (32) and (33) should be satisfied at same time to stop the iteration.

Fig. 1 shows the convergence of BDSRLCDL on the AR face dataset [42]. It shows that the curves decrease fast and almost tend to be stable after 30 iterations, which verifies our BDSRLCDL algorithm with good convergent property.

Algorithm 1 Calculate $\mathbf{W}$ while fixing $\mathbf{Z}_{ii}$ and $\mathbf{X}_{ii}$

Input: The training data set $\mathbf{Y}$, the current $\mathbf{W}^t$ and $\{\mathbf{X}_{ii}^{t+1}\}_{i=1}^C$.

1: Construct a strict block-diagonal coefficient matrix $\hat{\mathbf{X}} = \mathbf{diag}(\mathbf{X}_{11}^{t+1}, \mathbf{X}_{22}^{t+1}, ..., \mathbf{X}_{CC}^{t+1})$.
2: Let $\hat{\mathbf{W}} = \mathbf{W}^t$.
3: **for** each $k \in 1, 2, ..., K$ **do**
4: Calculate $\mathbf{E}_k$ by (27);
5: Calculate $\hat{w}_k$ by (28);
6: Update $\hat{w}_k$ by (29) .
7: **end for**
8: Let $\mathbf{W}^{t+1} = \hat{\mathbf{W}}$.

Output:
The matrix $\mathbf{W}^{t+1}$.

Algorithm 2 Block-Diagonal Sparse Representation based Linear Combination Dictionary Learning (BDSRLCDL)

Input: The training data set $\mathbf{Y}$, the parameters τ, λ, α.

1: **Initialize** : $\{\mathbf{Z}_{ii}^0\}_{i=1}^C = \{\mathbf{X}_{ii}^0\}_{i=1}^C = \mathbf{0}$, $\{\mathbf{F}_{ii}^0\}_{i=1}^C = \mathbf{1}$, initialize $\mathbf{W}^0$ randomly and let $w_i^0 = w_i^0 / \|\mathbf{Y}w_i^0\|_2$, $\gamma^0 = 10^{-4}$, $\rho = 1.1$, $\gamma^{max} = 10^{10}$, $\varepsilon_1 = \varepsilon_2 = 10^{-7}$, the number of maximum iteration $MaxIter = 1000$.
2: $t = 0$.
3: **while** not converged and $t \leq MaxIter$ **do**
4: Calculate $\{\mathbf{Z}_{ii}^{t+1}\}_{i=1}^C$ by (20);
5: Calculate $\{\mathbf{X}_{ii}^{t+1}\}_{i=1}^C$ by (23);
6: Calculate $\mathbf{W}^{t+1}$ by Algorithm 1;
7: Calculate $\{\mathbf{F}_{ii}^{t+1}\}_{i=1}^C$ by (30);
8: Calculate γ^{t+1} by (31);
9: $t = t + 1$.
10: **end while**

Output:
The matrices $\mathbf{W}^t$, $\mathbf{X}^t$.

V. Experimental Results

To evaluate the proposed method, we implement recognition experiments on various public datasets of different types. There are four types of datasets: i) Face image datasets, including Extended Yale B face dataset [43] and AR face dataset [42]; ii) Texture image sets, including KTH-TIPS texture dataset [44] and DynTex++ dataset [45]; iii) Scene image sets, including 15-Scene dataset [48] and UCF sports action dataset [51]; iv) Object image sets, including Oxford Flowers 17 dataset [49] and Pittsburgh Food Image dataset (PFID) [52]. These datasets are challenging for recognition applications.

The performance of the proposed method is compared with some state-of-the-art dictionary learning algorithms, such as SRC [1], DLSI [18], LC-KSVD [16], FDDL [17], MFL [13], DL-COPAR [19], DGSDL [20] and Discriminative Collaborative Representation Dictionary learning method (DCR) [21].

A. Face recognition

We test our proposed algorithm for face recognition on two face datasets: Extended Yale B [43] and AR face dataset [42]. The former has face images with various illuminations. The later has face images with different expressions and illuminations. Both of them are challenging data sets for face recognition.

1) Extended Yale B: This dataset contains 2,414 frontal face images of 38 subjects captured under various laboratory-controlled lighting conditions, i.e. the number of classes $C = 38$. Each subject has about 64 images. Fig.2 shows some

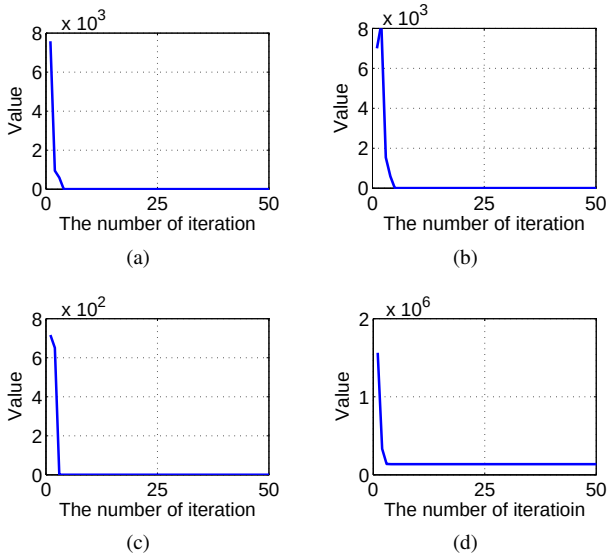

Fig. 1. The convergence curves of our BDSRLCDL method on the AR dataset [42]. (a) the convergence curve of $\|\mathbf{Z}^{t+1} - \mathbf{X}^{t+1}\|_\infty$; (b) the convergence curve of $\|\mathbf{Z}^{t+1} - \mathbf{Z}^{t}\|_\infty$; (c) the convergence curve of $\|\mathbf{X}^{t+1} - \mathbf{X}^{t}\|_\infty$; (d) the convergence curve of the objective function in (18).

samples of the dataset. In our experiment, we randomly select 20 images of each subject to compose the training set, and the rest images are used for testing. All the images are cropped and normalized to the size of 32×32 pixels. The parameters τ, α, λ are tuned manually by the experiment results, here $\tau = \lambda = \alpha = 0.001$. We simply set the atom number of all sub-dictionary as same number, here $K_i = 20$, $i = 1, ..., C$. The recognition experiments are repeated 10 times and the mean recognition rate is reported to evaluate the performance.

The experiment results are shown in Table I. It is shown that the proposed method obtains the highest recognition rate of 96.62% (in bold text), which is higher than the second (underlined) by 0.61%.

Generally, the atom number of the dictionary is critical to dictionary learning methods. So we further investigate the influence of the atom numbers on these dictionary learning methods. The recognition experiments are implemented with different atom numbers, here K_i is set from 8 to 20 for all methods. The experimental results are shown in Fig. 3. It can be observed that the recognition rate of the BDSRLCDL changes little with the standard deviation 0.49% compared with DLSI 1.59%, LC-KSVD 1.72%, FDDL 1.49%, MFL 2.28%, DL-COPAR 1.80%, DGSDL 1.66% and DCR 1.27%. Thus our method is robust to the scale of the learned dictionary. Even with small number of atoms in the dictionary, the recognition accuracy downgrades little. Note that the different number of atoms in the experiment is not implemented for SRC method as SRC uses the training data itself as the dictionary.

2) AR face dataset: The AR face dataset contains over 4,000 frontal face images from 126 persons (70 men and 56 women). These face images are captured under different facial expressions, illuminations, and occlusions (sun glasses and scarf). The face images under the same conditions were

Fig. 2. Some face images of the Extended Yale B dataset.

TABLE I
RECOGNITION RESULTS OF DIFFERENT METHODS ON EXTENDED YALE B DATASET WITH $K_i = 20$.

Algorithm	Recognition rate (%)
SRC	88.50
DLSI	94.03
LC-KSVD	94.42
FDDL	93.92
MFL	93.65
DL-COPAR	95.11
DGSDL	95.72
DCR	<u>96.01</u>
BDSRLCDL	**96.62**

captured in two sessions, which are separated by two weeks (14 days). Fig. 4 shows some samples of AR face dataset. In our experiment, we adopt the same way of [1], [17] and [20] to construct the training and testing data for recognition experiment, in which, a subset of images of 100 persons (50 men and 50 women) is selected ($C = 100$). For each person, seven images selected from Session 1 are used as training data, and seven images selected from Session 2 are used as testing data. All the selected images are cropped and normalized to the size of 50×40 pixels. The parameters $\tau = \lambda = 0.0002$, $\alpha = 0.00015$. The atom number of sub-dictionary $K_i = 7, i = 1, ..., C$. In the above data setting, the training set and test set are fixed, so there is no experiment repeated.

Table II reports the experimental results of the proposed method and other comparing methods. It can be seen that the proposed BDSRLCDL algorithm obtains the highest recognition rate of 95.22% (in bold text), which is higher than the second (underlined) by 0.80%. We also do the experiments with different number of the sub-dictionary atoms, here $K_i = 4$ to 7. The experimental results are shown in Fig. 5. It is shown that the result of our method is also robust to the variety of the atoms number of sub-dictionary.

The results of the experiments on the above two datasets have demonstrated that our proposed method has better performance on face recognition compared with other methods. The better performance may benefit from the introduction of the block-diagonal and low rank constraints. Additionally, our method shows robustness to the dictionary size.

B. Texture classification

In this experiment, we evaluate the proposed algorithm on two widely used texture datasets. The first is KTH-TIPS dataset [44], which is a static texture dataset containing various texture images. It is a challenging texture dataset for recognition as the images are captured in different scale, pose and illumination. The other is the dynamic texture dataset

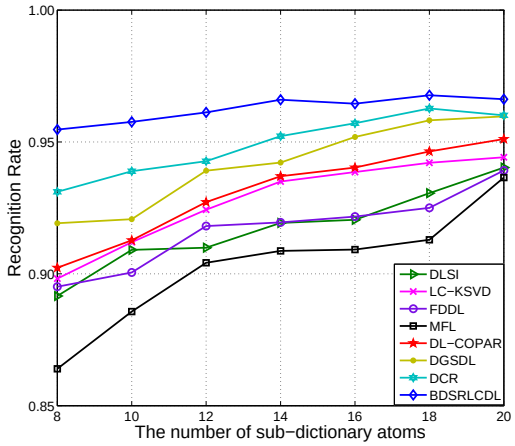

Fig. 3. Recognition results of different methods on Extended Yale B face dataset with the number of sub-dictionary atoms changing.

Fig. 4. Some face images of AR face dataset.

named DynTex++ [45], which is a set of dynamic texture videos captured in different complex scenarios. The dynamic texture video clips are divided into dozens of classes. It is also a challenging dataset for recognition.

1) KTH-TIPS texture dataset: This dataset contains 10 classes of static texture images ($C = 10$). Each class has 81 texture images captured at nine scales, three illumination directions and three poses. Some samples of this dataset are shown in Fig. 6. In our experiment, we randomly select 40 images from each class as training data and use the rest as test data. To represent the feature of these texture images, instead of using the original image raw data, we use the PRI-CoLBP$_0$ descriptor in [54] to construct the sparse model and implement recognition experiment. The parameters are set to $\tau = \lambda = \alpha = 10^{-6}$ and $K_i = 40$. Each experiment is also repeated 10 times.

The experimental results are shown in Table III. It is shown that our method has best recognition rate compared with other methods. The recognition experiments are also conducted with different numbers of dictionary atoms ($K_i = 20$ to 40). The results in Fig. 7 indicates that our method has high recognition accuracy with different number of atoms in the learned dictionary.

2) DynTex++ dataset: This dataset contains 345 video sequences, which are captured from different scenarios such as river water, fish swimming, smoke, cloud and so on. These videos are divided into 36 classes ($C = 36$) and each class has 100 subsequences (totally 3600 subsequences) with a fixed size of $50 \times 50 \times 50$ (50 gray frames). Some samples of the dataset are shown in Fig. 8.

Considering the high dimension of the original video clips, we adopt the method of Grassmann manifold in [46] to represent the dynamic texture samples. In this method, each video clip is firstly represented as the Local Binary Patterns from Three Orthogonal Plans (LBP-TOP) feature, which is proved an efficient way to capture the dynamic texture feature [47]. Then the LBP-TOP feature is used to construct the Grassmann Manifold points in form of a column orthogonal matrix in size of 177×14. To get a proper distance measurement, these Grassmann Manifold points are transformed into symmetrical matrics in size of 177×177, which have Euclid like distance and finally are used to construct the sparse model in our experiment. More detail can be found in [46]. In our experiment, we select 50 video clips randomly from each class used for dictionary learning and the rest for testing. The parameters $\tau = \lambda = 10^{-6}$, $\alpha = 10^{-4}$, and $K_i = 50$. Each experiment is also repeated 10 times.

The experimental results are shown in Table IV. Once again the proposed method performs excellent against all the other methods in terms of recognition rate. Similarly we also explore the influence of dictionary sizes and the results are shown in Fig. 9, here $K_i = 25$ to 50. It is obvious that our method has the best performance than all the other methods.

These two experiments show that the proposed method offers the best recognition performance with certain robustness for the static and dynamic texture datasets. In general, the dynamic texture contains more complex variety compared with static texture, which results in a lower recognition rate.

TABLE II
RECOGNITION RESULTS OF DIFFERENT METHODS ON AR FACE DATASET WITH $K_i = 7$.

Algorithm	Recognition rate (%)
SRC	89.14
DLSI	89.61
LC-KSVD	93.96
FDDL	93.00
MFL	90.12
DL-COPAR	94.12
DGSDL	94.42
DCR	93.43
BDSRLCDL	**95.22**

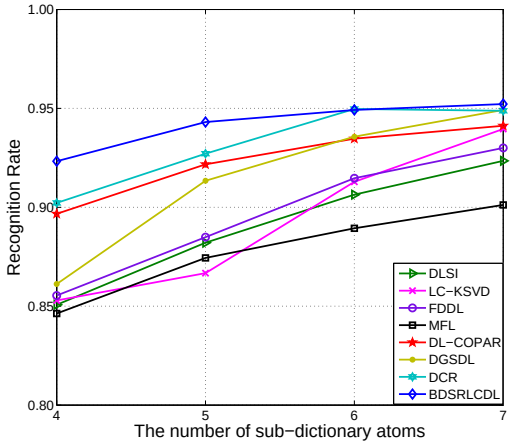

Fig. 5. Recognition results of different methods on AR face dataset with the number of sub-dictionary atoms changing.

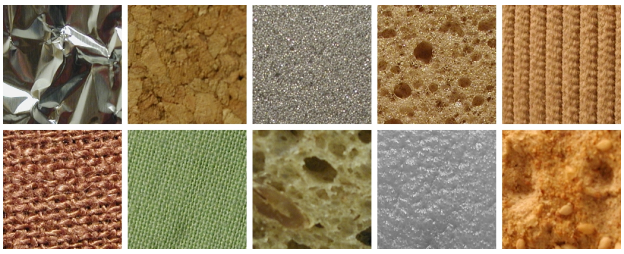

Fig. 6. Some texture images from KTH-TIPS dataset.

TABLE III
RECOGNITION RESULTS OF DIFFERENT METHODS ON KTH-TIPS DATASET WITH $K_i = 40$.

Algorithm	Recognition rate (%)
SRC	83.77
DLSI	96.00
LC-KSVD	96.21
FDDL	96.00
MFL	91.68
DL-COPAR	92.16
DGSDL	93.26
DCR	94.33
BDSRLCDL	**96.37**

Fig. 8. Samples images from DynTex++ dataset.

TABLE IV
RECOGNITION RESULTS OF DIFFERENT METHODS ON DYNTEX++ DATASET WITH $K_i = 50$.

Algorithm	Recognition rate (%)
SRC	86.20
DLSI	90.34
LC-KSVD	91.29
FDDL	92.03
MFL	90.02
DL-COPAR	91.77
DGSDL	90.43
DCR	90.27
BDSRLCDL	**92.35**

C. Scene recognition

The scene recognition or clustering is a classic problem in computer vision. As too much variety and influencing factors to make a proper descriptor for a complex scene, there are many challenges in scene recognition. To evaluate the proposed method, we implement scene recognition experiments on two public datasets, the 15-Scene dataset [48] and the UCF sports dataset [51]. The former contains a set of images with different indoor and outdoor scenes. The later is a specific scene dataset having different sport scenes in the form of video sequences.

1) The 15-Scene dataset: This dataset contains images of a wide range of outdoor and indoor scenes, such as bedrooms, kitchens, and country scenes, etc. There are totally 4485 images in this dataset and these images are divided into 15 categories ($C = 15$). Each category has 210 to 410 images with a size of 250×300 pixels. Some images of this dataset are shown in Fig. 10. For recognition experiment, we randomly select 100 images per category as training data and use the rest as test data. Here we adopted the spatial-pyramid feature and SIFT-descriptor in [16] to represent the images. The parameters are set to $\tau = 10^{-6}, \lambda = 10^{-5}, \alpha = 10^{-5}$ and $K_i = 50$. Each experiment is also repeated 10 times.

The experimental results are shown in Table V. It can be observed that the proposed method has 2+% improvement than other methods. The recognition experiment is also conducted to assess the robustness of the method with different number of sub-dictionary atoms. The results in Fig. 11 shows that our method is stable when the number of sub-dictionary atoms K_i varies from 50 to 100.

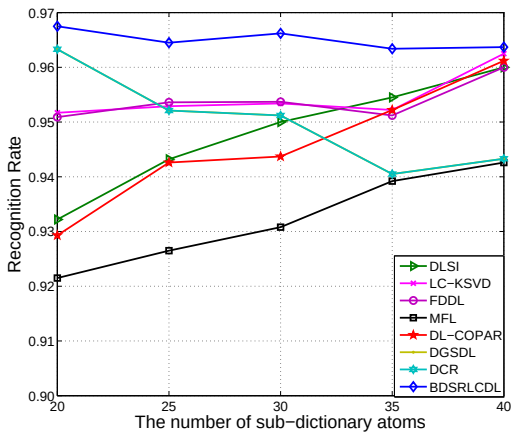

Fig. 7. Recognition rate on various methods with different number of sub-dictionary atoms on KTH-TIPS texture dataset

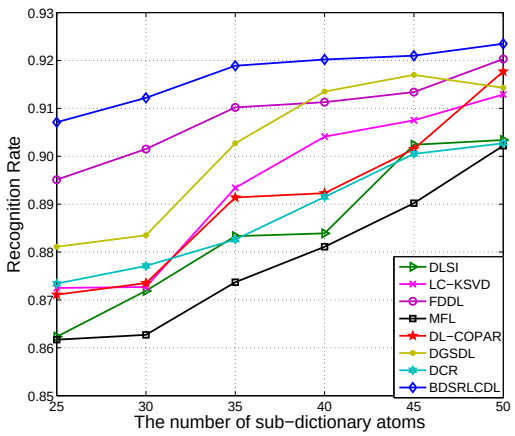

Fig. 9. Recognition results of different methods on DynTex++ dataset with the number of sub-dictionary atoms changing.

Fig. 10. Sample images from the 15-Scene dataset

TABLE V
RECOGNITION RESULTS OF DIFFERENT METHODS ON THE 15-SCENE DATASET WITH $K_i = 50$.

Algorithm	Recognition rate (%)
SRC	88.40
DLSI	94.22
LC-KSVD	93.17
FDDL	94.67
MFL	92.22
DL-COPAR	93.79
DGSDL	94.43
DCR	95.92
BDSRLCDL	**98.02**

2) The UCF sports dataset [51]: This dataset contains various video sequences collected from the sports channels of BBC and ESPN. These videos covers 10 classes of sport scenes ($C = 10$), including kicking, golfing, diving, horse riding, skateboarding, running, swinging, swinging highbar, lifting and walking. Each class contains 6 to 22 video sequences. Some examples of this dataset are shown in Fig. 12.

We implement recognition experiment in a fivefold cross validation manner, where four folds are used in training and the remaining one is used for testing. To obtain good recognition results and reduce the data dimension, the action bank features in [53] are extracted to represent the video sequences. The parameters are $\tau = \lambda = \alpha = 10^{-4}$ and $K_i = 10$. Each experiment is also repeated 10 times.

The experimental results are shown in Table VI which further confirm that the proposed method has the best performance in recognition rates compared with the other methods. However in this experiment no further experiments were conducted for other values of K_i as there are only few training samples for most of classes.

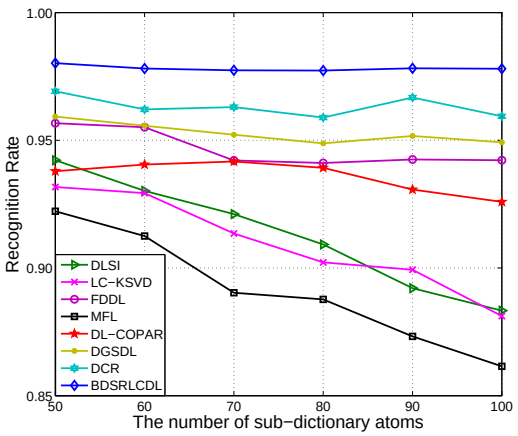

Fig. 11. Recognition results of different methods on the 15-Scene dataset with the number of sub-dictionary atoms changing.

Fig. 12. Some images of the UCF sports action dataset.

TABLE VI
RECOGNITION RESULTS OF DIFFERENT METHODS ON THE UCF SPORTS ACTION DATASET WITH $K_i = 10$.

Algorithm	Recognition rate (%)
SRC	92.72
DLSI	92.17
LC-KSVD	91.53
FDDL	94.33
MFL	90.34
DL-COPAR	90.73
DGSDL	91.27
DCR	92.35
BDSRLCDL	**94.52**

D. Object recognition

To further evaluate our method, we continue the recognition experiments on object data. Here we select two public image datasets of specific objects to do experiments. One is the Oxford Flowers 17 dataset [49] which contains 17 classes of flower images. The other is the Pittsburgh Food Image Dataset (PFID) [52] which is a recently constructed image set containing various food images. It is a challenging dataset for recognition as the difference between classes is minor. At present, many state-of-the-art recognition methods have poor performance on this dataset [16], [17], [21].

1) The Oxford Flowers 17 dataset: This dataset consists of 17 species of flowers with 80 images of each class. Some of species have very similar appearance. There are large viewpoint, scale, and illumination variations for the images. Some images of this dataset are shown in Fig. 13. This dataset is challenging for recognition as the large intra-class variability and sometimes small inter-class variability. We adopted the data splits provided on the website (www.robots.ox.ac.uk/~vgg/data/flowers) to construct our data setting, in which the training splits are used as our training data, and the validation and test splits are combined together as our test data. So there are totally 3 groups of data for experiments and each group has 40 training images and 40 test images for each class of 17 species ($C = 17$). All images are resized so that the smallest dimension is 500 pixels. To construct the sparse model, the recently proposed Frequent Local Histogram (FLH) [50] feature is adopted as the image representations.

Fig. 13. Some images of the Oxford Flowers 17 dataset.

TABLE VII
RECOGNITION RESULTS OF DIFFERENT METHODS ON THE OXFORD FLOWERS 17 DATASET WITH $K_i = 30$.

Algorithm	Recognition rate (%)
SRC	88.40
DLSI	88.87
LC-KSVD	90.20
FDDL	91.72
MFL	89.07
DL-COPAR	91.28
DGSDL	92.75
DCR	93.41
BDSRLCDL	**96.47**

The parameters are $\tau = \lambda = \alpha = 10^{-6}$ and $K_i = 30$. The mean recognition of the experiments on the 3 groups are shown in Table VII. The experiments with different size of dictionary are also implemented (Fig. 14 shows the results). We have the similar results as the above experiments.

2) The Pittsburgh Food Image Dataset: This Dataset is a recently released dataset containing a set of fast food images and videos captured from 13 chain restaurants. There are 61 categories of specific food items ($C = 61$) and each food category consists of three different instances (bought in different days from different branches of the restaurant chain). Each instance contains six images with different viewpoints. Some images of this dataset are shown in Fig.15. Similar to the experimental setting in [52], we randomly chosen two instances for training and the other one for testing for each category. To represent the food images, we also adopt the PRI-

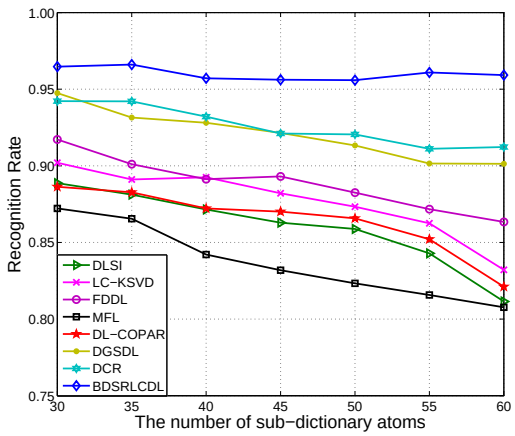

Fig. 14. Recognition results of different methods on the Oxford Flowers 17 dataset with the number of sub-dictionary atoms changing.

Fig. 15. Some images of the Pittsburgh Food Image Dataset.

TABLE VIII
RECOGNITION RESULTS OF DIFFERENT METHODS ON THE PITTSBURGH FOOD IMAGE DATASET WITH $K_i = 6$.

Algorithm	Recognition rate (%)
SRC	22.61
DLSI	21.65
LC-KSVD	23.15
FDDL	16.65
MFL	21.64
DL-COPAR	18.70
DGSDL	30.52
DCR	37.35
BDSRLCDL	**49.50**

CoLBP$_0$ feature descriptor [54] as the feature. The parameters are $\tau = \lambda = \alpha = 10^{-4}$ and $K_i = 6$. Each experiment is also repeated 10 times.

The experimental results are shown in Table VI. For this dataset, the proposed method significantly outperforms all the other methods with more than 12% gap in terms of recognition rates. Given that the high performance over the dataset, we omit the analysis on the robustness over the size of sub-dictionaries.

VI. CONCLUSION AND FUTURE WORKS

In this paper, we proposed a Block-Diagonal Sparse Representation model for recognition based on Linear Combination Dictionary Learning (BDSRLCDL) method. The method incorporates a parametric block-diagonal sparse representation in dictionary learning model to eliminate correlation between classes and to achieve better discriminative performance. To further enhance the sparse representational power, we enforce a low rank representation matrix to describe the correlation of the sparse representation within each class. Instead of using the conventional over-complete dictionary, a dictionary consisting of linear combinations of the training samples is adopted in the model. The proposed method is evaluated on wide range of public datasets in different recognition applications, such as face recognition, texture recognition, scene recognition and object recognition. It has been demonstrated that the proposed method outperforms most state-of-the-art dictionary learning methods. In future work, We will explore the extension of the method for the cases of data in non-Euclidean spaces such as kernel spaces and manifold spaces.

REFERENCES

[1] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust Face Recognition via Sparse Representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[2] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma, "Toward a Practical Face Recognition System: Robust Alignment and Illumination by Sparse Representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 2, pp. 372–386, Feb. 2012.

[3] M. Yang and L. Zhang, "Gabor feature based sparse representation for face recognition with gabor occlusion dictionary." in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2010. pp. 448–461.

[4] B. Zhang, A. Perina, V. Murino and A. D. Bue, "Sparse Representation Classification with Manifold Constraints Transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4557–4565.

[5] C. Zhang, J. Liu, Q. Tian, C. Xu, H. Lu, and S. Ma, "Image classification by non-negative sparse coding, low-rank and sparse decomposition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 1673–1680.

[6] B. Liu, Y. Wang, B. Shen, Y. Zhang, M. Hebert, "Self-explanatory sparse representation for image classification,", in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2014. pp. 600–616.

[7] T. Guha and R. K. Ward, "Learning sparse representations for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 8, pp. 1576–1588, Aug. 2012.

[8] X. Zhang, Y. Yang, L. C. Jiao, and F. Dong, "Manifold-constrained coding and sparse representation for human action recognition," in *Pattern Recognition.*, vol. 46, no. 7, pp. 1819–1831, Jul. 2013.

[9] Y. Zhu, X. Zhao, Y. Fu, and Y. Liu, "Sparse Coding on Local Spatial-Temporal Volumes for Human Action Recognition," in *Proc. Asian Conference on Computer Vision (ACCV)*, Nov. 2010. pp. 660–671.

[10] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," in *Proc. IEEE*, vol. 98, no. 6, pp. 1031–1044, Jun. 2010.

[11] K. Engan, S. O. Aase, and J. H. Husoy, "Frame based signal compression using method of optimal directions (MOD)," in *Proc. IEEE Int. Symp. Circuits Syst.*, Jul. 1999, pp. 1–4.

[12] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[13] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Metaface learning for sparse representation based face recognition," in *Proc. IEEE Conf. Image Process. (ICIP)*, Sep. 2010, pp. 1601–1604.

[14] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 2691–2698.

[15] Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent K-SVD," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 1697–1704.

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

[17] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *Proc. IEEE Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 543–550.

[18] I. Ramirez, P. Sprechmann, and G. Sapiro, "Recognitionand clustering via dictionary learning with structured incoherence and shared features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 3501–3508.

[19] S. Kong and D. Wang, "A dictionary learning approach for classification: Separating the particularity and the commonality," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2012, pp. 186–199.

[20] Y. Sun, Q. Liu, J. Tang, and D. Tao, "Learning Discriminative Dictionary for Group Sparse Representation," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3816–3828, Sep. 2014.

[21] Y. Wu, W. Li, M. Mukunoki, M. Minoh, and S. Lao, "Discriminative Collaborative Representation for Classification," in *Proc. Asian Conference on Computer Vision (ACCV)*, Nov. 2014, pp. 646–660.

[22] Y. Zhang, Z. Jiang, and L. S. Davis, "Learning structured low-rank representations for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun, 2013, pp. 676–683.

[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," in *AAAI Conf. on Artificial Intell.*, Jul. 2014, pp. 2810C2816.

[24] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proc. Int?l Conf. Machine Learning (ICML)*, Jun. 2010, pp. 663–670.

[25] R. Rubinstein, M. Zibulevsky, and M. Elad, "Double sparsity: Learning sparse dictionaries for sparse signal approximation," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1553–1564, Mar. 2010.

[26] H. Nguyen, V. M. Patel, N. M. Nasrabadi, and R. Chellappa, "Design of non-linear kernel dictionaries for object recognition," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5123–5135, Dec. 2013.

[27] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *IEEE Asilomar Conf. Signals, Syst. Comput.*, Nov. 1993, pp. 40–44

[28] J. Tropp and A. Gilbert, "Signal Recovery from Random Measurements via Orthogonal Matching Pursuit," *IEEE Trans. Information Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.

[29] M. Elad and M. Aharon, "Image Denoising Via Sparse and Redundant representations over Learned Dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.

[30] J. Mairal, M. Elad, and G. Sapiro, "Sparse Representation for Color Image Restoration," *IEEE Trans. Image Process.*, vol. 17, no. 1, pp. 53–69, Jan. 2008.

[31] Y. Romano and M. Elad, "Improving K-SVD denoising by post-processing its method-noise," in *Proc. IEEE Conf. Image Process. (ICIP)*, Sep. 2013, pp. 435–439.

[32] J. Rutterford and M. Davison, "An introduction to stock exchange investment," Palgrave Macmillan, 2007.

[33] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient Sparse Coding Algorithms," in *Proc. Neural Information Processing Systems Conf.*, vol. 19, pp. 801–808, Jun. 2006.

[34] G. Karol and Y. LeCun, "Learning Fast Approximations of Sparse Coding," in *Proc. Int?l Conf. Machine Learning (ICML)*, Jun. 2010, pp. 399–406.

[35] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, Jan, 2011.

[36] B. Recht, M. Fazel, and P. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, 2010.

[37] Z. Lin, R. Liu and Z. Su, "Linearized Alternating Direction Method with Adaptive Penalty for Low-Rank Representation," in *Proc. Advances in Neural Information Processing Systems*, 2011.

[38] Z. Lin, M. Chen, L. Wu, and Y. Ma, "The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices," arXiv preprint arXiv:1009.5055, 2010.

[39] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Convex optimization with sparsity-inducing norms," *Optimization for Machine Learning*, pp. 19–53, 2011.

[40] J. Liu and J. Ye, "Efficient l1/lq norm regularization," arXiv preprint arXiv:1009.4766, 2010.

[41] R. Liu, Z. Lin, and Z. Su, "Linearized Alternating Direction Method with Parallel Splitting and Adap-tive Penalty for Separable Convex Programs in Machine Learning," in *Proc. Asian Conf. Machine Learning.*, Nov. 2013, pp. 116–132.

[42] A. Martinez and R. Benavente. The AR face database. CVC Tech. Report No. 24, 1998.

[43] K. C. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 684–698, May. 2005.

[44] E. Hayman, B. Caputo, M. Fritz, and J. Eklundh, "On the significance of real-world conditions for material classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, May. 2004, pp. 253C-266.

[45] B. Ghanem and N. Ahuja, "Maximum margin distance learning for dynamic texture recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2010, pp. 223–236.

[46] B. Wang, Y. Hu, J. Gao, Y. Sun, and B. Yin, "Low Rank Representation on Grassmann Manifolds," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, Nov. 2014, pp. 653–668.

[47] G. Zhao and M, Pietikäinen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[48] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2006, pp. 2169–2178.

[49] M. Nlsback and A. Zisserman, "A visual vocabulary for flower classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2006, pp. 1447–1454.

[50] B. Fernando, E. Fromont, and T. Tuytelaars, "Effective use of frequent item set mining for image classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2012, pp. 214–227.

[51] M. Rodriguez, J. Ahmed, and M. Shah, " Action MACH: A Spatio-temporal Maximum Average Correlation Height Filter for Action Recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.
[52] M. Chen, K. Dhingra, W. Wu, L. Yang, R. Sukthankar, and J. Yang, "PFID: Pittsburgh fast-food image dataset," in *Proc. 16th IEEE Int. Conf. Image Process. (ICIP)*, Nov. 2009, pp. 289C-292.
[53] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 1234–1241.
[54] X. Qi, R. Xiao, C. Li, Y. Qiao, J. Guo, and X. Tang, "Pairwise Rotation Invariant Co-occurrence Local Binary Pattern," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2199–2213, Apr. 2014.