Multimed Tools Appl (2016) 75:8861–8874
DOI 10.1007/s11042-014-2257-y

Elastic net regularized dictionary learning for image classification

Bin Shen · Bao-Di Liu · Qifan Wang

Received: 3 April 2014 / Revised: 12 July 2014 / Accepted: 26 August 2014 /
Published online: 3 October 2014

Abstract Dictionary learning plays a key role in image representation for classification. A multi-modal dictionary is usually learned from feature samples across different classes and shared in the feature encoding process. Ideally each atom in dictionary corresponds to a single class of images, while each class of images corresponds to a certain group of atoms. Image features are encoded as linear combinations of selected atoms in a given dictionary. We propose to use elastic net as regularizer to select atoms in feature coding and related dictionary learning process, which not only benefits from the sparsity similar as ℓ_1 penalty but also encourages a grouping effect that helps improve image representation. Experimental results of image classification on benchmark datasets show that with dictionary learned in the proposed way outperforms state-of-the-art dictionary learning algorithms.

Keywords Dictionary learning · Elastic net regularization · Image classification

1 Introduction

With the increasing popularity of information technologies such as digital imaging and social network, there are huge amount of multimedia resources available in our daily life. The related tasks [7–9, 25] including management, processing and mining become critical issues to researchers and engineers. Image classification, which aims to automatically associate images with semantic labels, is one of the fundamental components for many of these tasks and paves the way for various types of following processing. The most common framework for image classification is the discriminative model [10, 15–18, 29, 30].

B. Shen (✉) · Q. Wang
Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA
e-mail: bshen@purdue.edu

B.-D. Liu (✉)
College of Information and Control Engineering, China University of Petroleum,
Qingdao 266580, China
e-mail: thu.liubaodi@gmail.com

It contains five main steps, including feature extraction, dictionary learning, image feature coding, image pooling, and SVM-based classification [22]. Dictionary learning plays a key role in image representation for various vision tasks [23, 31]. A dictionary is usually composed of visual words, which encode low level visual information of images across different classes. The primitive versions of vocabulary learning are typically K-means clustering on image patches combined with hard- or soft-assignment vector quantization (VQ) [27]. Spatial pyramid matching (SPM) is typically incorporated in the pipeline to compensate the loss of spatial information [10]. Yang et al. [30] introduced sparse representation algorithm for learning dictionary and coding images based on SPM, resulting in state-of-the-art performance in image classification.

Nowadays, various dictionary learning algorithms are emerged. First, more and more researchers focused on locality-preserving dictionary learning algorithms due that locality was more essential than sparsity [32]. Wang et al. [28] considered that each word in the vocabulary lied on a manifold, and utilized locally linear coding [21] for vector quantization to preserve the local information on vocabulary. Gao et al. [6] incorporated the histogram intersection kernel based laplacian matrix into the objective function of sparse coding to enforce the consistence in sparse representation of similar local features. Zheng et al. [34] explicitly considered the vector quantization based laplacian matrix in the objective function of sparse coding, and Lu et al. [19] considered the hypergraph (vertex, hyperedge, incidence matrix and hyperedge weights) regularization. Data points are assumed to be distributed on the a manifold in [20] and a manifold projection is used there to improve traditional sparse coding. In [24], sparse representation algorithm is adopted to construct the graph model which is embedded into non-negative matrix factorization algorithm. In [14], sparse representation in k-nearest neighbor to construct the graph model to improve the accuracy and robustness of image representation. Second, the learned dictionary becomes versatile with the introduction of the kernel technique. For the K-means based scheme, [29] learned dictionary in the histogram intersection kernel (HIK) space, while [27] learned it in the Gaussian radial basis function (RBF) kernel space. For the sparse representation based scheme, [26] proposed kernel K-SVD and kernel MOD methods. Gao et al. [4] proposed kernel sparse representation (KSR), and learned the dictionary directly in the Gaussian RBF kernel space using a fixed point iteration method. It generally outperforms the previous alternative extensions of sparse representation for image classification and face recognition.

On the one hand, traditional dictionary learning algorithms with ℓ_1 regularizer have achieved state-of-the-art performance for image classification. However, when a group of atoms in the dictionary are highly correlated, the ℓ_1 regularizer tends to select one atom from the group and ignore the other atoms. On the other hand, since images from different classes usually have different appearances, how to tackle the multi-modality in visual appearance of different classes becomes a problem urgent to solve. In practice, a dictionary includes all atoms for different classes, and, to resolve the multi-modality issue, a group of atoms are later selected from a given dictionary to encode image feature. Specifically, image features are approximated by a linear combination of atoms selected to result in compact sparse representation, either by nearest neighbor search or squared loss minimization with ℓ_1 norm regularization. Empirically, sparse representation by ℓ_1 norm regularizer achieves higher performance in image representation for classification at the cost of extra computational resource. However, ℓ_1 regularizer aims to sparsify the solution while ignores group effect. Note that ideally each atom in dictionary corresponds to a single class of images, while each class of images corresponds to a certain group of atoms. Inspired by [35], instead of use ℓ_1 regularizer, we use elastic net as regularizer in dictionary learning. Elastic net not only enjoys the sparsity as ℓ_1 regularizer, but also encourages a grouping effect in atom selection, which benefits the image feature encoding process considering the multi-modality in visual appearance of images from different classes.

The rest of the paper are organized as follows. Section 2 proposed Elastic Net regularized Dictionary Learning algorithm (ENDL). The solution to the minimization of the objective function and guaranteed convergence are elaborated in Section 3. Then, experimental results and analysis are shown in Section 4. Finally, discussions and conclusions are drawn in Section 5.

2 Proposed elastic net regularized dictionary learning

Let $\boldsymbol{X} \in \mathbb{R}^{D\times N}$ be the input data matrix, where D and N are the dimension and number of the data vectors, respectively. Let $\boldsymbol{B} \in \mathbb{R}^{D\times K}$ and $\boldsymbol{S} \in \mathbb{R}^{K\times N}$ denote the basis matrix and corresponding sparse codes (also called coefficient matrix), respectively, where K is the number of the bases. ℓ_1 regularizer based dictionary learning algorithm aims to solve the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{B},\boldsymbol{S}} f(\boldsymbol{B},\boldsymbol{S}) &= \|\boldsymbol{X}-\boldsymbol{BS}\|_F^2 + 2\alpha\|\boldsymbol{S}\|_1 \\ & s.t.\ \|\boldsymbol{B}_{\bullet i}\|_2 \le 1,\ \forall i = 1,2,\ldots,K. \end{aligned} \tag{1}$$

Here, $\boldsymbol{B}_{\bullet i}$ and $\boldsymbol{B}_{k\bullet}$ denotes the i-th column and k-th row vectors of matrix $\boldsymbol{B}$, respectively. The ℓ_1 norm regularization term is adopted to enforce sparsity of $\boldsymbol{S}$ and α is the regularization parameter to control the tradeoff between fitting goodness and sparseness.

ℓ_1 regularizer based dictionary learning algorithm has achieved superior performance for image classification. However, for each image feature, ℓ_1 regularizer ignores group effect in atom selection, which means that the image feature could not finds the atoms which tends to correspond to same class of images, while the atoms could not capture a notion of the same class. Inspired by the advantages of ℓ_2 norm regularizer [33], elastic net regularized dictionary learning algorithm is proposed and the objective function of the algorithm is as follows,

$$\begin{aligned} \min_{\boldsymbol{B},\boldsymbol{S}} f(\boldsymbol{B},\boldsymbol{S}) &= \|\boldsymbol{X}-\boldsymbol{BS}\|_F^2 + 2\alpha\|\boldsymbol{S}\|_1 + \beta\|\boldsymbol{S}\|_F^2 \\ & s.t.\ \|\boldsymbol{B}_{\bullet i}\|_2 \le 1,\ \forall i = 1,2,\ldots,K. \end{aligned} \tag{2}$$

There are several advantages by this elastic net regularizer. First, the ℓ_2 norm regularizer helps remove the limitation on the number of selected atoms from dictionary; second, it encourages grouping effect, which makes the image feature finds the atoms which tends to correspond to same class of images; third, it also stabilize the ℓ_1 norm regularization path.

3 Optimization of the objective function

In this section, we focus on solving the optimization of the objective function proposed in the last section. Similar to [11], this optimization problem is not jointly convex in both $\boldsymbol{B}$ and $\boldsymbol{S}$, while it is separately convex in either $\boldsymbol{B}$ or $\boldsymbol{S}$ with $\boldsymbol{S}$ or $\boldsymbol{B}$ fixed. So the objective function can be optimized by alternating minimization to two optimization subproblems as follows.

- Fixed $\boldsymbol{B}$, the objective function of finding sparse codes $\boldsymbol{S}$ can be written as an Elastic Net regularized Least Square ($EN-LS$) minimization subproblem:

$$f(\boldsymbol{S}) = \|(\boldsymbol{X}) - \boldsymbol{BS}\|_F^2 + 2\alpha\|\boldsymbol{S}\|_1 + \beta\|\boldsymbol{S}\|_F^2 \tag{3}$$

- Fixed $\boldsymbol{S}$, the objective function of learning weight $\boldsymbol{B}$ can be written as an ℓ_2 norm constrained least square ($L2-LS$) minimization subproblem:

$$\begin{aligned} f(\boldsymbol{B}) &= \|(\boldsymbol{X}) - \boldsymbol{BS}\|_F^2 \\ &\quad s.t.\ \|\boldsymbol{B}_{\bullet k}\|_2^2 \le 1,\ \forall k = 1, 2, \cdots, K. \end{aligned} \tag{4}$$

3.1 $EN-LS$ minimization subproblem

The (3) can be simplified as

$$\begin{aligned} f(\boldsymbol{S}) &= \|\boldsymbol{X} - \boldsymbol{BS}\|_F^2 + 2\alpha\|\boldsymbol{S}\|_1 + \beta\|\boldsymbol{S}\|_F^2 \\ &= tr\left\{\boldsymbol{X}^T\boldsymbol{X} - 2\boldsymbol{X}^T\boldsymbol{BS} + \boldsymbol{S}^T\boldsymbol{B}^T\boldsymbol{BS}\right\} + 2\alpha\|\boldsymbol{S}\|_1 + \beta\|\boldsymbol{S}\|_F^2 \\ &= tr\left\{\boldsymbol{X}^T\boldsymbol{X}\right\} - 2\sum_{n=1}^{N}[\boldsymbol{X}^T\boldsymbol{B}]_{n\bullet}\boldsymbol{S}_{\bullet n} + \sum_{n=1}^{N}\boldsymbol{S}_{\bullet n}^T\boldsymbol{B}^T\boldsymbol{BS}_{\bullet n} \\ &\quad +2\alpha\sum_{k=1}^{K}\sum_{n=1}^{N}|\boldsymbol{S}_{kn}| + \beta\sum_{k=1}^{K}\sum_{n=1}^{N}\boldsymbol{S}_{kn}^2, \end{aligned} \tag{5}$$

where $tr\left\{\boldsymbol{X}^T\boldsymbol{X}\right\}$ represents the trace of matrix $\{\boldsymbol{X}^T\boldsymbol{X}\}$.

Ignoring the constant term $tr\left\{\boldsymbol{X}^T\boldsymbol{X}\right\}$, the objective function of $\boldsymbol{S}_{\bullet n}$ reduces to (6) with $\boldsymbol{B}$ fixed.

$$f(\boldsymbol{S}_{\bullet n}) = \boldsymbol{S}_{\bullet n}^T\boldsymbol{B}^T\boldsymbol{BS}_{\bullet n} - 2[\boldsymbol{X}^T\boldsymbol{B}]_{n\bullet}\boldsymbol{S}_{\bullet n} + 2\alpha||\boldsymbol{S}_{\bullet n}||_1 + \beta||\boldsymbol{S}_{\bullet n}||_2^2. \tag{6}$$

And then the objective function of $\boldsymbol{S}_{kn}$ in (6) reduces to (7) with $\boldsymbol{B}$ and $\{\boldsymbol{S}_{1n}, \boldsymbol{S}_{2n}, \ldots, \boldsymbol{S}_{kn}\}/\boldsymbol{S}_{kn}$ fixed.

$$\begin{aligned} f(\boldsymbol{S}_{kn}) &= \boldsymbol{S}_{kn}^2[\boldsymbol{B}^T\boldsymbol{B}]_{kk} + 2\alpha\,|\boldsymbol{S}_{kn}| + \beta\boldsymbol{S}_{kn}^2 \\ &\quad +2\boldsymbol{S}_{kn}\left\{\sum_{l=1,l\neq k}^{K}[\boldsymbol{B}^T\boldsymbol{B}]_{kl}\boldsymbol{S}_{ln} - [\boldsymbol{B}^T\boldsymbol{X}]_{kn}\right\} \\ &= \boldsymbol{S}_{kn}^2\left([\boldsymbol{B}^T\boldsymbol{B}]_{kk} + \beta\right) + 2\alpha\,|\boldsymbol{S}_{kn}| - 2\boldsymbol{S}_{kn}\boldsymbol{H}_{kn}, \end{aligned} \tag{7}$$

where $\boldsymbol{H}_{kn} = [\boldsymbol{B}^T\boldsymbol{X}]_{kn} - \sum_{l=1,l\neq k}^{K}[\boldsymbol{B}^T\boldsymbol{B}]_{kl}\boldsymbol{S}_{ln}$.

When $\|\boldsymbol{B}_{\bullet k}\|_1 > 0$, $f(\boldsymbol{S}_{kn})$ is piece-wise parabolic function with $[\boldsymbol{B}^T\boldsymbol{B}]_{kk} + \beta = 1 + \beta$.

Based on the convexity and monotonic property of the parabolic function, it is not difficult to know that $f(\boldsymbol{S}_{kn})$ reaches the minimum at the unique point.

$$\boldsymbol{S}_{kn} = \{\max\{\boldsymbol{H}_{kn}, \alpha\} + \min\{\boldsymbol{H}_{kn}, -\alpha\}\} / (1+\beta)\,. \tag{8}$$

Furthermore,given that the optimal value for $\boldsymbol{S}_{kn}$ does not depend on the other entries in the same row, each whole row of $\boldsymbol{S}$ can be optimized simultaneously. That is

$$\boldsymbol{S}_{k\bullet} = \{\max\{\boldsymbol{H}_{k\bullet}, \alpha\}+\min\{\boldsymbol{H}_{k\bullet}, -\alpha\}\} / (1+\beta)\,. \tag{9}$$

3.2 $L2 - LS$ minimization subproblem

Without the Elastic Net regularization term in (3) and additional constraints in (4), $\boldsymbol{S}_{k\bullet}$ and $\boldsymbol{B}_{\bullet k}$ are dual in objective function $\|\boldsymbol{X} - \boldsymbol{B}\boldsymbol{S}\|_F^2$ for $\forall k \in \{1, 2, \ldots, K\}$.

With $\boldsymbol{S}$ fixed, the objective function of $L2$-LS minimization subproblem can be simplified as

$$\begin{aligned} f(\boldsymbol{B}) &= \|\boldsymbol{X} - \boldsymbol{B}\boldsymbol{S}\|_F^2 \\ &= tr\left\{\boldsymbol{X}^T\boldsymbol{X} - 2\boldsymbol{S}\boldsymbol{X}^T\boldsymbol{B} + \boldsymbol{S}\boldsymbol{S}^T\boldsymbol{B}^T\boldsymbol{B}\right\} \\ &= tr\left\{\boldsymbol{X}^T\boldsymbol{X}\right\} - 2\sum_{k=1}^{K}[\boldsymbol{S}\boldsymbol{X}^T]_{k\bullet}\boldsymbol{B}_{\bullet k} + \sum_{k=1}^{K}[\boldsymbol{S}\boldsymbol{S}^T\boldsymbol{B}^T]_{k\bullet}\boldsymbol{B}_{\bullet k} \\ s.t.\ \ &\|\boldsymbol{B}_{\bullet k}\|_2^2 \le 1,\ \ \forall k = 1, 2, \cdots, K. \end{aligned} \tag{10}$$

The objective function of $\boldsymbol{B}_{\bullet k}$ reduces to equation (11) with $\boldsymbol{S}$ fixed

$$f(\boldsymbol{B}_{\bullet k}) = \left[\boldsymbol{S}_{k\bullet}[\boldsymbol{S}_{k\bullet}]^T\right]\left[[\boldsymbol{B}_{\bullet k}]^T\boldsymbol{B}_{\bullet k}\right] + 2[\boldsymbol{B}_{\bullet k}]^T\left\{\tilde{\boldsymbol{B}}^k\boldsymbol{S}[\boldsymbol{S}_{k\bullet}]^T - \boldsymbol{X}[\boldsymbol{S}_{k\bullet}]^T\right\} \tag{11}$$

where $\tilde{\boldsymbol{B}}^k = \begin{cases} \boldsymbol{B}_{\bullet p}, & p \neq k \\ \boldsymbol{0}, & p = k \end{cases}$.

When imposing the ℓ_2 norm constraint, i.e. $\|\boldsymbol{B}_{\bullet k}\|_2^2 = [\boldsymbol{B}_{\bullet k}]^T\boldsymbol{B}_{\bullet k} = 1$, (11) becomes (12)

$$f(\boldsymbol{B}_{\bullet k}) = 2[\boldsymbol{B}_{\bullet k}]^T\left\{\tilde{\boldsymbol{B}}^k\boldsymbol{S}[\boldsymbol{S}_{k\bullet}]^T - \boldsymbol{X}[\boldsymbol{S}_{k\bullet}]^T\right\} + \boldsymbol{S}_{k\bullet}[\boldsymbol{S}_{k\bullet}]^T \tag{12}$$

Hence, the original constrained minimization problem becomes a linear programming under a unit norm constraint, whose solution is as follows.

$$\begin{aligned} \boldsymbol{B}_{\bullet k} &= \arg\min_{\boldsymbol{B}_{\bullet k}}\|\boldsymbol{X} - \boldsymbol{B}\boldsymbol{S}\|_F^2 \\ &= \frac{\boldsymbol{X}[\boldsymbol{S}_{k\bullet}]^T - \tilde{\boldsymbol{B}}^k\boldsymbol{S}[\boldsymbol{S}_{k\bullet}]^T}{\left\|\boldsymbol{X}[\boldsymbol{S}_{k\bullet}]^T - \tilde{\boldsymbol{B}}^k\boldsymbol{S}[\boldsymbol{S}_{k\bullet}]^T\right\|_2} \end{aligned} \tag{13}$$

where $\tilde{\boldsymbol{B}}^k_{\bullet p} = \begin{cases} \boldsymbol{B}_{\bullet p}, & p \neq k \\ \boldsymbol{0}, & p = k \end{cases}$.

3.3 Overall algorithm

Our algorithm for sparse coding and bases learning is shown in Algorithm 1. Here, $\boldsymbol{1} \in \mathbb{R}^{K\times K}$ is a square matrix with all elements 1, $\boldsymbol{I} \in \mathbb{R}^{K\times K}$ is the identity matrix, and $\odot$ indicates element dot product. By iterating $\boldsymbol{S}$ and $\boldsymbol{B}$ alternately, the sparse codes are obtained, and the corresponding bases are learned.

Algorithm 1 Elastic Net Regularized Dictionary Learning

```
Require: Data matrix $\boldsymbol{X} \in \mathbb{R}^{D \times N}$ and $K$
$\boldsymbol{B} \leftarrow rand(D, K), \boldsymbol{B}_{\bullet k} = \frac{\boldsymbol{B}_{\bullet k}}{\|\boldsymbol{B}_{\bullet k}\|_2} \forall k, \boldsymbol{S} \leftarrow zeros(K, N)$
$iter = 0$
while $(f(iter) - f(iter + 1))/f(iter) > 1e{-}6$ do
    $iter \leftarrow iter + 1$
    Update $\boldsymbol{S}$:
    Compute $\boldsymbol{A} = (\boldsymbol{B}^T\boldsymbol{B}) \odot (\mathbf{1} - \boldsymbol{I})$ and $\boldsymbol{E} = \boldsymbol{B}^T\boldsymbol{X}$
    for $k = 1; k \leq K; k{+}{+}$ do
        $\boldsymbol{S}_{k\bullet} = \left\{ \max\{\boldsymbol{E}_{k\bullet} - \boldsymbol{A}_{k\bullet}\boldsymbol{S}, \alpha\} + \min\{\boldsymbol{E}_{k\bullet} - \boldsymbol{A}_{k\bullet}\boldsymbol{S}, -\alpha\} \right\}/(1 + \beta)$
    end for
    Update $\boldsymbol{B}$:
    Compute $\boldsymbol{G} = (\boldsymbol{S}\boldsymbol{S}^T) \odot (\mathbf{1} - \boldsymbol{I})$, $\boldsymbol{W} = \boldsymbol{X}\boldsymbol{S}^T$
    for $k = 1; k \leq K; k{+}{+}$ do
        $\boldsymbol{B}_{\bullet k} = \frac{\boldsymbol{W}_{\bullet k} - \boldsymbol{B}\mathbf{G}_{\bullet k}}{\|\boldsymbol{W}_{\bullet k} - \boldsymbol{B}\mathbf{G}_{\bullet k}\|_2}$
    end for
    Update the objective function:
    $f(iter) = \|\boldsymbol{X} - \boldsymbol{B}\boldsymbol{S}\|_F^2 + 2\alpha\|\boldsymbol{S}\|_1 + \beta\|\boldsymbol{S}\|_F^2$
end while
return $\boldsymbol{B}$, and $\boldsymbol{S}$
```

Unlikely traditional dictionary learning methods, which use the alternating iteration optimization strategy for the dictionary $\boldsymbol{B}$ and the corresponding sparse codes $\boldsymbol{S}$, we propose method to divide the dictionary learning problem into $2K$ subproblems with close form solution for each subproblem.

3.4 Analysis and comment

Theorem 1 *The objective function in* (1) *is nonincreasing under the update rules given in* (9) *and* (13).

Proof Since the exact minimization point is obtained by (9) or (13), each operation updates $\boldsymbol{S}_{1\bullet}, \ldots, \boldsymbol{S}_{K\bullet}, \boldsymbol{B}_{\bullet 1}, \ldots, \boldsymbol{B}_{\bullet K}$ alternately, it monotonically decreases the objective function in (1). Considering that the objective function is obviously bounded below, it converges. □

In respect that the optimal value for a given block of variables is uniquely attained by the solution (9) and (13) due to the strict convexity of subproblems at each iteration, the limit points are stationary points [2].

Our proposed optimization strategy is suitable for dictionary learning with huge of samples, because the samples can be handled parallel.

4 Experimental results

In this section, three benchmark datasets, such as UIUC-Sports dataset [13], Scene 15 dataset [10], and Caltech-101 dataset [12] are utilized to evaluate the performance of our proposed elastic net regularized dictionary learning (ENDL) algorithm.

4.1 Experimental settings

For each dataset, the data are randomly split into training set and testing set based on published protocols. To make the results more convincing, the experimental process is repeated 8 times, and the mean and standard deviation of the classification accuracy are record. Each image is resized with maximum side 300 pixels firstly, except 400 pixels for UIUC-Sports dataset due to the high resolution of original images. As for the image features, densely sampled SIFT features are used to demonstrate the effectiveness of ENDL. The feature is extracted under three scales 16 × 16, 24 × 24, and 32 × 32, and the step size 8 pixels. 128 dimensional SIFT descriptors are obtained and normalized to 1 with ℓ_2 norm. The samples used for learning dictionary are about 80,000. The dictionary size is 512 and 1024 (512 size dictionary is used to adjust the parameter). Spatial pyramid matching kernel is embedded with 1, 4, and 16 segments. We use a max pooling strategy [30]. An image is represented by the concatenation of each segment and normalized to 1 with ℓ_2 norm. Linear kernel SVM classifier and one-vs-respectively multi-classification strategy are used, and the LIBSVM [3] package is adopted.

4.2 ENDL algorithm for image classification

The proposed ENDL algorithm is applied for image classification on three benchmark datasets, such as UIUC-Sports dataset, Scene 15 dataset and Caltech-101 dataset.

4.2.1 UIUC-Sports dataset

For UIUC-Sports dataset [13], there are 8 classes with totally 1,579 images: rowing (250 images), badminton (200 images), polo (182 images), bocce (137 images), snow boarding (190 images), croquet (236 images), sailing (190 images), and rock climbing (194 images). For each class, the sizes and even the number of instances are quite different, while the poses of the objects vary a lot. The background is also highly clutter and discrepant. Some images from different classes may have similar background. Figure 1 shows some example

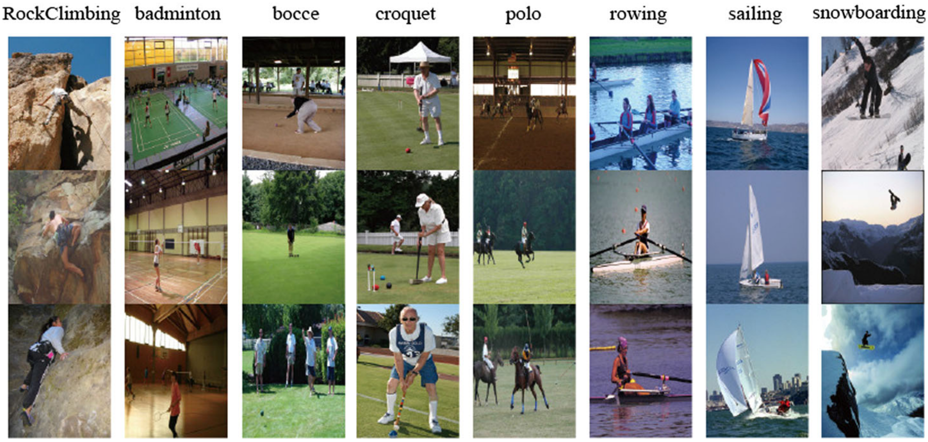
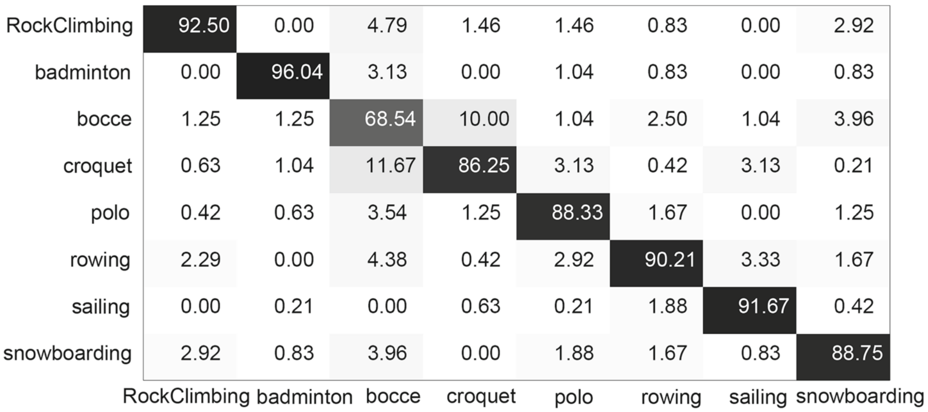


Fig. 1 Example images from the UIUC-Sports data set



	RockClimbing	badminton	bocce	croquet	polo	rowing	sailing	snowboarding
RockClimbing	92.50	0.00	4.79	1.46	1.46	0.83	0.00	2.92
badminton	0.00	96.04	3.13	0.00	1.04	0.83	0.00	0.83
bocce	1.25	1.25	68.54	10.00	1.04	2.50	1.04	3.96
croquet	0.63	1.04	11.67	86.25	3.13	0.42	3.13	0.21
polo	0.42	0.63	3.54	1.25	88.33	1.67	0.00	1.25
rowing	2.29	0.00	4.38	0.42	2.92	90.21	3.33	1.67
sailing	0.00	0.21	0.00	0.63	0.21	1.88	91.67	0.42
snowboarding	2.92	0.83	3.96	0.00	1.88	1.67	0.83	88.75

Fig. 2 Confusion matrix on UIUC-Sports data set (%)

images. We follow the common setup: 70 images per class are randomly selected as the training data, and 60 images per class for testing. Figure 2 shows the confusion matrices. Table 1[1] shows performance of different methods. The classification rate for image classification of our proposed ENDL algorithm outperforms that of traditional ℓ_1 norm regularized dictionary learning by 5.05 %.

4.2.2 *Scene 15 dataset*

For Scene 15 dataset, there are 15 classes with totally 4,485 images. Each class varies from 200 to 400 images. The images contain not only indoor scenes, such as bedroom, living room, PARoffice, kitchen, and store, but also outdoor scenes, such as industrial, forest, mountain, tallbuilding, highway, street, opencountry, and so on. Figure 3 shows some example images. We use an identical experimental setup as [10]: 100 images per class are randomly selected as the training data, and the rest for testing. Figure 4 shows the confusion matrices. Table 1[1] shows performance of different methods. The classification rate for image classification of our proposed ENDL algorithm outperforms that of traditional ℓ_1 norm regularized dictionary learning by 1.97 %.

4.2.3 *Caltech-101 dataset*

Caltech-101 dataset introduced in [12] contains 102 classes, one of which is the background. After removing the background class, the rest 101 classes with totally 8,677 images are used for classification, with each class varying from 31 to 800 images. We follow the common experimental setup for this data set, where 15 and 30 images per category are selected as the training set, and the rest for the testing set (the maximum is 50 images per category for testing). Table 2 shows the performance of different methods. Our proposed ENDL algorithm outperforms the traditional sparse representation based image recognition by 3.42 % and 4.35 % for 15 and 30 training images per class, respectively.

[1] All the results of OCSVM and HIKVQ are based on step size 8 and without concatenated Sobel images.

Table 1 Performance comparisons on UIUC-Sports data set and Scene 15 data set (%)

Methods	UIUC-Sports	Scene 15
ScSPM [5, 30]	82.74 ± 1.46	80.28 ± 0.93
OCSVM [29]	81.33 ± 1.56	82.02 ± 0.54
HIKVQ [29]	81.87 ± 1.14	81.77 ± 0.49
EMK [1]	74.56 ± 1.32	NA
ENDL	87.79 ± 0.94	82.25 ± 0.56

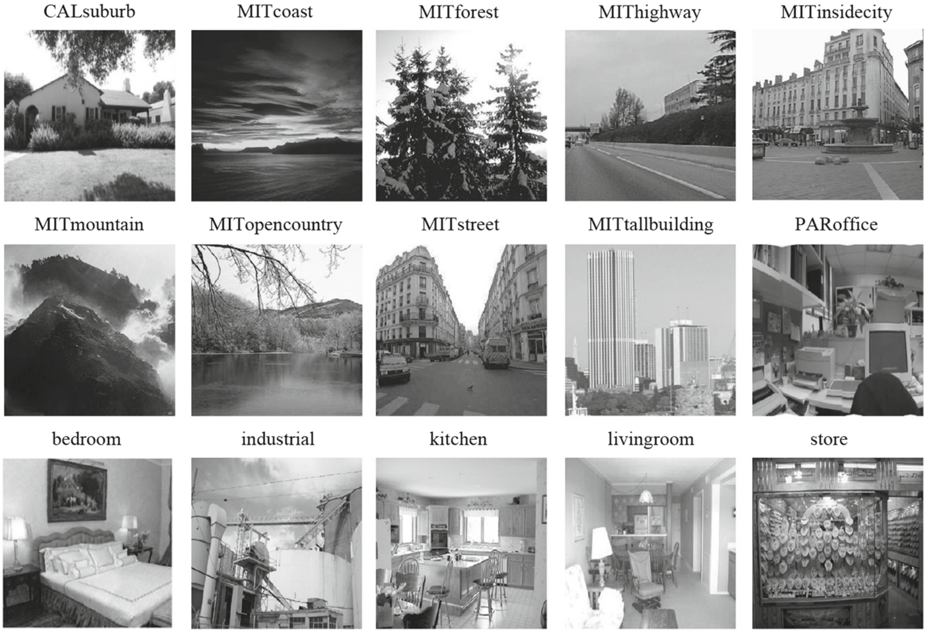
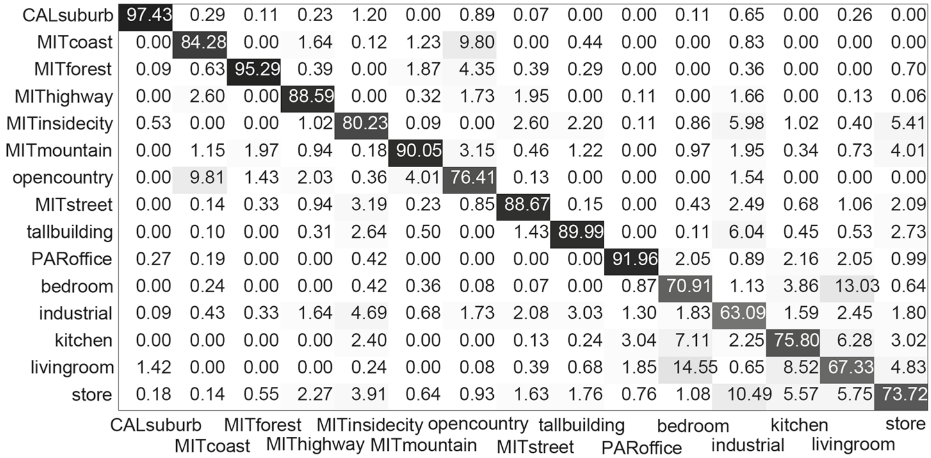


Fig. 3 Example images from the Scene 15 data set



	CALsuburb	MITcoast	MITforest	MIThighway	MITinsidecity	MITmountain	opencountry	MITstreet	tallbuilding	PARoffice	bedroom	industrial	kitchen	livingroom	store
CALsuburb	97.43	0.29	0.11	0.23	1.20	0.00	0.89	0.07	0.00	0.00	0.11	0.65	0.00	0.26	0.00
MITcoast	0.00	84.28	0.00	1.64	0.12	1.23	9.80	0.00	0.44	0.00	0.00	0.83	0.00	0.00	0.00
MITforest	0.09	0.63	95.29	0.39	0.00	1.87	4.35	0.39	0.29	0.00	0.00	0.36	0.00	0.00	0.70
MIThighway	0.00	2.60	0.00	88.59	0.00	0.32	1.73	1.95	0.00	0.11	0.00	1.66	0.00	0.13	0.06
MITinsidecity	0.53	0.00	0.00	1.02	80.23	0.09	0.00	2.60	2.20	0.11	0.86	5.98	1.02	0.40	5.41
MITmountain	0.00	1.15	1.97	0.94	0.18	90.05	3.15	0.46	1.22	0.00	0.97	1.95	0.34	0.73	4.01
opencountry	0.00	9.81	1.43	2.03	0.36	4.01	76.41	0.13	0.00	0.00	0.00	1.54	0.00	0.00	0.00
MITstreet	0.00	0.14	0.33	0.94	3.19	0.23	0.85	88.67	0.15	0.00	0.43	2.49	0.68	1.06	2.09
tallbuilding	0.00	0.10	0.00	0.31	2.64	0.50	0.00	1.43	89.99	0.00	0.11	6.04	0.45	0.53	2.73
PARoffice	0.27	0.19	0.00	0.00	0.42	0.00	0.00	0.00	0.00	91.96	2.05	0.89	2.16	2.05	0.99
bedroom	0.00	0.24	0.00	0.00	0.42	0.36	0.08	0.07	0.00	0.87	70.91	1.13	3.86	13.03	0.64
industrial	0.09	0.43	0.33	1.64	4.69	0.68	1.73	2.08	3.03	1.30	1.83	63.09	1.59	2.45	1.80
kitchen	0.00	0.00	0.00	0.00	2.40	0.00	0.00	0.13	0.24	3.04	7.11	2.25	75.80	6.28	3.02
livingroom	1.42	0.00	0.00	0.00	0.24	0.00	0.08	0.39	0.68	1.85	14.55	0.65	8.52	67.33	4.83
store	0.18	0.14	0.55	2.27	3.91	0.64	0.93	1.63	1.76	0.76	1.08	10.49	5.57	5.75	73.72

Fig. 4 Confusion matrix on Scene 15 data set (%)

Table 2 Performance comparison on Caltech-101 data set

Methods	15 training	30 training
KSPM [10]	–	64.6 ± 0.8
KC [27]	–	64.1 ± 1.2
LLC[a] [28]	63.92 ± 0.46	70.63 ± 0.99
ScSPM [30]	67.0 ± 0.45	73.2 ± 0.54
ENDL	70.42 ± 0.4	77.55 ± 0.54

[a] For LLC, we adopt the code of local feature coding provided by [28] and do experiment on our data set with single scale features and the size of dictionary 1024

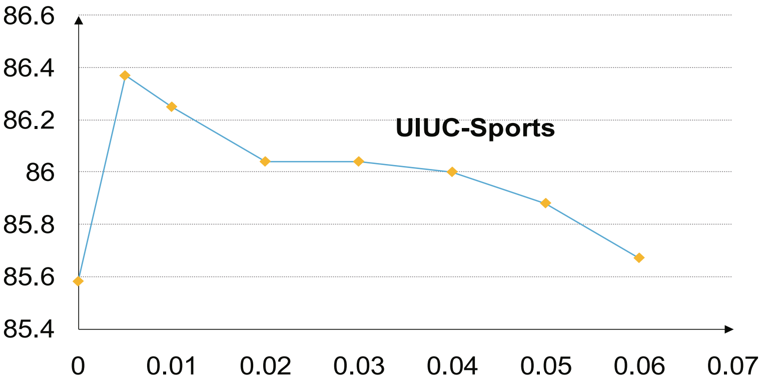


Fig. 5 Selection of parameters for UIUC-Sports dataset. The classification rate under different β with $\alpha = 0.15$ and dictionary size 512

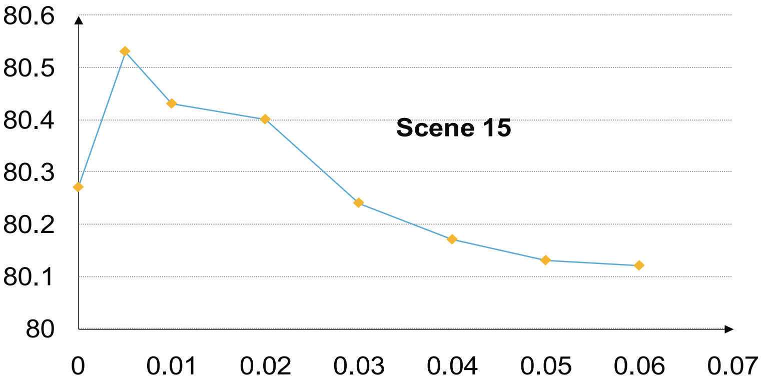


Fig. 6 Selection of parameters for Scene 15 dataset. The classification rate under different β with $\alpha = 0.15$ and dictionary size 512

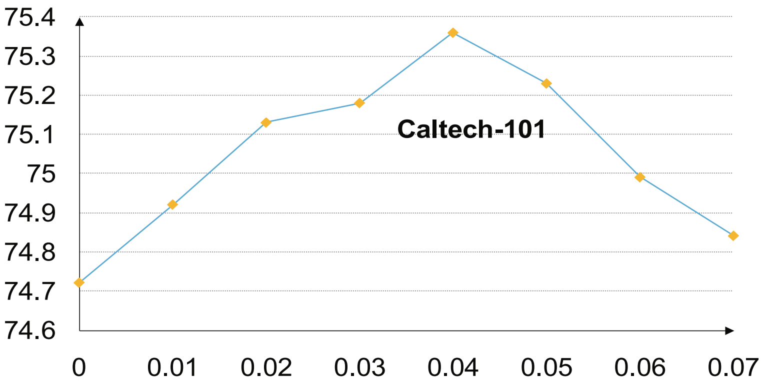


Fig. 7 Selection of parameters for Caltech-101 dataset. The classification rate under different β with $\alpha = 0.15$ and dictionary size 512

4.3 Selection of parameters

There are two parameters α and β. α controls the tradeoff between fitting goodness and sparseness. With α increasing, the codes become sparser and sparser, more and more salient, easier to be distinguished; on the contrary, the reconstruction error becomes larger and larger, leading to inaccurate description of the codes. To be consistent with the empirical conclusion in the paper [30], the parameter α is set to 0.15. β is used to reflect a grouping effect in atom selection. With the increasing value of β, the reconstruction error becomes larger, the sparsity becomes less sparse, and atom selection becomes more sense. We studied the effect of different β for these three datasets. Figure 5 lists the performance when $\beta = \{0.0, 0.005, 0.01, 0.015, 0.02, 0.025, 0.03, 0.035\}$ with $\alpha = 0.15$ for UIUC-Sports dataset. As can be seen, when $\beta = 0.0$, ENDL are equivalent to sparse coding based dictionary learning method. With β growing, the classification rate increases. The best classification accuracy can be obtained when $\beta = 0.005$. After that, the performance starts to degenerate. Figure 6 lists the performance when $\beta = \{0.0, 0.005, 0.01, 0.015, 0.02, 0.025, 0.03, 0.035\}$ with $\alpha = 0.15$ for Scene-15 dataset. The best classification accuracy can be obtained when $\beta = 0.005$. Figure 7 lists the performance when $\beta = \{0.0, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07\}$ with $\alpha = 0.15$ for Caltech-101 dataset. The best classification accuracy can be obtained when $\beta = 0.04$.

5 Conclusion

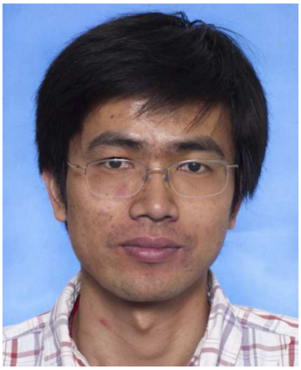
Dictionary learning is a fundamental component for image representation. Image features are coded as sparse linear combinations of atoms in dictionary. Sparse pattern is believed to a critical property of encoding, since it correspond to the atom selection. Considering that a dictionary is of multi-modality since it represents images from multiple classes, the atom selection makes more sense there is a grouping effect, i.e., the atoms for the same class of images should be selected for any image feature. To improve the grouping effect and enjoy the merits of sparsity, elastic net is introduced as a regularizer for dictionary learning. Experimental results on benchmark datasets show that the proposed dictionary outperforms the state-of-art sparse dictionary learning algorithm, which only considers ℓ_1 penalty.

Acknowledgments This work was supported by the National Natural Science Foundation of P.R. China (No. 61402535), Qingdao Science and Technology Project (No. 14-2-4-111-jch), the Fundamental Research Funds for the Central Universities (No. R1405012A), and the Talent Acquisition Project (No.Y1305024).

References

1. Bo L, Sminchisescu C (2009) Efficient match kernel between sets of features for visual recognition. In: Proceedings of Advances in neural information processing systems, pp. 135–143
2. Bertsekas DP (1999) Nonlinear programming. Athena Scientific, Belmont
3. Chang C-C, Lin C-J (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Systs Technol 2(3):27:1–27:27
4. Gao S, Tsang IW-H, Chia L-T (2010) Kernel sparse representation for image classification and face recognition. In: Proceedings of the 11th ECCV. Springer, pp 1–14
5. Gao S, Tsang IWH, Chia LT (2013) Laplacian sparse coding, hypergraph laplacian sparse coding, and applications. IEEE Trans Pattern Anal Mach Intell 35(1):92–104
6. Gao S, Tsang IW, Chia L-T, Zhao P (2010) Local features are not lonely–laplacian sparse coding for image classification. In: Proceedings of the 23rd CVPR. IEEE, pp 3555–3561
7. Gao Y, Wang M, Tao D, Ji R, Dai Q (2012) 3-d object retrieval and recognition with hypergraph analysis. IEEE Trans Image Process 21(9):4290–4303
8. Gao Y, Wang M, Zha Z-J, Shen J, Li X, Wu X (2013) Visual-textual joint relevance learning for tag-based social image search. IEEE Trans Image Process 22(1):363–376
9. Gao Y, Wang M, Zha Z-J, Tian Q, Dai Q, Zhang N (2011) Less is more: efficient 3-d object retrieval with query view selection. IEEE Trans Multimed 13(5):1007–1018
10. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proceedings of the 19th CVPR, pp 2169–2178
11. Lee H, Battle A, Raina R, Ng AY (2006) Efficient sparse coding algorithms. In: Proceedings of advances in neural information processing systems, pp 801–808
12. Li F-F, Fergus R, Perona P (2004) Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In: Workshop of the 17th CVPR, vol 12, p 178
13. Li L-J, Fei-Fei L (2007) What, where and who? Classifying events by scene and object recognition. In: Proceedings of the 11th ICCV. IEEE, pp 1–8
14. Liu B-D, Wang Y-X, Zhang Y-J, Shen B (2013) Learning dictionary on manifolds for image classification. Pattern Recog 46(7):1879–1890
15. Liu B-D, Wang Y-X, Shen B, Zhang Y-J, Hebert M (2014) Self-explanatory sparse representation for image classification. In: Proceedings of the 13th ECCV. Springer, pp 600–616
16. Liu B-D, Wang Y-X, Shen B, Zhang Y-J, Wang Y-J, Liu W-F (2013) Self-explanatory convex sparse representation for image classification. In: 2013 IEEE international conference on systems, man, and cybernetics (SMC). IEEE, 2120–2125
17. Liu B-D, Wang Y-X, Shen B, Zhang Y-J, Wang Y-J (2014) Blockwise coordinate descent schemes for sparse representation. In: Proceedings of the 39th ICASSP. IEEE, pp 5267–5271
18. Liu B-D, Wang Y-X, Zhang Y-J, Zheng Y (2012) Discriminant sparse coding for image classification. In: 2012 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 2193–2196
19. Lu Z, Peng Y (2011) Latent semantic learning by efficient sparse coding with hypergraph regularization. In: Proceedings of the 25th AAAI, pp 411–416
20. Ramamurthy KN, Thiagarajan JJ, Spanias A (2011) Improved sparse coding using manifold projections. In: Proceedings of the 18th ICIP. IEEE, pp 1237–1240

21. Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500):2323–2326
22. Shen B, Liu B-D, Allebach J (2014) Tisvm: large margin classifier for misaligned image classification. In: Proceedings of the 21st ICIP. IEEE
23. Shen B, Liu B-D, Wang Q, Ji R (2014) Robust nonnegative matrix factorization via l1 norm regularization by multiplicative updating rules. In: Proceedings of the 21st ICIP. IEEE
24. Shen B, Si L (2010) Non-negative matrix factorization clustering on multiple manifolds. In: Proceedings of the 24th AAAI, pp 575–580
25. Shen B, Wei H, Zhang Y, Zhang Y-J (2009) Image inpainting via sparse representation. In: IEEE international conference on acoustics, speech and signal processing, 2009. ICASSP 2009. IEEE, pp 697–700
26. Van Nguyen H, Patel VM, Nasrabadi NM, Chellappa R (2012) Kernel dictionary learning. In: Proceedings of the 37th ICASSP. IEEE, pp 2021–2024
27. van Gemert JC, Veenman CJ, Smeulders AWM, Geusebroek J-M (2010) Visual word ambiguity. IEEE Trans Pattern Anal Mach Intell 32(7):1271–1283
28. Wang J, Yang J, Yu K, Lv F, Huang T, Y Gong (2010) Locality-constrained linear coding for image classification. In: Proceedings of the 23rd CVPR. IEEE, pp 3360–3367
29. Wu J, Rehg JM (2009) Beyond the euclidean distance: creating effective visual codebooks using the histogram intersection kernel. In: Proceedings of the 12th ICCV, pp 630–637
30. Yang J, Kai Y, Gong Y, Huang TS (2009) Linear spatial pyramid matching using sparse coding for image classification. In: Proceedings of the 22nd CVPR, pp 1794–1801
31. Yi W, Shen B, Ling H (2014) Visual tracking via online non-negative matrix factorization. IEEE Trans Circ Syst Video Technol 24(3):374–383
32. Yu K, Zhang T, Gong Y (2009) Nonlinear learning using local coordinate coding. In: Proceedings of advances in neural information processing systems, pp 2223–2231
33. Zhang D, Yang M, Feng X (2011) Sparse representation or collaborative representation: which helps face recognition? In: Proceedings of the 13th ICCV. IEEE, pp 471–478
34. Zheng M, Bu J, Chen C, Wang C, Zhang L, Qiu G, Cai D (2011) Graph regularized sparse coding for image representation. IEEE Trans Image Process 20(5):1327–1336
35. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc: Ser B (Stat Methodol) 67(2):301–320



Bin Shen is a Ph.D candidate in Department of Computer Science, Purdue University, West Lafayette, Indiana, US 47907. Before joining Purdue, he got B.S. and M.S. degrees from EE, Tsinghua University, Beijing, in 2007 and 2009, respectively. His research interests include image processing, machine learning and data mining.

Bao-Di Liu was born in Shandong, China. He received the Ph.D. degree in Electronic Engineering from Tsinghua University, China. Currently, he is an assistant professor in College of Information and Control Engineering, China University of Petroleum, China. His research interests include computer vision and machine learning.

Qifan Wang is a PhD candidate in the Department of Computer Science, Purdue University. His research interests include machine learning, information retrieval, data mining and computer vision. He received both his BE and ME in computer science from Tsinghua University.