Pattern Recognition 59 (2016) 168–175

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Class relatedness oriented-discriminative dictionary learning for multiclass image classification

Dongyu Zhang [a,*], Pengju Liu [b], Kai Zhang [b], Hongzhi Zhang [b], Qing Wang [a], Xiaoyuan Jing [c]

[a] Sun Yat-Sen University, Guangzhou 510275, China
[b] Harbin Institute of Technology, Harbin 150001, China
[c] State Key Laboratory of Software Engineering of Wuhan University, WuHan 430079, China

ARTICLE INFO

Article history:
Received 4 September 2015
Received in revised form
10 November 2015
Accepted 7 December 2015
Available online 19 December 2015

Keywords:
Dictionary learning
Joint sparsity
$\ell_{1,\infty}$-norm
Support vector machine
Class relatedness

ABSTRACT

Dictionary learning (DL) has recently attracted intensive attention due to its representative and discriminative power in various classification tasks. Although much progress has been reported in the existing supervised DL approaches, it is still an open problem that how to build the relationship between dictionary atoms and the class labels in multiclass classification. In this paper, based on the assumption that the relevance of dictionary atoms could be helpful in multiclass classification task, we proposed a class relatedness oriented (CRO) discriminative dictionary learning method for sparse coding. Utilizing the $\ell_{1,\infty}$-norm regularization on the coding coefficient matrix, the proposed method can adaptively learn the class relatedness between dictionary atoms and the multiclass labels. Experimental results of face recognition, object classification, and action recognition demonstrate that our proposed method is comparable to many state-of-the-art DDL methods.

1. Introduction

As a hot topic in computer vision community, image classification and recognition have inspired many interesting works [1–6]. In recent years, sparse representations based images classification have received considerable interest [7–9]. By using over complete dictionary, sparse representation represents a signal or image as the sparse linear combination of the dictionary atoms. Unlike the principle component analysis based decomposition, sparse representation do not impose the dictionary elements to be orthogonal, which allows more flexibility to adapt the representation to the input signal or image. Recent publications show that sparse representation has been successfully applied to different kinds of images classification tasks, such as face recognition [9,36], handwriting digits recognition [18,37], natural images classification [10], and so on [11–13,40,41].

In sparsity based classification, a query sample $\mathbf{x} \in \mathcal{R}$ is first represented over a dictionary $\mathbf{D} \in \mathbb{R}^{m \times K}$ with a sparse coefficients vector α as $\mathbf{x} \approx \mathbf{D}\alpha$. Then the classification is performed on the coefficients vector α and dictionary $\mathbf{D}$. Thus, the effectiveness of the sparse representation model highly rely on the design of the overcomplete dictionary. One possible route to design the dictionary is to use a prespecified transform matrix, such as FFT bases or wavelet bases, which often leads to simple and fast algorithms for the evaluation of the sparse representation. Although taking these kinds of analytically designed off-the-shelf dictionary is universal to all types of images, it might be not effective enough for specific classification task, such as fine-categories flowers classification [37] and face recognition [9,36]. Another possible way to design the dictionary is to learn the dictionary elements/atoms from the input training data with sparsity regularization. Wright et al. [9] employed all of training samples as the dictionary for sparse representation and achieved impressive performances on face recognition. In [12], by generalizing k-means clustering, Aharon et al. proposed K-SVD algorithm to efficiently learn an overcomplete dictionary from a set of training samples. Recent works show that the dictionary learning methods have been received considerable interest and led to state-of-the-art results in image reconstruction [7,8,12,14] and image classification [9,11,15,16,36,37].

The current dictionary learning methods can be grouped into two categories, unsupervised dictionary learning [12,13] and supervised dictionary learning [15–17,37]. In unsupervised dictionary learning method, the dictionary is designed to minimize the residual error of reconstructing the training samples without using the classification labels. The dictionaries produced in such way can faithfully represent the training samples, which are useful for image reconstruction. However, they are not advantageous for image classification tasks. Although some of unsupervised dictionary learning were applied for classification tasks, recent

* Corresponding author.
E-mail address: cszhangdy@163.com (D. Zhang).

http://dx.doi.org/10.1016/j.patcog.2015.12.005
0031-3203/

research indicated that the supervised dictionary learning can yield better classification performance by exploiting the class discrimination information [15–17].

The current supervised dictionary learning methods can be roughly grouped into two categories. In the first category, a shared dictionary by all classes is learned with discriminative representation coefficients simultaneously [16,17,21]. In the second category, multiple dictionaries or class-specific dictionaries are learned [18,20,23,27,38]. Ramirez et al. [18] proposed a structured dictionary learning scheme by promoting the discriminative ability of different class-specific sub-dictionaries. Zhou et al. [27] proposed to learn multiple dictionaries for visually correlated object classification. However, there are potential problems in both of the two categories of supervised dictionary learning methods. For the first categories, each of the dictionary atom is associated to all the classes, while the possible mixed information of different classes may reduce the discrimination of the learned dictionary. For the second categories, each dictionary atom is assigned to a single class, but without exploiting the possible correlation of different class dictionary atoms which may help in promoting the discrimination performance. Both of the two cases have ignored the fact that the relationship between the dictionary atoms and class labels needs to be updated during the dictionary learning process. Although much progress have been made in dictionary learning, it is still an open problem to adaptively build the relationship between dictionary atoms and class labels.

In this paper, we proposed a new discriminative dictionary learning (DDL) scheme to adaptively learn the relationship between the dictionary atoms and the class labels by using a joint sparsity constraint on the coding vectors of each class of training samples. Specifically, the joint sparsity is enforced on the representation coefficients with the $\ell_{1,\infty}$-norm regularization, which has been widely applied as penalties in signal processing [24] and machine learning [25,26]. Besides, in order to make the coding vectors more discriminative, we also add a linear classifier to the objective function. Thus, we can learn a class related dictionary and a multiclass linear classifier simultaneously. For a test sample, we could use the learned dictionary to obtain the corresponding coding vector, and then predict its label with the linear classifier. Similar work of exploiting the relation of dictionary atoms and class labels has been studied recently. For example, in [37] Gao et al. learned a shared dictionary and category specific dictionaries for fine-grained flower images categorization, in which the shared dictionary used to describe the correlated relation of different classes. However, in their method the structures of shared dictionary and category specific dictionaries are pre-specified. We argue that the relationship between dictionary atoms and class labels should not be predefined. Instead, they should be learned adaptively during the dictionary learning process. This piece of work of is an extension of our conference paper, in which we further analyze the motivation and the principle of the proposed class relatedness dictionary method. More experiments have also been performed to comprehensively evaluate the proposed method.

The remainder of this paper is organized as follows. In Section 2, we briefly introduce the related work about DDL methods. Then we propose our class relatedness oriented (CRO) DDL model, together with the corresponding optimization method in Section 3. Experimental results of the proposed method are discussed in Section 4, and finally Section 5 concludes this paper.

2. Related work

Discriminative dictionary learning (DDL), has been successfully applied in pattern recognition applications such as image classification [18,19,29] and face recognition [9,20,36]. DDL methods concentrate on the discriminative classification capability of the dictionary as its goal is to assign correct class labels to the test samples. To enrich such capability, how to build the relationship between dictionary atoms and class labels plays a crucial role in the dictionary learning process.

Based on the relationship between dictionary atoms and class labels, prevailing DDL methods can be roughly divided into two main categories: global dictionary learning methods and class-specific dictionary learning methods. In global dictionary learning, the dictionary atoms are shared by all class and the coding vectors are generally explored for classification. Mairal et al. [21] proposed a DDL method by learning a shared dictionary and training a classifier on coding vectors simultaneously for handwriting digit recognition and texture classification. Zhang and Li [17] proposed a joint learning algorithm based on K-SVD for face recognition. Pham and Venkatesh [16] proposed to jointly train the dictionary and classifier for face recognition and object categorization. Cai et al. [22] proposed a support vector guided dictionary method (SVGDL) to jointly optimize the dictionary and classifier. Even though a global dictionary can be powerful to represent training data, all the above methods fail to adaptively learn the correspondence between dictionary atoms and class labels. As each dictionary atom is shared by all classes, the mixed information from different class samples may reduce the discrimination of the learned dictionary.

In class-specific DDL methods, each dictionary atom is assigned to a single class and the dictionary atoms associated with different classes are encouraged to be as independent as possible. Ramirez et al. [18] proposed a structured dictionary learning scheme by promoting the discriminative ability between different class-specific sub-dictionaries. Castrodad and Sapiro [23] learned a set of class-specific sub-dictionaries with non-negative penalty on both dictionary atoms and coding vectors. Yang et al. [20] proposed a DDL framework which employs Fisher discrimination criterion to learn class-specific dictionaries. Since each dictionary atom has a single label, the reconstruction error with respect to each class could be used for classification. However, those methods ignored the cross relatedness of different dictionary atoms and class labels, e.g., sometimes it is helpful in promoting the performance by assigning some dictionary atoms to different class labels in multiclass classification task.

The above DDL approach associates dictionary atoms and class labels in two extreme manners: the dictionary atom is either associated to all classes, or assigned to a single one. In order to adaptively build the relationship between dictionary atoms and class labels, we propose a well-principled DDL scheme by applying joint sparsity constraint on the coding vectors of each class with $\ell_{1,\infty}$-norm regularization, respectively. Since the $\ell_{1,\infty}$-norm is a matrix norm that encourages entire rows of a matrix to be zeros, the resultant coding vector of a certain class should be row sparse. Besides, by incorporating a classifier into the objective function to promote the discriminative ability of coding vectors, our method would adaptively build the relatedness between class labels and the dictionary atoms in the training phase.

3. The CRO-DDL method

In this section, we first briefly describe the general DDL model, and then propose our CRO-DDL method. The process of parameters optimization for the proposed method is also presented, and the classification rule is discussed in the end of this section.

3.1. The general DDL model

Let $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_C]$ be a set of **N** training samples with class label $y_i \in [1, \dots, C]$, where $\mathbf{X}_k$ is the subset containing n_k samples from the k-th class. The general DDL model can be described as follows:

$$\min_{\mathbf{D},\mathbf{A}} \mathcal{R}(\mathbf{X}, \mathbf{D}, \mathbf{A}) + \lambda_1 \|\mathbf{A}\| + \lambda_2 \mathbf{L}(\mathbf{A}), \tag{1}$$

where $\mathcal{R}(\mathbf{X}, \mathbf{D}, \mathbf{A})$ is the reconstruction error, $\mathbf{D} \in \mathbb{R}^{m \times K}$ is the learned dictionary of **C** classes, $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_C] \in \mathbb{R}^{K \times n}$ is the sparse coding coefficients of input training samples $\mathbf{X}$, λ_1 and λ_2 are the trade off parameters, $\|\mathbf{A}\|$ denotes certain norm for **A**. In general, $\|\mathbf{A}\|$ is set to be $\|\mathbf{A}\|_1$ to ensure the sparsity of coding vectors which tends to produce better classification results [31]. However, the ℓ_1-norm classification sparse coding suffers from high computation burden. To tackle this problem, some researchers attempt to use ℓ_2-norm regularization and their results can be very competitive with well-designed classification rule or classifier [22].

3.2. Formulation of CRO-DDL

Instead of learning a global dictionary without class specific property or a class-specific dictionary without class relatedness property, we propose to adaptively learn the relationship between dictionary atoms and the class labels.

Fig. 1 shows the motivation of our proposed methods. Supposes that dictionary **D** to be a collection of different kinds of features of images, where each dictionary atom corresponding a single category of features. For the samples of i-th class $\mathbf{X}_i$, it can be represented by the corresponding sparse coding coefficients matrix $\mathbf{A}_i$ over **D** as $\mathbf{X} \approx \mathbf{D}\mathbf{A}_i$. This progress can be regards as the course of feature selection. As the samples of $\mathbf{X}_i$ are with the same class label, they must share similar features. Thus, for the coding vectors $a_k, a_j \in \mathbf{A}_i$, the non-zero elements of vector a_k and a_j should be at the same rows. Thus, the coding matrix $\mathbf{A}_i$ should be group sparse. Besides, different class samples may share some common features as shown in Fig. 1. The i-th class samples share some atoms/features with class k, while class i may also share other features with class j. These shared features may be helpful in the classification of different classes, such as class k and class j. These shared atoms are not pre-specified for different class samples. They are learned from the training samples adaptively, by adopting the $\ell_{1,\infty}$-norm regularization together with the discriminative classifier.

Based on the above assumption, we adopt the $\ell_{1,\infty}$-norm for group sparse regularization and replace $\|\mathbf{A}\|$ with $\sum_{k=1}^{C} \|\mathbf{A}_k\|_{1,\infty}$ in Eq. (1). The $\sum_{k=1}^{C} \|\mathbf{A}_k\|_{1,\infty}$ penalizes the sum of maximum absolute values of each row of the k-th class coding matrix $\mathbf{A}_k$, and encourages entire row to be zero, which forcing $\mathbf{X}_k$ to be sparsely represented by the linear combination of atoms of dictionary **D**. Besides, in order to enlarge the discriminative ability of coding vectors, we use a multi-class support vector machine (SVM) as the discrimination term, i.e.,

$$\mathbf{L}(\mathbf{A}) = 2 \sum_{k=1}^{C} \mathcal{L}\left(\mathbf{A}, \mathbf{y}^k, \mathbf{w}_k, b_k\right), \tag{2}$$

where $\mathbf{w}_k$ denotes the normal to the k-th class hyperplane of SVM, b_k denotes the corresponding bias, and $\mathbf{y}^k = [y_1^k, y_2^k, \dots, y_n^k]$ is defined as $y_i^k = 1$ if class labels $y_i = k$ and otherwise $y_i^k = -1$. Specifically, the discrimination term is $\mathcal{L}(\mathbf{A}, \mathbf{y}^k, \mathbf{w}_k, b_k) = \frac{1}{2}\|w_k\|_2^2 + \theta \ell(\mathbf{A}, \mathbf{y}^k, \mathbf{w}_k, b_k)$, where $\ell(\mathbf{A}, \mathbf{y}^k, \mathbf{w}_k, b_k)$ is the hinge loss function, and θ is a predefined constant. Thus, the proposed CRO-DDL model can be formulated as follows:

$$\min_{\mathbf{D},\mathbf{A},\mathbf{w},\mathbf{b}} \|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2 + \lambda_1 \sum_{k=1}^{C} \|\mathbf{A}_k\|_{1,\infty} + 2\lambda_2 \sum_{k=1}^{C} \mathcal{L}\left(\mathbf{A}, \mathbf{y}^k, \mathbf{w}_k, b_k\right) \tag{3}$$

where $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_C]$, $\mathbf{b} = [b_1, b_2, \dots, b_C]$. According to [22], the solution of SVM classifier, can be represented as the linear combination of a few coding coefficient vectors, i.e., support vectors, which indicates that only the coding vectors near the hyperplane play the dominant role in the discriminative dictionary learning. The dictionary learned in this way is adaptive to the training data, and the optimal parameters of SVM classifier can also be simultaneously learned for classification.

3.3. Optimization of CRO-DDL

In this section, we describe our algorithm to find the optimal solution for all parameters of Eq. (3). As a joint optimization problem, Eq. (3) can be solved by optimizing the objective function with respect to **D**, **A** and $\langle \mathbf{W}, \mathbf{b} \rangle$ alternatively.

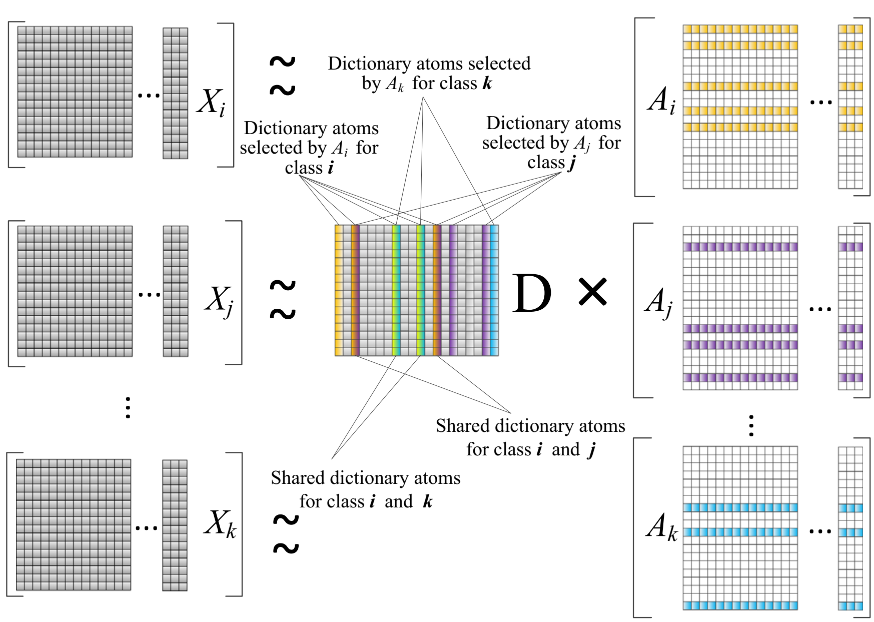

Fig. 1. Illustrations of class relatedness dictionary learned in the proposed method. Some of dictionary atoms are shared with different classes. Besides, the coding coefficient matrix of each class is group sparse.

Fig. 2. Sample illustrations from Extended Yale B dataset.

3.3.1. Update the coding coefficient matrix ***A***

Note that $\|\mathbf{X}-\mathbf{D}\mathbf{A}\|_F^2=\sum_{k=1}^{C}\|\mathbf{X}_k-\mathbf{D}\mathbf{A}_k\|_F^2$, and then the proposed CRO-DDL model can be reformulated as

$$\min_{\mathbf{D},\mathbf{A},\mathbf{W},\mathbf{b}}\sum_{k=1}^{C}\|\mathbf{X}_k-\mathbf{D}\mathbf{A}_k\|_F^2+\lambda_1\sum_{k=1}^{C}\|\mathbf{A}_k\|_{1,\infty}+2\lambda_2\sum_{k=1}^{C}\mathcal{L}(\mathbf{A},\mathbf{y}_k,\mathbf{w}_k,b_k). \tag{4}$$

By fixing dictionary $\mathbf{D}$ and classifier parameter $\langle\mathbf{W},\mathbf{b}\rangle$, the coding coefficient matrix $\mathbf{A}$ can be obtained by separately calculating $\mathbf{A}_k$, $(k=1,\dots,C)$ as:

$$\langle\mathbf{A}_k\rangle=\arg\min_{\mathbf{A}_k}\|\mathbf{X}_k-\mathbf{D}\mathbf{A}_k\|_F^2+\lambda_1\|\mathbf{A}_k\|_{1,\infty}+2\lambda_2\sum_{k=1}^{C}\mathcal{L}(\mathbf{A}_k,\mathbf{y}_k,\mathbf{w}_k,b_k). \tag{5}$$

To efficiently solve Eq. (5), we adopt the commonly used augmented Lagrange multiplier (ALM) method by introducing auxiliary variable $\mathbf{A}_k'$, $(k=1,\dots,C)$, and then iteratively solved the following equivalent problem:

$$\langle\mathbf{A}_k,\mathbf{A}_k'\rangle=\arg\min_{\mathbf{A}_k,\mathbf{A}_k'}\|\mathbf{X}_k-\mathbf{D}\mathbf{A}_k\|_F^2+\lambda_1\|\mathbf{A}_k'\|_{1,\infty}+2\lambda_2\sum_{k=1}^{C}\mathcal{L}(\mathbf{A}_k,\mathbf{y}_k,)+\frac{u}{2}\|\mathbf{A}_k-\mathbf{A}_k'\|_F^2, \tag{6}$$

where u is a positive penalty parameter. We then adopted the augmented Lagrangian method to alternatively optimize $\mathbf{A}_k$ and $\mathbf{A}_k'$ until convergence as follow:

(a) First, we fix parameter $\mathbf{A}_k'$ to solve $\mathbf{A}_k$. Let $\mathbf{a}_{k_i}$ denote the coding vector of i-th sample from k-th class and $\mathbf{a}_{k_i}'$ denote the corresponding auxiliary variable, we can optimize $\mathbf{A}_k$ in columns as,

$$\langle\mathbf{a}_{k_i}\rangle=\arg\min_{\mathbf{a}_{k_i}}\|\mathbf{x}_{k_i}-\mathbf{D}\mathbf{a}_{k_i}\|_2^2+2\lambda_2\theta\sum_{k=1}^{C}\ell^2(\mathbf{a}_{k_i},\mathbf{y}_k^i)+\frac{u}{2}\|\mathbf{a}_{k_i}-\mathbf{a}_{k_i}'\|_2^2. \tag{7}$$

For computational simplicity and better smooth property, we adopt the quadratic hinge loss function $\ell^2\left(\mathbf{a}_{k_i},y_{k_i},\mathbf{w}_k,b_k\right)=\left[\max(y_k^i(\mathbf{w}_k^T\mathbf{a}_{k_i}+b_k)-1,0)\right]^2$ in Eq. (7). Thus, Eq. (7) became a least square problem with respect to $\mathbf{a}_{k_i}$ which can be easily solved.

(b) After $\mathbf{A}_k$ is fixed, we can get $\mathbf{A}_k'$ by solving the following problem:

$$\langle\mathbf{A}_k'\rangle=\arg\min_{\mathbf{A}_k'}\lambda_1\|\mathbf{A}_k'\|_{1,\infty}+\frac{u}{2}\|\mathbf{A}_k-\mathbf{A}_k'\|_F^2,\quad k=1,2,\dots,C. \tag{8}$$

Problem (8) is a $\ell_{1,\infty}$ regularized problem for joint sparse dictionary learning. In this paper, we employed the projected gradient method developed by Quattoni et al. [30] to solve it.

3.3.2. Update the dictionary ***D***

By fixing $\mathbf{A}$, $\mathbf{W}$, and $\mathbf{b}$, the optimization solution of problem of (4) with respect to $\mathbf{D}$ equal to the following problem:

$$\langle\mathbf{D}\rangle=\arg\min_{\mathbf{D}}\|\mathbf{X}-\mathbf{D}\mathbf{A}\|_F^2,\ \text{s.t.}\ \|\mathbf{d}_i\|_2^2\le 1,\quad i=1,\dots,K, \tag{9}$$

where the constraint on $\mathbf{d}_i$ is to avoid the scaling issue of dictionary atoms. Problem (9) can be solved effectively by introducing a variable $\mathbf{S}$:

$$\langle\mathbf{D}\rangle=\arg\min_{\mathbf{D}}\|\mathbf{X}-\mathbf{D}\mathbf{A}\|_F^2\ \text{s.t.}\ \mathbf{D}=\mathbf{S},\ \|\mathbf{s}_i\|_2^2\le 1,\quad i=1,\dots,K. \tag{10}$$

The optimal solution of problem (10) can be obtained by the alternating direction method of multipliers (ADMM) [28] algorithm as

$$\begin{cases}\mathbf{D}_{(t+1)}=\arg\min_{\mathbf{D}}\|\mathbf{X}-\mathbf{D}\mathbf{A}\|_F^2+\rho\|\mathbf{D}-\mathbf{S}_{(t)}+\mathbf{T}_{(t)}\|_F^2,\\ \mathbf{S}_{(t+1)}=\arg\min_{\mathbf{b}}\rho\|\mathbf{D}_{(t+1)}-\mathbf{S}+\mathbf{T}_{(t)}\|_F^2,\ \text{s.t.}\ \|\mathbf{s}_i\|_2^2\le 1,\\ \mathbf{T}_{(t+1)}=\mathbf{T}_{(t)}+\mathbf{D}_{(t+1)}-\mathbf{S}_{(t+1)},\quad \text{update }\rho\text{ if appropriate.}\end{cases} \tag{11}$$

3.3.3. Update the classifier parameters ***W*** *and* ***b***

By fixing parameters $\mathbf{D}$ and $\mathbf{A}$, we can update $\mathbf{W}$ and $\mathbf{b}$ by solving the following problem:

$$\langle\mathbf{W},\mathbf{b}\rangle=\arg\min_{\mathbf{W},\mathbf{b}}\sum_{k=1}^{C}\mathcal{L}(\mathbf{A}_k,\mathbf{y}_k,\mathbf{w}_k,b_k). \tag{12}$$

Problem (12) is actually a multiclass linear SVM problem. We adopted the SVM solver in [10] to solve it.

By repeating the above steps, we can obtain the optimal parameters $\mathbf{D}$, $\mathbf{A}$, $\mathbf{W}$, $\mathbf{b}$ of problem (4) iteratively. Algorithm 1 summarizes the optimization procedure of above parameters.

Algorithm 1. Optimization procedure of CRO-DDL model.

Require: Train samples $\mathbf{X}$, Class label $\mathbf{Y},\lambda_1,\lambda_2,\theta,u$.
while not converge **do**
  for $k=1$ **to** C **do**
    for $i=1$ **to** n_k **do**
      $\mathbf{a}_{k_i}\leftarrow\arg\min_{\mathbf{a}_{k_i}}\|\mathbf{x}_{k_i}-\mathbf{D}\mathbf{a}_{k_i}\|_2^2+2\lambda_2\theta\sum_{k=1}^{C}\ell^2(\mathbf{a}_{k_i},\mathbf{y}_k^i)+\frac{u}{2}\|\mathbf{a}_{k_i}-\mathbf{a}_{k_i}'\|_2^2$
    end for
    $\mathbf{A}_k'\leftarrow\arg\min_{\mathbf{A}_k'}\lambda_1\|\mathbf{A}_k'\|_{1,\infty}+\frac{u}{2}\|\mathbf{A}_k-\mathbf{A}_k'\|_F^2$
  end for
  $\mathbf{D}\leftarrow\arg\min_{\mathbf{D}}\|\mathbf{X}-\mathbf{D}\mathbf{A}\|_F^2$, s.t. $\|\mathbf{d}_i\|_2^2\le 1, i=1,\dots,K$,
  for $k=1$ **to** C **do**
    Update $\mathbf{w}_k,b_k$ by solving Eq. (12) with multi-class SVM solver in [10].
  end for
end while
Output: dictionary $\mathbf{D}$, coding matrix $\mathbf{A}$, classifier parameters $\mathbf{W}$ and $\mathbf{b}$.

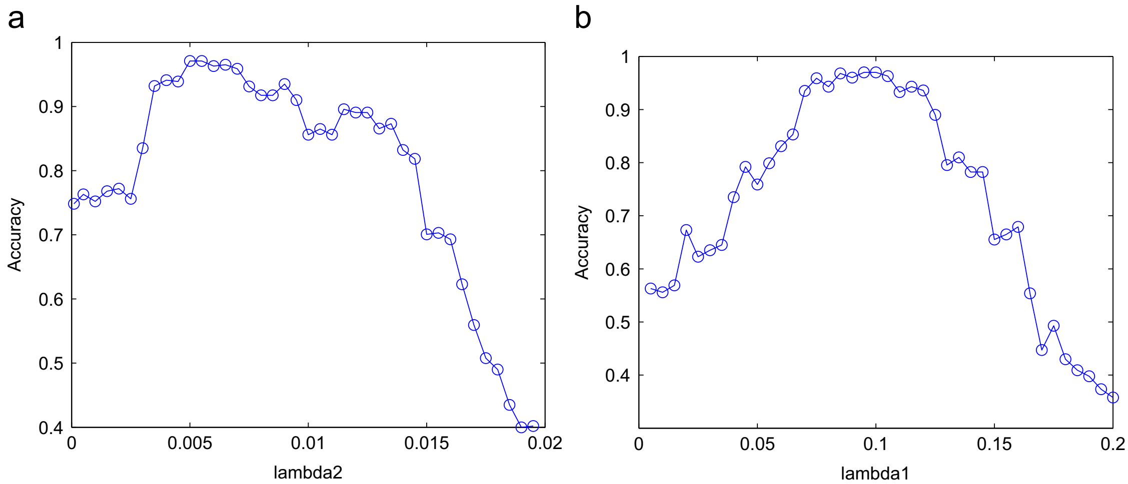

Fig. 3. Classification accuracy vs. (a) λ_1 and (b) λ_2, respectively. In (a) λ_1 is fixed to 0.1 and in (b) λ_2 is fixed to 0.005.

Table 1
The recognition rates of competing methods on Extended Yale B dataset.

SRC [9]	DKSVD [17]	LC-KSVD [36]	FDDL [20]	SVGDL [22]	CRO-DDL
0.9	0.753	0.906	0.919	0.961	**0.971**

Fig. 4. Sample illustrations from AR dataset.

Table 2
The face recognition rates of competing methods on AR dataset.

SRC [9]	DKSVD [17]	LC-KSVD [36]	FDDL [20]	SVGDL [22]	CRO-DDL
0.888	0.854	0.897	0.920	0.946	**0.949**

Table 3
The gender recognition rates of competing methods on AR dataset.

SRC [9]	DKSVD [17]	LC-KSVD [36]	FDDL [20]	SVGDL [22]	CRO-DDL
0.93	0.861	0.86	**0.954**	0.943	0.951

3.4. Classification rule of CRO-DDL

Once the dictionary $\mathbf{D}$ and the classifier parameters $\langle \mathbf{W}, \mathbf{b} \rangle$ are learned, we perform classification as follows: for a test sample $\mathbf{x}$, we first calculate the sparse coding vector. As $\ell_{1,\infty}$-norm is a matrix norm and thus cannot tackle a single vector, we use ℓ_1-norm regularization to get the coding vector, resulting in the following tractable problem [31,39]:

$$\langle \mathbf{a}^* \rangle = \arg\min_{\mathbf{a}} \| \mathbf{x} - \mathbf{D}\mathbf{a} \|_2^2 + \lambda_1 \| \mathbf{a} \|_1. \tag{13}$$

After obtained the coding vector $\mathbf{a}^*$, we can directly use the SVM classifier for classification as

$$identity(\mathbf{x}) = \arg\max_k \mathbf{w}_k^T \mathbf{a}^* + b_k, \quad k = 1, \ldots, C. \tag{14}$$

4. Experimental results

In this section, we experimentally evaluated the proposed CRO-DDL algorithm for objection recognition, action recognition, and face recognition on four publicly available datasets, i.e., Extended Yale B dataset [34], AR dataset [33], Caltech-101 object dataset [32], and UCF sports action dataset [35], respectively. We compared the performance of the proposed model with the sparse representation based classification (SRC) method [9], K-SVD [12] and state-of-the-art DDL methods including DKSVD [17], LC-KSVD [36], FDDL [20] and SVGDL [22].

4.1. Experiments on face recognition

We applied our algorithm for face recognition problems on two public datasets, i.e., Extended Yale B [34] and AR [33]. We will show the experimental results in the following subsections.

4.1.1. Face recognition on Extended Yale B

We test the proposed method on the Extended Yale B database [34], which consists of 2414 frontal face images from 38 persons.

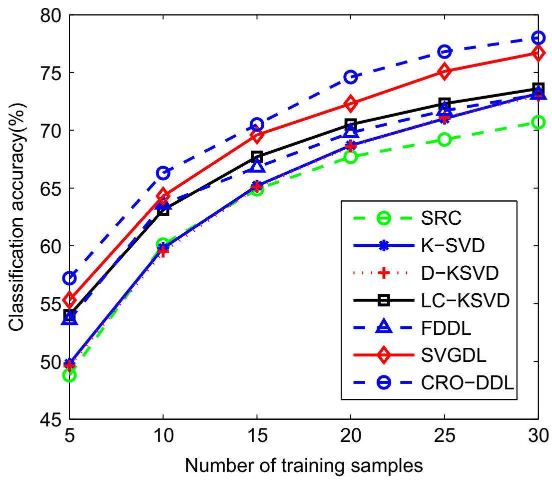

Fig. 5. Comparison results on Caltech-101 with different training samples.

All the images were cropped into the size of 54×48. Fig. 2 shows some samples from the Extended Yale B dataset.

Following the experimental setting in [22] for comparison, we randomly selected 20 images of each person for training and the rest were used for testing. In this experiment, the number of the dictionary atoms was 380.

As there are four parameters λ_1, λ_2, u, and θ in the proposed CRO-DLL method, we first analyze their sensitivity. For simplicity, we set $\theta = 0.2$ as [22], $u = \lambda_2$, and tune λ_1 and λ_2. Fig. 3 shows the variation of classification accuracy along with the parameters. In Fig. 3, we first set $\lambda_1 = 0.1$ and change λ_2 to see the variation of classification rate. The result shows the proposed method achieved the best accuracy when $\lambda_2 = 0.005$. Next, we set $\lambda_2 = 0.005$ and tune the parameter λ_1. Fig. 3(b) shows that, with $\lambda_1 = 0.1$, we can get satisfactory classification rate. Thus, for face recognition on Extended Yale B dataset, λ_1 and λ_2 were set to 0.1 and 0.005, respectively.

The corresponding classification accuracy is shown in Table 1. For comparison, Table 1 also presents the classification rates of SRC [9], DKSVD [17], LC-KSVD [36], FDDL [20], and SVGDL [22]. As one can see that the proposed CRO-DDL achieved better recognition accuracy than all the competing methods.

4.1.2. Face recognition on AR dataset

We further test our method on the AR dataset [33] for Face recognition. The dataset is consists of over 4000 face images collected from 126 individuals. Following the experimental setting in [33], we used a subset containing 1400 face images from 50 females and 50 males subjects in the experiment. The face images were resized to 60×43 as shown in Fig. 4. For each subject, we selected the 7 images with illumination and expression changes from Session 1 for training, and the other 7 images with the same condition from Session 2 for testing. The dictionary size was set to 500 in the experiment. Parameters were set as $\lambda_1 = 0.2, \lambda_2 = 0.002, u = 0.002$, and $\theta = 0.2$, respectively.

We evaluated our approach and compared with SRC [9], DKSVD [17], LC-KSVD [36], FDDL [20], and SVGDL [22]. Experimental results of different methods are summarized in Table 2. The recognition accuracy of proposed CRO-DDL is 0.949, which has improvement over all the competing methods. SVGDL performs the second best in this experiment, followed by FDDL and LC-KSVD.

4.1.3. Gender recognition on AR dataset

We also evaluate the proposed method for gender classification. We selected a subset of 25 males and 25 females to construct the training dataset, and select other 25 males and 25 females to

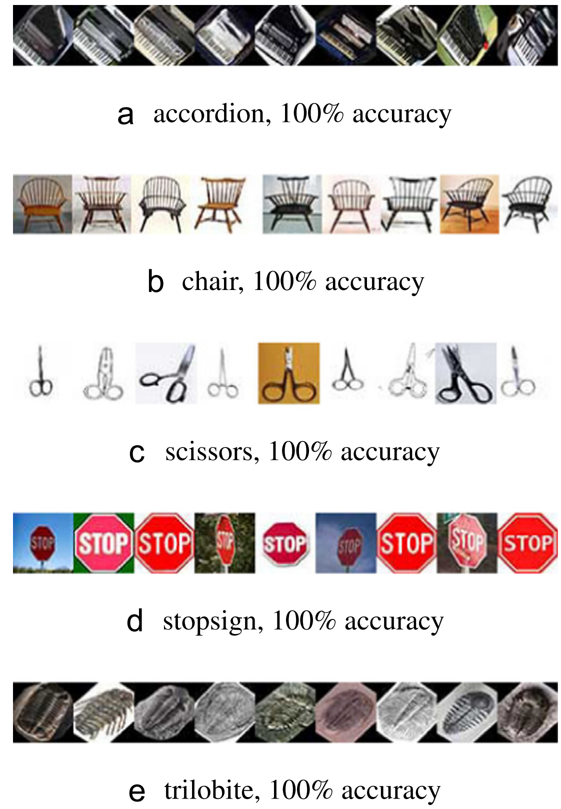

Fig. 6. Example images of classes with high classification accuracy from Caltech-101: (a) accordion, 100% accuracy; (b) chair, 100% accuracy; (c) scissors, 100% accuracy; (d) stopsign, 100% accuracy; (e) trilobite, 100% accuracy.

construct the test dataset. The images of the training dataset and test dataset has no-overlap. For each image, we used PCA to reduce its dimension to 300. The dictionary size was fixed to 500. Parameters were set as $\lambda_1 = 0.3, \lambda_2 = 0.005, u = 0.005$, and $\theta = 0.2$, respectively.

We compared our method with SRC [9], DKSVD [17], LC-KSVD [36], FDDL [20], SVGDL [22], and the comparison results are shown in Table 3. We can see that our method performs the second best classification rate, which is slightly lower than that of FDDL. This maybe because that, in gender classification, there are enough training samples for each class (only two classes). Thus, in FDDL, the learned dictionary of each class is representative enough to represent the test sample.

4.2. Objection classification on Caltech-101 dataset

The Caltech-101 object dataset [32] contains 9144 images from 102 object classes, i.e., 101 object classes and a background class. The contents of images includes flowers, trees, animals, and vehicles. The samples from each class have significant shape variability. For each class, its number varies from 31 to 800. For fairly comparison, we followed the experimental setting as [22]. For each category we randomly selected 5, 10, 15, 20, 25 and 30 images for training, used the rest for testing. The number of dictionary atoms was set to be 510 in all the cases. We repeated the experiments 10 times with different random splits of the training and testing images. The final recognition rates were reported as the average of each run.

We evaluated the effectiveness of proposed CRO-DDL on Caltech-101 by comparing it with six state-of-the-art sparse representation methods, i.e., SRC [9], K-SVD [12], DKSVD [17], LC-KSVD [36], FDDL [20] and SVGDL [22]. For observation convenience, the comparison results are shown in Fig. 5.

Fig. 7. Sample illustrations from UCF sports action dataset.

Table 4
Confusion matrix of CRO-DDL on UCF sport action dataset.

		Predicted									
		Diving	Golf	High-bar	Kicing	Lifting	Swing	Riding	Running	Skating	Walking
Actual	Diving	0.83	0	0	0	0	0.17	0	0	0	0
	Golf	0	0.94	0	0	0	0.06	0	0	0	0
	High-bar	0	0.05	0.90	0	0	0.05	0	0	0	0
	Kicking	0	0	0	1.00	0	0	0	0	0	0
	Lifting	0	0	0	0	0.83	0	0	0	0.17	0
	Swing	0	0	0	0	0	1.00	0	0	0	0
	Riding	0	0	0	0	0	0	1.00	0	0	0
	Running	0	0	0	0	0	0	0	1.00	0	0
	Skating	0	0	0	0	0	0	0	0	1.00	0
	Walking	0	0	0	0	0	0.04	0	0	0	0.96

Table 5
The recognition rates of competing methods on UCF sports action dataset.

SRC [9]	DKSVD [17]	LC-KSVD [36]	FDDL [20]	SVGDL [22]	CRO-DDL
0.929	0.881	0.912	0.943	0.944	**0.968**

As shown in Fig. 5, SRC achieved the worst accuracy which was possibly attributed to the lack of discriminative dictionary learning. With class-specific dictionaries, FDDL outperformed K-SVD and LC-KSVD. However, when the training number was high (about 25, 30) per class, there was no significant gain over K-SVD and LC-KSVD. By learning a discriminative dictionary under the guidance of SVM, SVGDL had a better classification accuracy than LC-KSVD and FDDL. With the possible learned cross class relatedness, our method outperformed all the competing methods. Fig. 6 shows some examples from the classes with high classification accuracy.

4.3. Action recognition on UCF sports action dataset

We evaluated the proposed method on the UCF sports action dataset [35] which has been widely used for numerous applications, such as action recognition, action localization, and saliency detection. The dataset consists of a set of actions collected from various sports from broadcast television channels, such as the BBC and ESPN, including a total of 150 sequences of 10 action classes, such as diving, golf, kicking, lifting, horse riding, running, skateboarding, swinging (pommel horse), swinging (high bar), and walking. Fig. 7 shows different class samples of UCF sports action dataset.

Following the common experimental settings in [22], we evaluated our approach by using the five-fold cross validation, in which four folds of samples were used for training, and the rest samples were used for test. The dictionary size was set to 50. Parameters were set as $\lambda_1 = 0.2, \lambda_2 = 0.02, u = 0.02$, and $\theta = 0.2$, respectively. The confusion matrix of proposed method on UCF sports action dataset is shown in Table 4. It can be seen that our proposed method receives 100% correct on 5 classes.

For comparison, we list the classification accuracies of our CRO-DDL and the comparison methods, including SRC [7], DKSVD [17], LC-KSVD [36], FDDL [20], and SVGDL [22] in Table 5. Our proposed method has the best classification result. FDDL and SVGDL have very comparable results, i.e., 0.943 vs. 0.944.

5. Conclusions

In this paper, based on the assumption that relevance between dictionary atoms of different class labels could be helpful in multiclass classification task, we proposed a new dictionary method, CRO-DDL algorithm, for sparse representation. Utilizing the $\ell_{1,\infty}$-norm regularization on the coding matrix of each class, the proposed method can adaptively builds the relevance between dictionary atoms and the class labels, by learning a class related dictionary and a SVM classifier simultaneously. Experimental results on face recognition, object classification, and action classification demonstrate that our method is comparable to many state-of-the-art DDL methods in multi-class image classification.

Conflict of interest

None declared.

Acknowledgements

This work is partially supported by the NSFC funds of China under Contract Nos. 61271093, 61401125, 61471146, Science and Technology Plan of Guangdong Province, contract no. 2015B010128009 and National High Technology Research and Development Program of China (863 program) under Contract No. 2013AA013801.

References

[1] L. Lin, X. Wang, W. Yang, J. Lai, Discriminatively trained and-or graph models for object shape detection, IEEE Trans. Pattern Anal. Mach. Intell. (T-PAMI) 37 (5) (2015) 959–972.
[2] J. Mairal, M. Elad, G. Sapiro, Sparse representation for color image restoration, IEEE Trans. Image Process. 17 (1) (2008) 53–69.
[3] M. Ranzato, C. Poultney, S. Chopra, Y. Cun, Efficient learning of sparse representations with an energy-based model, in: Proceedings of the Conference on Neural Information Processing Systems, 2006.
[4] L. Lin, X. Liu, S. Peng, H. Cao, Y. Wang, B. Jiang, Object categorization with sketch representation and generalized samples, Pattern Recognit. 45 (10) (2012) 3648–3660.
[5] J. Mairal, F. Bach, J. Ponce, G. Sapiro, Online learning for matrix factorization and sparse coding, J. Mach. Learn. Res. 11 (2010) 19–60.
[6] L. Lin, P. Luo, X. Chen, K. Zeng, Representing and recognizing objects with massive local image patches, Pattern Recognit. 45 (1) (2012) 231–240.
[7] J. Yang, J. Wright, T. Huang, Y. Ma, Image super-resolution as sparse representation of raw patches, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008.
[8] M. Elad, M. Aharon, Image denosing via sparse and redundant representations over learned dictionaries, IEEE Trans. Image Process. 54 (December (12)) (2006) 3736–3745.
[9] J. Wright, M. Yang, A. Ganesh, S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (February (2)) (2009) 210–227.
[10] J. Yang, K. Yu, Y. Gong, T. Huang, Linear spatial pyramid matching using sparse coding for image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009.
[11] B.A. Olshausen, D.J. Field, Sparse coding with an overcomplete basis set: a strategy employed by v1?, Vis. Res. 37 (1997) 3311–3325.
[12] M. Aharon, M. Elad, A. Bruckstein, K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation, IEEE Trans. Signal Process. 54 (November (1)) (2006) 4311–4322.
[13] R. Raina, A. Battle, H. Lee, B. Packer, A.Y. Ng, Self-taught learning: transfer learning from unlabeled data, in: Proceedings of the Conference on Machine Learning, 2007.
[14] K. Kavukcuoglu, M. Ranzato, R. Fergus, Y. LeCun, Learning invariant features through topographic filter maps, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009.
[15] J. Mairal, F. Bach, J. Ponce, G. Sapiro, A. Zisserman, Discriminative learned dictionaries for local image analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008.
[16] D. Pham, S. Venkatesh, Joint learning and dictionary construction for pattern recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008.
[17] Q. Zhang, B. Li, Discriminative K-SVD for dictionary learning in face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010.
[18] I. Ramirez, P. Sprechmann, G. Sapiro, Classification and clustering via dictionary learning with structured incoherence and shared features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 3501–3508.
[19] I. Tosic, P. Frossard, Dictionary learning, what is the right representation for my signal? IEEE Signal Process. Mag. 28 (2011) 27–38.
[20] M. Yang, L. Zhang, X. Feng, D. Zhang, Fisher discrimination dictionary learning for sparse representation, in: Proceedings of the IEEE Conference on Computer Vision, 2011, pp. 543–550.
[21] J. Mairal, F. Bach, J. Ponce, Task-driven dictionary learning, IEEE Trans. Pattern Anal. Mach. Intell. (2012) 791–804.
[22] S. Cai, W. Zuo, L. Zhang, X. Feng, P. Wang, Support vector guided dictionary learning, in: European Conference on Computer Vision, vol. 8692, 2014, pp. 624–639.
[23] A. Castrodad, G. Sapiro, Sparse modeling of human actions from motion imagery, Int. J. Comput. Vis. (2012) 1–15.
[24] M. Fornasier, H. Rauhut, Recovery algorithm for vector-valued data with joint sparsity constraints, SIAM J. Numer. Anal. 46 (2) (2008) 577–613.
[25] H. Liu, M. Palatucci, J. Zhang, Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery, in: Proceedings of the IEEE Conference on Machine Learning, 2009, pp. 649–656.
[26] J. Zhang, A probabilistic framework for multi-task learning, Technical Report CMU-LTI-06-006, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 2006.
[27] N. Zhou, Y. Shen, J. Peng, J. Fan, Learning inter-related visual dictionary for object recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012.
[28] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Found. Trends in Mach. Learn. 3 (1) (2011) 1–122.
[29] S. Gao, I.W.H. Tsang, L.T. Chia, Laplacian sparse coding, hypergraph Laplacian sparse coding, and applications, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 92–104.
[30] A. Quattoni, X. Carreras, M. Collins, T. Darrell, An efficient projection for l1,∞, regularization, in: Proceedings of the IEEE Conference on Machine Learning, vol. 382, 2009.
[31] W. Zuo, D. Meng, L. Zhang, X. Feng, D. Zhang, A generalized iterated shrinkage algorithm for non-convex sparse coding, in: Proceedings of the IEEE Conference on Computer Vision, 2013, pp. 217–224.
[32] F. Li, R. Fergus, P. Perona, Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories, in: Proceedings of the IEEE Conference on Computer Society Conference on Computer Vision and Pattern Recognition, 2004, pp. 178–186.
[33] A. Martinez, R. Benavente, The AR face database, CVC Technical Report, 1998.
[34] K.C. Lee, J. Ho, D. Kriegman, Acquiring linear subspaces for face recognition under variable lighting, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 684–698.
[35] M.D. Rodriguez, J. Ahmed, M. Shah, Action MACH: a spatio-temporal maximum average correlation height filter for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008.
[36] Z. Jiang, Z. Lin, L.S. Davis, Label consistent K-SVD: learning a discriminative dictionary for recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (11) (2013) 2651–2664.
[37] S. Gao, I.W.H. Tsang, Y. Ma, Learning category-specific dictionary and shared dictionary for fine-grained image categorization, IEEE Trans. Image Process. 23 (2) (2014) 623–634.
[38] M. Yang, L. Zhang, X. Feng, D. Zhang, Sparse representation based fisher discrimination dictionary learning for image classification, Int. J. Comput. Vis. 109 (September (3)) (2014) 209–232.
[39] David L. Donoho, Y. Tsaig, Fast Solution of ℓ_1-norm minimization problems when the solution may be sparse, IEEE Trans. Inf. Theory 54 (11) (2008) 4789–4812.
[40] Mairal Julien, Bach Francis, Ponce Jean, Sapiro Guillermo, Online learning for matrix factorization and sparse coding, J. Mach. Learn. Res. 11 (2010) 16–19.
[41] F. Alhussein, D. Mike, F. Pascal, Dictionary learning for fast classification based on soft-thresholding, Int. J. Comput. Vis. (2014) 1–16.

Dongyu Zhang received his B.S. and Ph.D. degree from Northeastern University and Harbin Institute of Technology, China, in 2003 and 2010, respectively. His current research interests include image restoration, image segmentation, sparse representation and metric learning.

Pengju Liu received his B.S. degree from Northeast Forestry University, China, in 2014. He is working towards his M.S. degree in Harbin Institute of Technology, China. His major research interests include image analysis and sparse representation.

Kai Zhang received the M.Sc. degree in applied mathematics from China Jiliang University, Hangzhou, China. He is currently pursuing the Ph.D. degree from Harbin Institute of Technology, Harbin, China. His research interests include machine learning, image processing and pattern recognition.

Hongzhi Zhang currently is an Associate Professor in School of Computer Science and Technology, Harbin Institute of Technology, China. He received the Ph.D. degree in Computer Science and Technology from the Harbin Institute of Technology in 2007. His current research interests include pattern recognition, computer vision, and their applications in medical diagnosis.

Qing Wang currently is an Associate professor in School of Software of Sun Yat-sen University. His current research interests include human computer interaction, user experience, collaborative software, and interoperability of enterprise software and application.

Xiaoyuan Jing received the M.Sc. and Ph.D. degrees in pattern recognition and intelligence systems from the Nanjing University of Science and Technology, Nanjing, China, in 1995 and 1998, respectively. He is currently a Professor and Doctoral Supervisor with the State Key Laboratory of Software Engineering of Wuhan University, Wuhan, China. His current research interests include pattern recognition, machine learning, data mining, image processing, neural networks, and artificial intelligence.