Neurocomputing 169 (2015) 110–118

Contents lists available at ScienceDirect

Neurocomputing

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

Image classification based on low-rank matrix recovery and Naive Bayes collaborative representation

Xu Zhang [a,c], Shijie Hao [a], Chenyang Xu [b], Xueming Qian [b], Meng Wang [a,*], Jianguo Jiang [a]

[a] School of Computer and Information, Hefei University of Technology, Hefei, China
[b] School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China
[c] Computer teaching and research section, Army Officer Academy of PLA, Hefei, China

ARTICLE INFO

Article history:
Received 1 May 2014
Received in revised form
29 December 2014
Accepted 1 January 2015
Available online 28 May 2015

Keywords:
Image classification
Bag of visual words
Collaborative representation
Low rank

ABSTRACT

Most image classification methods require an expensive learning/training phase to gain high performances. But they frequently encounter problems such as overfitting of parameters and scarcity of training data. In this paper, we present a novel learning-free image classification algorithm under the framework of Naive-Bayes Nearest-Neighbor (NBNN) and collaborative representation, where non-negative sparse coding, low-rank matrix recovery and collaborative representation are jointly employed to obtain more robust and discriminative representation. First, instead of using general sparse coding, non-negative sparse coding combined with max pooling is introduced to further reduce information loss. Second, we use the low-rank matrix recovery technique to decompose the training data of the same class into a discriminative low-rank matrix, in which more structurally correlated information is preserved. As for testing images, a low-rank projection matrix is also learned to remove possible image corruptions. Finally, the classification process is implemented by simply comparing the responses over the different bases. Experimental results on several image datasets demonstrate the effectiveness of our method.

1. Introduction

Image classification is one of the fundamental problems in computer vision and pattern recognition, and plays a key role in many applications, e.g. image retrieval [1], image annotation [2], image quality assessment [3] and so on. In the last few years many novel image classification methods have been proposed with the development of artificial intelligence and machine learning. In general, those methods can be roughly classified into two groups: the first group is based on the learning classifiers (called as the learning-based classifiers) which require learning/training classifier parameters; the second group is based on non-parametric classifiers (called as the learning-free classifiers), where the decision rules usually depend on the datasets [4]. Non-parametric classifiers have many advantages over learning-based ones such as avoiding overfitting of parameters and requiring no learning/training phase. However, they often provide inferior performances compared to the learning-based classifiers.

From another viewpoint, feature organization is also a key point in image classification. The Bag of Visual Words (BoVW) model [5] has been widely used in many vision applications and shows satisfying performances especially in the task of image classification. In the BoVW model, codebook generating and feature encoding are crucial steps which determine the completeness of image representation. But there are some limitations existing in the classical BoVW model for image classification: i) Lacking semantic information in codebooks. We cannot ensure the generated visual words containing independent semantic meanings as text words in natural languages. ii) Large quantization errors. Discriminative features may be discarded during clustering as the clustering centers are often determined by high density data points which are less discriminative.

To alleviate the above problems, Yang et al. [6] proposed an extension of spatial pyramid matching kernel (SMP), i.e. ScSPM, which combines with the sparse coding model. Based on lowest reconstruction error, optimal codebook and coding coefficients can be learned through the sparse coding instead of K-means clustering. And the multi-scale max pooling is then used to generate the final image representation. In the max pooling process, however, low responses (small coefficients) are suppressed and only the max response is preserved. Also, the sparse coding model has no constraints on the signs of coding coefficients. So the negative coefficients would be discarded along with large amounts of zero coefficients, although they may potentially contain useful information. These factors are likely to pose negative effects on the classification performance.

[*] Corresponding author.

http://dx.doi.org/10.1016/j.neucom.2015.01.085
0925-2312/© 2015 Elsevier B.V. All rights reserved.

Apart from representing images, sparse coding can also be widely used for other tasks such as recognition and classification. Wright et al. [7] proposed a sparse representation based classification (SRC) algorithm for face recognition and impressive results were reported. In that method, a testing image is represented as a linear and sparse combination of all the training samples via ℓ^1-norm minimization. As another view, it has been argued in [8] that, instead of the ℓ^1-norm sparse representation, it is the collaborative representation (i.e., representing the test image collaboratively by samples from all the classes) that makes SRC effective. Collaborative representation based classification (CRC) can be also regarded as a regularized linear regression problem, where non-sparse ℓ^2-norm is used to regularize the representation coefficients. Compared with SRC, CRC achieves similar results but performs better in terms of the speed. Although SRC and CRC achieve impressive results, these methods are still not robust enough while training images are largely corrupted by factors such as occlusions and disguises. It is a key factor to choose a proper dictionary for the image representation, which greatly impacts performance of image classification. In [8], the original training images are directly used as the dictionary. However, this inefficient strategy would greatly degenerate the classification performance faced with corrupted training images.

A number of approaches have been proposed to address the problem. For example, Wright et al. [9] presented the low-rank matrix recovery method, which is capable of recovering a low-rank matrix from arbitrarily corrupted input data. This method has been successfully adopted in many applications such as background subtraction [10], tracking [11], and web data mining [12]. Chen et al. [13] used this technique to decompose the training data of the same class into a low-rank matrix and an associated sparse error matrix for robust face recognition. In their model, the sparse errors are treated as noises and are removed from the training data. Zhang et al. [14] proposed an image classification method by leveraging the non-negative sparse coding and the low-rank recovery to obtain more discriminative bases. As equal weights of the low-rank matrix and the sparse matrix are used for coding the BoVW representation in their model, this heuristic strategy limits the capacity of enhancing the representing power of the bases.

We also note that both SRC and CRC do not fully take advantage of the common structural information from multiple images of a certain class. It is obvious that the contents of natural images with a same classifying label can be highly diversified. On one hand, they often contain varied backgrounds and multiple objects with different poses and occlusions. On the other hand, target objects usually have large intra-class appearance variability and corruptions such as lighting variations and pixel contamination. Despite these factors, images of a certain class also share lots of common features and correlations. In this sense, the matrix of stacked sparse coding representation of images within the same class may be low-rank, and can be decomposed into a low-rank matrix and a sparse error matrix via low-rank matrix recovery. This representation would maintain more information of a certain object with higher semantic consistency, which is potential in improving classification accuracy.

Motivated by Naive-Bayes Nearest-Neighbor (NBNN) [4] and the above-mentioned observations, in this paper we present a new learning-free image classification algorithm by leveraging low-rank matrix recovery, low-rank projection matrix and collaborative representation. The main contributions in this paper are as follows: (a) the matrix constructed by intra-class images, which are expressed by non-negative sparse coding with max pooling, is low rank; (b) test images can be collaboratively represented by a more discriminating dictionary and a low-rank projection matrix; (c) we derive the Naive Bayes Collaborative Representation, i.e. NBCR, for the classification problem, where the coding coefficients obtained from the ℓ^2-regularized least square formulation are used. Moreover, the proposed method is learning-free, which avoids problems in learning-based method.

The rest of this paper is organized as follows. Section 2 briefly introduces the related work. We present the image representation via non-negative sparse coding and low-rank matrix recovery in Section 3. The proposed image classification method is given in Section 4. Experimental results on several publicly available datasets are reported in Section 5. Section 6 finally concludes the paper.

2. Related work

BoVW model has been proven to be effective for image classification. Over the past years, many improved versions such as discriminative codebook learning [15,16], feature encoding [17,18] and classifier learning [19,20] have been proposed. Among of the extensions, Lazebnik et al. [21] proposed the spatial pyramid matching kernel (SPM) to model the spatial layout of the local features, which achieves impressive performances and has been widely used. Yang et al. [6] presented a method called ScSPM by combining the SPM with sparse coding. The model is highlighted in its speed (only linear SVM is needed) and state-of-the-art results on several benchmark datasets. As a step further, Wang et al. [22] improved ScSPM by introducing the locality constraint, which further speeds up the algorithm and increases the accuracy.

Sparse coding [23] has been demonstrated to be a powerful image representation. The idea is to represent an image as a linear combination of a few bases from an over-complete codebook. Liu et al. [24] proposed to learn sparse and non-negative representations of image and applied the method into several applications, e.g. face recognition and image classification. Zhou [25] presented a label consistent K-SVD (LC-KSVD) algorithm to learn a compact and discriminative dictionary for sparse coding. Sparse methods also be used in classifying tasks. Sparse representation-based classification (SRC) [7] has shown very promising results on face recognition where training images are often directly chosen as the bases for sparse representation and testing images can be classified by the minimal reconstruction error. Recently, Zhang et al. [8] argued that, instead of the ℓ^1-norm sparsity on α, the collaborative representation actually makes the vital contribution in SRC. However, when both training and testing images are corrupted, the performances of these methods would be degenerated. Chen et al. [13] utilized low-rank matrix recovery to address the SRC problem, where the training data is decomposed into a set of representative bases and a sparse error matrix and the noises are thus effectively removed.

As a useful tool, low-rank matrix recovery keeps attracting extensive attentions from many research communities and promising results have been achieved in many applications [26,27]. Lin et al. [28] proposed the Accelerated Proximal Gradient approach to solve a relaxed convex form of the problem. Applying augmented Lagrange multipliers (ALM), Lin et al. [29] introduced RPCA though the Exact and Inexact ALM method. As for the applications, Chen et al. [13] used a low-rank technique to remove noise from training data for face recognition. Zhang et al. [30] presented a new image classification approach to learn a structural low-rank and sparse image representation. Chen [31] presented a method to learn a low-rank projection matrix between the training images and the recovery results. Zhang et al. [14] represent images by combining the low rank matrix and the sparse matrix with equal weights. The LLC method in [22] is then adopted to achieve the classification.

Different from the above-mentioned methods, Boiman et al. [4] proposed a novel non-parametric method named Naive-Bayes Nearest-Neighbor (NBNN) for image classification. The NBNN method is simple and requires no learning/training phase. Meanwhile, it achieves satisfying performance. In the method, instead of image-to-image (I2I) distance, image-to-class (I2C) distance is employed for the classification task. Avoidance of descriptor quantization and using of I2C distance are the key ingredients of the NBNN method.

Compared to the previous works, our approach distinguishes itself in fully utilizing the structural information of images within the same class. Also, by promoting the structural incoherence for collaborative representation, the dictionary elements between different classes can be independent as much as possible. Inherited from the characteristic of the NBNN method, our method is also learning-free.

3. Learning sparse and low-rank representation

In order to take advantages of the underlying data structure, non-negative sparse coding and low-rank matrix recovery are applied for the image representation in our method. First, non-negative sparse coding is introduced in Section 3.1. Then, low-rank matrix recovery with structural incoherence and low-rank projection matrix are described in Section 3.2 and Section 3.3 respectively.

3.1. Non-negative sparse coding

K-means clustering is wildly used for the codebook generation in BoVW model. Let $\boldsymbol{X}$ be a set of local features $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_N]$ $(\boldsymbol{x}_i\in R^{d\times 1}, i=1,\ldots,N)$, e.g., $d=128$ for the SIFT descriptor. The data matrix of local feature space is partitioned into k clusterings, with the corresponding centers $\boldsymbol{V}=[\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_k]\in R^{d\times k}$ forming the visual words. To represent $\boldsymbol{x}_i$, the vector quantization (VQ) by K-means clustering method is applied to solve the following optimization problem:

$$\min_{\mathbf{U},\mathbf{V}}\sum_{i=1}^{N}\|\boldsymbol{x}_i-\boldsymbol{u}_n\boldsymbol{V}\|^2 \quad \text{Subject to } Card(\boldsymbol{u}_i)=1, |\boldsymbol{u}_i|=1, \boldsymbol{u}_i 0, \forall i \qquad (1)$$

where $\mathbf{U}=[\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_N]$ $(\mathbf{u}_i\in R^{k\times 1}, i=1,\ldots,N)$ is the cluster membership indicators. The cardinality constraint of $Card(\mathbf{u}_i)=1$ indicates that each local feature can be assigned to only one visual word. Yang et al. [6] relaxed the constraint using sparse coding by a ℓ^1-norm regularization and the problem of Eq. (1) can be reformed as the following optimization:

$$\min_{\mathbf{U},\mathbf{V}}\sum_{i=1}^{N}\|\boldsymbol{x}_i-\boldsymbol{u}_n\boldsymbol{V}\|^2+\lambda\|\boldsymbol{u}_i\|_1 \quad \text{Subject to } \|\boldsymbol{v}_k\|^2\le 1, \forall k \qquad (2)$$

where λ is the regularization parameter and $\|\bullet\|_1$ denotes the ℓ^1-norm. In max pooling operation of ScSPM, the negative coefficients are suppressed by zero coefficients, although they have less meaningful responses than negative coefficients. This possibly limits the classifying performance due to the potential information loss.

To address this problem, we employ the non-negative sparse coding method, which tries to solve the following optimization problem as:

$$\min_{\mathbf{U},\mathbf{V}}\sum_{i=1}^{N}\|\mathbf{x}_i-\boldsymbol{u}_n\boldsymbol{V}\|^2+\lambda\|\boldsymbol{u}_i\|_1 \quad \text{Subject to } \|\boldsymbol{v}_k\|^2\le 1, \boldsymbol{u}_i 0, \forall k, i \qquad (3)$$

Compared with Eq. (2), the coding coefficients are restricted to be non-negative. Non-negative sparse coding maintains the characteristics of standard sparse coding and is also in consistency with mammal's visual mechanism.

The optimization in Eq. (3) is not convex for $\boldsymbol{U}$ and $\boldsymbol{V}$ simultaneously, but it is convex for $\boldsymbol{U}$ when $\boldsymbol{V}$ is fixed and vice versa. Following the work in [23], we can optimize $\boldsymbol{U}$ and $\boldsymbol{V}$ in an alternative style.

3.2. Low-rank matrix recovery with structural incoherence

Low-rank matrix recovery has attracted extensive attentions recently. It seeks to decomposed a data matrix $\boldsymbol{D}$ into two matrices, i.e. $\mathbf{D}=\boldsymbol{A}+\boldsymbol{E}$, where $\boldsymbol{A}$ is a low-rank matrix and $\boldsymbol{E}$ is the associated sparse matrix. Aiming at obtaining the low-rank approximation of $\boldsymbol{D}$, the method minimizes the rank of matrix $\boldsymbol{A}$ and reduces $\|\mathbf{E}\|_0$ in the meanwhile. Although this problem is theoretically NP-hard, Candes et al. [32] reformulated the problem as the following equation to solve it

$$\min_{\mathbf{A},\mathbf{E}}\|\mathbf{A}\|_*+\beta\|\mathbf{E}\|_1 \quad \text{Subject to } \boldsymbol{D}=\boldsymbol{A}+\boldsymbol{E} \qquad (4)$$

The nuclear norm $\|\bullet\|_*$ (the sum of the singular values) is used to approximate the rank of $\boldsymbol{A}$, and the ℓ^0-norm $\|\boldsymbol{E}\|_0$ is replaced by ℓ^1-norm $\|\boldsymbol{E}\|_1$, which sums up the absolute values of entries in $\boldsymbol{E}$. The technique of inexact augmented Lagrange multipliers [29] is applied to solve the target function (4) efficiency.

Specifically, the low-rank matrix recovery has been used to alleviate the limitation in SRC, i.e. less robustness to image corruptions, for improving the performance. Denoting the training data as $\boldsymbol{D}=[\boldsymbol{D}_1,\boldsymbol{D}_2,\ldots,\boldsymbol{D}_N]$, $\boldsymbol{D}_i$ is the set of training data from class i. By performing low-rank matrix recovery, the training data matrix is decomposed into a low rank matrix $\boldsymbol{A}=[\boldsymbol{A}_1,\boldsymbol{A}_2,\ldots,\boldsymbol{A}_N]$ and a sparse error matrix $\boldsymbol{E}=[\boldsymbol{E}_1,\boldsymbol{E}_2,\ldots,\boldsymbol{E}_N]$. The former one is a representative base matrix and the latter contains the associated spares error.

In order to maintain sufficient discriminative information and improve the representation ability of the low-rank matrix, the structural incoherence is useful to constrain the resulting low-rank matrices as independent as possible. Therefore the common features across different classes are reduced while the discriminative ones are preserved. Specifically, a regularization term is added to the objective function of formulation (4) to enforce the incoherence between the obtained low-rank matrices [13]. The optimization problem is therefore improved as

$$\min_{\mathbf{A},\mathbf{E}}\sum_{i=1}^{N}\{\|\boldsymbol{A}_i\|_*+\beta\|\boldsymbol{E}_i\|_1\}+\eta\sum_{j\ne i}\|\boldsymbol{A}_j^T\boldsymbol{A}_i\|_F \quad \text{Subject to } \boldsymbol{D}_i=\boldsymbol{A}_i+\boldsymbol{E}_i \qquad (5)$$

where η is a penalty parameter. This modified model improves the discriminating ability to the original low-rank matrix recovery. Given this characteristic, we adopt this technique into our natural image classifying framework to enhance the inter-class discriminability. Besides, to make the above problem more tractable, the property that $\|\boldsymbol{A}_j^T\boldsymbol{A}_i\|_F^2\le\|\boldsymbol{A}_j\|_F^2\|\boldsymbol{A}_i\|_F^2$ is used and the formulation (5) is relaxed into the following formulation:

$$\min_{\mathbf{A}_i,\mathbf{E}_i}\|\boldsymbol{A}_i\|_*+\beta\|\boldsymbol{E}_i\|_1+\eta'\|\boldsymbol{A}_i\|_F^2 \quad \text{Subject to } \boldsymbol{D}_i=\boldsymbol{A}_i+\boldsymbol{E}_i \qquad (6)$$

where $\eta'=\eta\sum_{j\ne i}\|\boldsymbol{A}_j\|_F^2$ is a constant when deriving $\mathbf{A}_i$ and $\mathbf{E}_i$. The extend Augmented Lagrange multipliers (ALM) can be employed to solve the formulation with regularization on structural incoherence, which reformulates the problem as follows:

$$\begin{aligned}L(\boldsymbol{A}_i,\boldsymbol{E}_i,\mathbf{Y}_i,u,\eta') &= \|\boldsymbol{A}_i\|_*+\beta\|\boldsymbol{E}_i\|_1+\eta'\|\boldsymbol{A}_i\|_F^2\\ &\quad + <\mathbf{Y}_i,\boldsymbol{D}_i-\boldsymbol{A}_i-\boldsymbol{E}_i> +\frac{\mu}{2}\|\boldsymbol{D}_i-\boldsymbol{A}_i-\boldsymbol{E}_i\|_F^2 \qquad (7)\end{aligned}$$

Details for solving the problem and updating of the above variables can be found in [13].

3.3. Low-rank projection matrix

In this paper, training images within the same class are firstly represented by the non-negative sparse coding and then stacked to a low-rank matrix for low-rank matrix recovery. As for the testing images, they are also likely to suffer from the image corruptions. Therefore, it is necessary to remove the corruptions from the samples for classification. As mentioned above, the low-rank matrix recovery method aims at recovering a low-rank matrix from corrupted data based on the hypothesis that the underlying data structure is approximately a single low-rank subspace. Following the work in [33] and [31], a low-rank projection matrix can be learned for testing images, which projects corrupted data onto their corresponding underlying subspace and removes possible corruptions of testing instances.

Denoting the original training images $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n] \in R^{d \times n}$, its principal components $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_n] \in R^{d \times n}$ can be efficiently obtained by removing the sparse corruption. The projection matrix $\boldsymbol{P}$ links $\boldsymbol{X}$ and $\boldsymbol{Y}$ by means of projecting the data points onto their underlying subspace. The low-rank projection matrix $\boldsymbol{P}$ can be learned by solving the following optimization problem:

$$\min_{\mathbf{P}} rank(\boldsymbol{P}) \text{ Subject to } \boldsymbol{Y} = \boldsymbol{PX} \tag{8}$$

As discussed in the above subsection, it is easy to see that the solution of problem (8) may be unique because of nature of the rank function. Similar to solving the low-rank matrix recovery problem in Section 3.2, the rank function is replaced by the nuclear norm, and the target problem turns into the following convex optimization problem:

$$\min_{\mathbf{P}} \|\mathbf{P}\|_* \text{ Subject to } \mathbf{Y} = \boldsymbol{PX} \tag{9}$$

It has been proven in [34] that if $\mathbf{P} \neq 0$, $\boldsymbol{Y} = \boldsymbol{PX}$ has a meaningful solution and $\boldsymbol{P}^* = \boldsymbol{YX}^+$ is the unique minimization to the formulation (9), where $\boldsymbol{X}^+$ is the pseudo-inverse of $\boldsymbol{X}$. Suppose the skinny SVD of $\mathbf{X}$ is $\mathbf{U\Sigma V}^T$, then the pseudo-inverse of $\mathbf{X}$ can be uniquely defined by $\boldsymbol{X}^+ = \boldsymbol{V\Sigma}^{-1}\boldsymbol{U}^T$. Through the low-rank projection matrix, the principle components $\boldsymbol{y}$ and error $\mathbf{e}$ can be expressed by $\boldsymbol{P^*x}$ and $\boldsymbol{x} - \boldsymbol{P^*x}$ respectively.

4. Naive Bayes collaborative representation based image classification

In this section, we first briefly introduce the NBNN algorithm and then propose our image classification algorithm combining with low-rank matrix recovery and collaborative representation under the NBNN framework.

4.1. NBNN

In NBNN, all the local features are retained in their original form without quantization. And the I2C distance measurement is adopted instead of image-to-image for good generalization. The steps of NBNN can be summarized as follows: i) compute local descriptors of the query image. ii) Find the nearest neighbors (NN) for every descriptor in each class. iii) Calculate and sum up the distance between every descriptor and its NN of each class, i.e. I2C distances. The predicted label of the query is finally assigned to the class with the minimum distance.

4.2. The proposed algorithm

Natural images of real world have complicated characters, which often contain multiple objects with different poses and occlusion even within the same class. However they still share a lot of similarities and correlate with each other. In this way, the algorithm proposed in this paper makes use of this potential relationship via integrating low-rank matrix recovery with collaborative representation.

Formally, let $\boldsymbol{D}_i = [\boldsymbol{d}_{i,1}, \boldsymbol{d}_{i,2}, ..., \boldsymbol{d}_{i,n}]$ be the stacked column vectors of the BoVW representations of n training images in the i-th class. Because of high similarity and relativity, the low-rank $\boldsymbol{A}_i$ and the sparse matrix $\boldsymbol{E}_i$ of each class can be obtained. Naturally if one image belongs to the i-th class, it can be well reconstructed by vectors of the i-th $\boldsymbol{A}_i$ instead of other classes. Under the framework of NBNN, we present an image classification algorithm based on Naive Bayes collaborative representation (NBSC) combining with low-rank matrix recovery. All the procedures of the NBCR algorithm are summarized in Algorithm 1:

```
Algorithm 1 Classification method based on Low-rank matrix
  and NBCR
Input: labeled (training) data $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]$ for $N$ classes, the
  testing sample, parameters $\lambda, \beta, \eta$ and $\gamma$.
  Step 1: perform non-negative sparse coding of training and
  testing sample
      $\min_{\mathbf{U,V}} \sum_{i=1}^{N} \|\mathbf{x} - \mathbf{uV}\|^2 + \lambda\|\mathbf{u}\|_1$ Subject to $\|\mathbf{v}\|^2 \leq 1, \mathbf{u}0$
    $\mathbf{U}$ and $\mathbf{y}$ are the sparse representation of training and
  testing sample respectively, where $\mathbf{U} = [\mathbf{U}_1, \mathbf{U}_2, ..., \mathbf{U}_N]$, $\mathbf{U}_i$ is
  the stacked BoVW representation of class $i$.
  Step 2: perform low-rank matrix recovery on $\mathbf{U}$
      For $i = 1{:}N$ do
      $\min_{\mathbf{A,E}} \sum_{i=1}^{N} \{\|\mathbf{A}_i\|_* + \beta\|\mathbf{E}_i\|_1\} + \eta\sum_{j \neq i}\|\mathbf{A}_j^T\mathbf{A}_i\|_F$ Subject to
  $\mathbf{U}_i = \mathbf{A}_i + \mathbf{E}_i$
      End for
  Step 3: Obtain the recovery results from training samples
        $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, ..., \mathbf{A}_N]$
  Compute the skinny SVD of $\mathbf{U}$ and the pseudo-inverse of $\mathbf{U}$
        $\mathbf{U} = \mathbf{u\Sigma v}^T$
        $\mathbf{U}^+ = \mathbf{v\Sigma}^{-1}\mathbf{u}^T$
  Calculate the low-rank projection matrix
        $\mathbf{P}^* = \mathbf{AU}^+$
  Project the BoVW representation of testing sample $\mathbf{y}$ on $\mathbf{P}^*$
        $\mathbf{y}_t = \mathbf{P}^*\mathbf{y}$
  Step 4: Perform NBCR to classify $\mathbf{y}_t$ over $\mathbf{A}$ by solving the
  following problem:
        $\min_{\mathbf{w}} \|\mathbf{y}_t - \mathbf{Aw}\|_2^2 + \gamma\|\mathbf{w}\|_2$
        $\hat{\mathbf{w}} = |\mathbf{w}| / \|\mathbf{w}\|_1$
      For $i = 1{:}N$ do
        $\hat{c}(i) = \sum_i \hat{\mathbf{w}}_i$
      End for
Output:
  Identity($\mathbf{q}$)$=\arg\max_i \hat{\mathbf{c}}(i)$
```

In our method, the low-rank projection matrix $\mathbf{P}^*$ is learned to recover the principle components of the testing sample for classification. But the recovery model is based on the hypothesis that the data is drawn from a single low-rank subspace rather than a mixture of several low-rank subspaces. This can be unsuitable for real-world instances since the projection matrix may not exactly recover the testing sample. However, if we assume that a batch of testing samples of the same class can be simultaneously obtained, we can make an alternative improvement over the proposed NBCR and the limits just mentioned above can be well addressed. For example, in a video surveillance system, it is difficult to accurately recognize face image from a single shot because of variety of gestures and illumination. If an image set constituted by all face images from cameras in different positions is employed to describe the target, the recognition accuracy will be greatly improved. Moreover, scene analysis based on image set is also considered to be very important. For instance, a robot equipped with an image acquisition device can achieve more accurate scene recognition results by analyzing image set composed of a plurality of images acquired during moving. Instead of using the projection matrix to recovery the testing sample, the low-rank matrix recovery is also employed to recovery the testing samples following formulation (10), which is denoted by NBCR_T. Under this alternative improvement, the step 3 in algorithm 1 is replaced by the following formulation:

$$\min_{\mathbf{M},\mathbf{N}} \|\mathbf{M}\|_* + \beta\|\mathbf{N}\|_1 \text{ Subject to } \mathbf{T} = \mathbf{M} + \mathbf{N} \qquad (10)$$

In this equation, $\mathbf{T} = [\mathbf{t}_1 \ldots \mathbf{t}_n]$ denotes the stacked BoVW representation of testing images within the same class, where $\mathbf{t}_i$ is the i-th testing image. $\mathbf{M}$ is the low-rank matrix, whose column vectors are considered as new representation of testing images. And $\mathbf{N}$ is the sparse error matrix.

5. Experimental results

In this section, we evaluate the proposed method on four public datasets: the Scene-15 dataset, the caltech-101 dataset, the caltech-256 dataset and the UIUC-Sport dataset.

In all experiments, we adopt the wildly used 128 dimensional SIFT descriptor as local features. For the fair comparisons with other methods, we use the same setup as [21] did. The dense SIFT descriptors extracted from 16 × 16 pixel patches are sampled from each image with a step length of 8 pixels. The descriptors are normalized with ℓ^2-norm. All images are preprocessed into gray scale and the max sides of each image are resized to 300 pixels. For the Scene-15, the Caltech 101 and UIUC-Sport dataset, codebook size for non-negative sparse coding is set to 1024. And for the Caltech 256 dataset, the codebook is set to 2048. As [14] did, the SIFT descriptors are randomly chosen to generate the codebook by solving the optimization of problem in Eq. (3) iteratively. Following the common benchmarking procedures, we repeat the experimental process for 5 times by randomly choosing training images and testing images for each dataset. The average per-class classification accuracy for each time is recorded and the mean accuracy and its standard deviation are taken as the indices for comparison.

There are four parameters in our method, i.e. λ and γ for the non-negative sparse coding and the collaborative representation respectively, β and η for the low-rank matrix recovery. We empirically set $\lambda = 0.15$ and $\gamma = 0.001$ for all the fours datasets. The parameter β balances the low-rank term and the sparsity error term while η balances the low-rank matrix approximation and matrix incoherence, which play key role in the method. Generally, the value β should be smaller with the dimension reducing. We carefully tune the parameters for good results and the specific parameters values are reported for each dataset.

5.1. Scene-15 dataset

The Scene-15 dataset consists of 4485 images from 15 categories, each of which contains 200 to 400 images with the average size of 300 × 250. The categories vary from outdoor scenes such as Building and mountains to indoor scenes like living room and bedroom. Fig. 1 shows some images from the Scene-15 dataset. We randomly select 100 images per category as training data and the rest are used for testing. The parameters of low-rank matrix recovery are $\beta = 0.05$ and $\eta = 1e^{-5}$.

Table 1 gives the performances of the proposed method as well as several other methods. Generally, it shows that our method achieves good performances. The performances are comparable with or better than that of LLC and the original ScSPM, and they are also better than the learning-free NN approaches, e.g. five percent over the NBNN method. Non-negative sparse coding with max pooling is adopted, which reduces quantization loss in original sparse coding. In addition, by leveraging the low-rank matrix recovery technique, better bases instead of using training images directly can be learned for coding, which is more discriminative in collaborative representation for classification. LScSPM [35] made use of Laplacian sparse coding framework combining with learning phases, which achieved extremely high performance. We also give the classification performances of NBCR_T with different numbers of testing instances, where the digits are the number of available testing samples. It is obvious and reasonable that the rate of classification accuracy improves accordingly with number of testing samples increasing.

5.2. Caltech-101 dataset

The Caltech-101 dataset consists of 101 categories. The number of images per category varies from about 30 to 800, where most of these images are medium resolution, i.e. 300 × 300 pixels. The dataset is more challenging due to high intra-class appearance variability and large number of category. Fig. 2 shows some images of the Caltech-101 dataset. Following the widely adopted configuration, we randomly select 15 and 30 images per-class for training and also select 15 images per-class for testing. The parameters of low-rank matrix recovery for the 15 training images are $\beta = 0.05$, $\eta = 1e^{-3}$ and $\beta = 0.05$, $\eta = 1e^{-5}$ for 30 training images.

Table 2 gives the performances of the proposed method as well as several other works on this dataset. As shown, our method achieves performances outperforming other NN-image (learning-free) methods (original NBNN, Local NBNN) and several learning-based methods (KCSPM, SVM-KNN, ScSPM) in the case of 15 training samples. Besides, it outperforms LLC by 2 percent for 15 training images, but it is inferior to LLC for 30 training images. The reason may be that the low-rank matrix recovery is essentially based on the hypothesis that the data is approximately drawn

Fig. 1. Example images of the Scene-15 Dataset.

from a low-rank subspace. In contrary, the LLC method does not have this assumption. One work worth mentioning is LR-Sc+SPM, which leveraged non-negative sparse coding, low-rank and sparse matrix decomposition techniques. In this method, the information loss through quantization can be effectively avoided. And the discriminative training step provides an additional benefit. As description of the aforementioned sub-section, we also give the classification performances of NBCR_T with different numbers of testing samples.

5.3. Caltech-256 dataset

The Caltech-256 dataset contains 29,780 images within 256 categories. Each category has at least 80 images. Compared with Caltech-101 dataset, Caltech-256 is more challenging dataset with higher intra-class variability and higher object location variability within the image. Fig. 3 shows some example images of this dataset. Like the setup in Section 5.2, we also randomly choose 15 images and 30 images per-class for training and up to 30 images for testing. The size of vocabulary for non-negative sparse coding is set to 2048 given the higher intra-class variability. The parameters of low-rank matrix recovery for the 15 training images are $\beta = 0.1$, $\eta = 1e^{-3}$ and $\beta = 0.1$, $\eta = 1e^{-5}$ for 30 training images.

Table 1
performance comparison on Scene-15 Dataset.

Algorithm	Performance	Learned?
KSPM[21]	81.40 ± 0.50	Yes
ScSPM[6]	80.28 ± 0.93	Yes
LLC[22]	81.5 ± 0.47	Yes
LScSPM[35]	89.75 ± 0.50	Yes
NBNN[4]	75.00 ± 3.30	No
Local NBNN[36]	79.28 ± 2.34	No
NBINN+NIMBLE[37]	78.23 ± 1.00	No
NBCR	80.12 ± 2.84	No
NBCR_T(10)	75.87 ± 4.71	No
NBCR_T(15)	90.67 ± 6.48	No
NBCR_T(20)	92.00 ± 5.76	No

The experimental results on this dataset are listed in Table 3. From this table, we observe that our method also achieves good performance on the dataset on the whole. On one hand, our method outperforms the listed learning-free methods, under the same experimental setting, except Local NBNN with training number of 30. On the other hand, the results of our method are better than most learning-based methods except LR-Sc+SPM with 15 training images. For example, the NBCR outperforms the LLC by 6 percent for 15 training number and 5.5% for 30 training number. And we also notice that with the number of training data increases, improvement of our method descends compared with ScSPM and LLC. The reason may be that the ignored sparse error matrix also includes certain useful information, which is beneficial to the final classification.

5.4. UIUC-Sport dataset

The UIUC-Sport dataset consists of 1792 images within 8 categories. The number of images per-class varies from 137 to 250. The eight categories are badminton, bocce, croquet, polo, rock climbing, rowing, sailing and snowboarding. Fig. 4 shows some example images of this dataset. For experimental settings, in each category 70 images are randomly selected for training and 60 images

Table 2
Performance comparison on Caltech-101 Dataset.

Algorithm	15 training	30 training	Learned?
KSPM[21]	56.40	64.40 ± 0.80	Yes
KCSPM[17]	–	64.14 ± 1.18	Yes
SVM-KNN[38]	59.10 ± 0.60	66.20 ± 0.50	Yes
ScSPM[6]	67.00 ± 0.45	73.20 ± 0.54	Yes
LLC[22]	65.43	73.44	Yes
LR-Sc+SPM[14]	69.58 ± 0.97	75.68 ± 0.89	Yes
NBNN[4]	65.00 ± 1.14	70.40	No
Local NBNN[36]	66.1 ± 1.17	71.9 ± 0.6	No
NBCR	67.42 ± 1.56	71.74 ± 1.78	No
NBCR_T(10)	63.67 ± 3.54	68.61 ± 2.12	No
NBCR_T(15)	69.64 ± 2.41	73.61 ± 3.23	No
NBCR_T(20)	73.12 ± 3.12	76.79 ± 2.62	No

Fig. 2. Example images of the Caltech-101 Dataset.

Fig. 3. Example images of the Caltech-256 Dataset.

randomly selected as testing data. The parameters of low-rank matrix recovery are $\beta = 0.1$ and $\eta = 1e^{-3}$.

Table 4 reports the performances of all the methods for comparison on the UIUC-Sport dataset. Different from the former three datasets, we can see that the proposed method is only superior to ScSPM, where the absolute values of the negative coefficients in the sparse representation are directly used. As pointed in [14], with difference of non-negative sparse coding, this behavior impairs the consistency between sparse coding and max pooling. In this experiment, more training images (70 images) are used as sub-dictionary for collaborative representation, while the data with more complicated structure cannot be considered as being drawn from a single underlying subspace. The low-rank matrix recovery technique and the learned projection matrix incur certain information loss. We also give the classification performances of NBCR_T with different numbers of testing samples. From the last three rows in Table 4 we can see that this method achieves much better results with number of testing image increasing, which makes full use homogeneity of testing images.

6. Discussion

As is known to all, in many tasks such as image classification and other multimedia content analysis [40,41], the number of training images has large influences on the final performance. From the above experiments, we can see that the performances of all the methods consistently increase with larger number of training images. However, during the process of training classifier in real-world applications, it is sometimes difficult to obtain enough training images, which could dampen the classifier's performances. This limitation is especially severe for datasets with large number of category and diversity.

To further evaluate our method, we randomly select p images ($p = 2$ –10) for training on Scene-15 and Caltech-101 dataset and the number of testing images are set to 20. The NBNN, ScSPM and LLC are employed for comparison with the proposed method including NBCR and NBCR_T. The performances of all the methods for comparison are shown in Figs. 5 and 6. From the results, we observe that all the classification accuracies consistently increase with the number of training image for all methods. Our methods are comparable or better than other three methods along with the increasing training number on the Scene-15 dataset. As for the Caltech-101 dataset, our method performs much better than their counterparts.

Among the methods mentioned in Figs. 5 and 6, ScSPM reduces quantization loss by sparse coding, but the ignored structural information from the same class and the max pooling strategy also introduce information loss during the encoding process for image representation. And similar features may vary a lot after encoding because of over-complete dictionary. LLC integrates locality constraint during sparse coding, where K nearest codes are used to encode a descriptor. Generally K is smaller than the descriptor dimension, so the reconstruction of a descriptor may lead to large deviation, which is an under-determined problem and affects the final classification performance.

Our method effectively avoids the above problems due to the following aspects: on one hand, the proposed method better

Table 3
Performance comparison on Caltech-256 Dataset.

Algorithm	15 training	30 training	Learned?
KSPM[21]	–	34.10	Yes
KCSPM[17]	–	27.17 ± 0.46	Yes
ScSPM[6]	27.73 ± 0.51	34.02 ± 0.35	Yes
LLC[22]	27.74 ± 0.32	32.07 ± 0.24	Yes
LR-Sc+SPM[14]	35.31 ± 0.70	–	Yes
NBNN[4]	30.5	37.00	No
Local NBNN[36]	33.5 ± 0.9	40.1 ± 0.1	No
NBCR	33.89 ± 0.87	37.51 ± 0.74	No
NBTCR(10)	27.11 ± 3.34	31.76 ± 2.71	No
NBTCR(15)	32.08 ± 1.87	35.54 ± 3.11	No
NBTCR(20)	34.65 ± 2.34	38.38 ± 3.54	No

Table 4
Performance comparison on UIUC-Sport Dataset.

Algorithm	Performance	Learned?
ScSPM[6]	82.74 ± 1.46	Yes
HIK+OCSVM[39]	83.54 ± 1.13	Yes
LScSPM[35]	85.31 ± 0.51	Yes
LR-Sc+SPM[14]	86.69 ± 1.66	Yes
NBCR	82.96 ± 2.43	No
NBCR_T(15)	82.67 ± 3.48	No
NBCR_T(20)	84.12 ± 2.76	No
NBCR_T(60)	88.53 ± 1.26	No

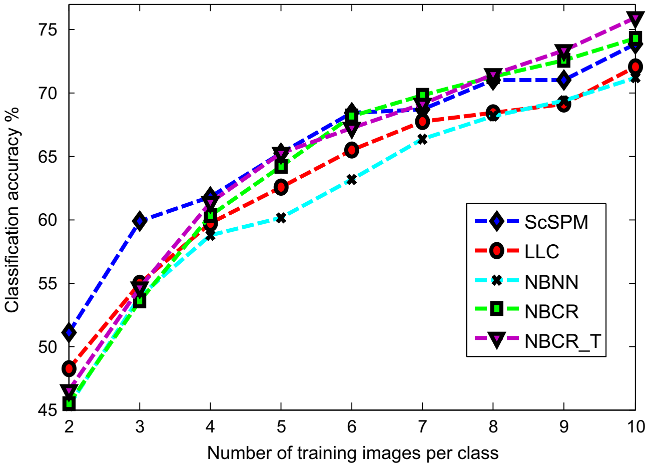


Fig. 5. Performance comparison on Scene-15.

Fig. 4. Example images of the UIUC Dataset.

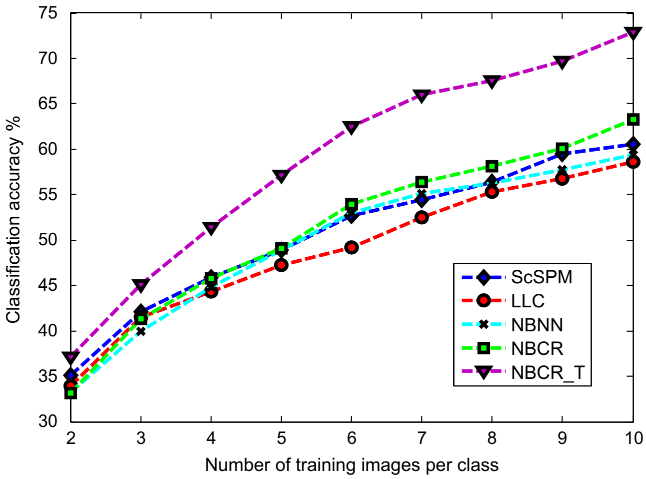


Fig. 6. Performance comparison on Caltech-101.

mines the structurally consistent information of training images within the same class. By low-rank matrix recovery, a more discriminative dictionary can be obtained for collaborative representation, which also has a closed-form solution. On the other hand, as a variant of NBNN, the proposed algorithm is learning-free, which can naturally handle dataset with a large number of categories and avoid common problems in learning-based methods such as parameter overfitting or non-balance between training and testing data.

7. Conclusion

In this paper, we present a novel learning-free image classification algorithm, i.e. NBCR, by leveraging the low-rank matrix recovery and the collaborative representation jointly under the NBNN framework. To reduce the information loss introduced by sparse coding and max pooling, non-negative sparse coding is adopted to obtain robust representation of images. Moreover, we use low-rank matrix recovery technique to get more discriminative dictionary for collaborative representation, which makes full use of underlying structural information of training images within the same class. The testing image recovered by a learned projection matrix is classified by responses over the bases. The experimental results on several public datasets demonstrate the effectiveness of the proposed method. Specially, if a batch of images of the same class is simultaneously tested, higher classification accuracy can be achieved through the proposed NBCR_T method. So our method is also suitable for image set classification.

References

[1] Richang Hong, Jinhui Tang, Hung-Khoon Tan, Chong-Wah Ngo, Shuicheng Yan, Tat-Seng Chua, Beyond search: event-driven summarization for web videos, Trans. Multimed. Comput. Commun. Appl. 7 (4) (2011) 35.

[2] Richang Hong, Meng Wang, Y.u.e. Gao, Dacheng Tao, Xuelong Li, Xindong Wu, Image annotation by multiple-instance learning with discriminative feature mapping and selection, IEEE Trans. Cybern. 44 (5) (2014) 669–680.

[3] R. Hong, J. Pan, S. Hao, et al., Image quality assessment based on matching pursuit[J], Inf. Sci. 273 (2014) 196–211.

[4] O. Boiman, E. Shechtman, M. Irani In defense of nearest-neighbor based image classification[C], in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8.

[5] J. Sivic, A. Zisserman Video google: a text retrieval approach to object matching in videos, in: Proceedingsof the Ninth IEEE International Conference on Computer Vision, 2003, pp. 1470–1477.

[6] J. Yang, K. Yu, Y.Gong , et al. Linear spatial pyramid matching using sparse coding for image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 1794–1801.

[7] J. Wright, Y. Ma, J. Mairal, et al., Sparse representation for computer vision and pattern recognition, Proc. IEEE 98 (6) (2010) 1031–1044.

[8] D. Zhang, M. Yang, X. Feng Sparse representation or collaborative representation: which helps face recognition?, in: Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV), 2011, pp. 471–478.

[9] J. Wright, A. Ganesh, S. Rao, et al., Robust principal component analysis: exact recovery of corrupted low-rank matrices via convex optimization, Adv. Neural Inf. Process. Syst. (NIPS) (2009) 2080–2088.

[10] X. Cui, J. Huang, S. Zhang, et al., Background Subtraction Using Low Rank and Group Sparsity Constraints Computer Vision–ECCV 2012, Springer, Berlin Heidelberg (2012) 612–625.

[11] T. Zhang, B. Ghanem, S. Liu, et al., Low-rank sparse Learning for Robust Visual Tracking Computer Vision–ECCV 2012, Springer, Berlin Heidelberg (2012) 470–484.

[12] G. Zhu, S. Yan, Y. Ma Image tag refinement towards low-rank, content-tag prior and error sparsity, in: Proceedings of the International Conference on Multimedia, ACM, 2010, pp. 461–470.

[13] C.F. Chen, C.P. Wei, Y.C.F. Wang Low-rank matrix recovery with structural incoherence for robust face recognition, in: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2012, pp. 2618–2625.

[14] C. Zhang, J. Liu, Q. Tian, et al. Image classification by non-negative sparse coding, low-rank and sparse decomposition, in: Proceedings of the Computer Vision and Pattern Recognition, CVPR, 2011, pp. 1673–1680.

[15] F. Perronnin, Universal and adapted vocabularies for generic visual categorization, IEEE Trans. Pattern Anal. Mach. Intell. 30 (7) (2008) 1243–1256.

[16] F. Jurie, B. Triggs Creating efficient codebooks for visual recognition, in: Proceedings of the Tenth IEEE International Conference on Computer Vision, ICCV, 2005, 1, pp. 604–610.

[17] J.C. van Gemert, C.J. Veenman, A.W.M. Smeulders, et al., Visual word ambiguity, IEEE Trans. Pattern Anal. Mach. Intell. 32 (7) (2010) 1271–1283.

[18] F. Perronnin, J. Sánchez, T. Mensink, Improving the Fisher Kernel for large-scale image classification Computer Vision–ECCV 2010, Springer, Berlin Heidelberg (2010) 143–156.

[19] Y. Lin, F. Lv, S. Zhu, et al. Large-scale image classification: fast feature extraction and svm training, in: Proceedings 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2011, pp. 1689–1696.

[20] W. Jun, W. Shitong, F. Chung, Positive and negative fuzzy rule system, extreme learning machine and image classification, Int. J. Mach. Learn. Cybern. 2 (4) (2011) 261–271.

[21] S. Lazebnik, C. Schmid, J. Ponce Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, 2, pp. 2169–2178.

[22] J. Wang, J. Yang, K. Yu, et al. Locality-constrained linear coding for image classification, in: Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2010, pp. 3360–3367.

[23] H. Lee, A. Battle, R. Raina, et al., Efficient sparse coding algorithms, Adv. Neural Inf. Process. Syst. 19 (2007) 801.

[24] Y. Liu, F. Wu, Z. Zhang, et al. Sparse representation using nonnegative curds and whey, in: Proceedings of the 2010 IEEE Conference onComputer Vision and Pattern Recognition, CVPR, 2010, pp. 3578–3585.

[25] Z. Jiang, Z. Lin, L.S. Davis Learning a discriminative dictionary for sparse coding via label consistent K-SVD, in: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2011, pp. 1697–1704.

[26] X. Zhou, C. Yang, W. Yu, Moving object detection by detecting contiguous outliers in the low-rank representation, IEEE Trans. Pattern Anal. Mach. Intell. 35 (3) (2013) 597–610.

[27] M. Mardani, G. Mateos, G.B. Giannakis, Recovery of low-rank plus compressed sparse matrices with application to unveiling traffic anomalies, IEEE Trans. Inf. Theory 59 (8) (2013) 5186–5205.

[28] Z. Lin, A. Ganesh, J. Wright, et al., Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix, Comput. Adv. Multi-Sens. Adapt. Process. (2009), 3rd IEEE International Workshop on pp. 213–216.

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

[30] Y. Zhang, Z. Jiang, L.S. Davis Learning structured low-rank representations for image classification, in: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2013, pp. 676–683.

[31] J. Chen, Z. Yi, Sparse representation for face recognition by discriminative low-rank matrix recovery, J. Vis. Commun. Image Represent. 25 (5) (2014) 763–773.

[32] E.J. Candès, X. Li, Y. Ma, et al., Robust principal component analysis? J. ACM 58 (3) (2011) 11.

[33] B.K. Bao, G. Liu, C. Xu, et al., Inductive robust principal component analysis, IEEE Trans. Image Process. 21 (8) (2012) 3794–3800.

[34] G. Liu, Z. Lin, S. Yan, et al., Robust recovery of subspace structures by low-rank representation, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 171–184.

[35] S. Gao, I.W. Tsang, L.T. Chia, et al. Local features are not lonely–Laplacian sparse coding for image classification, in: Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2010 pp. 3555–3561.

[36] S. McCann, D.G. Lowe Local naive bayes nearest neighbor for image classification, in: Proceedings of the 2012 IEEE Conference onComputer Vision and Pattern Recognition, CVPR, 2012 pp. 3650–3656.

[37] R. Timofte, L. Van Gool Iterative nearest neighbors for classification and dimensionality reduction, in: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2012 pp. 2456–2463.

[38] H. Zhang, A.C. Berg, M. Maire, et al. SVM-KNN: discriminative nearest neighbor classification for visual category recognition, in: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, 2, pp. 2126–2136.

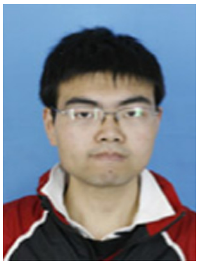
[39] J. Wu, J.M. Rehg Beyond the euclidean distance: creating effective visual codebooks using the histogram intersection kernel, in: Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, 2009 pp. 630–637.

[40] R. Hong, Z.J. Zha, Y. Gao, et al., Multimedia encyclopedia construction by mining web knowledge[J], Signal Process. 93 (8) (2013) 2361–2368.

[41] R. Hong, M. Wang, G. Li, et al., Multimedia question answering, IEEE MultiMedia 19 (4) (2012) 72–78.

Xu Zhang received the B.E. degree in computer science and technology from the University of Science and Technology of China, Hefei, China, in 2004. He is currently pursuing the Ph.D. degree in the School of Computer and Information at Hefei University of Technology, Hefei, China. His research interests include computer vision, machine learning, and pattern recognition.

Shijie Hao is a lecturer of School of Computer and Information Science, Hefei University of Technology (HFUT). He received the Ph.D. degree in Singal and Information Processing from HFUT in 2012. His research interests include image processing, multimedia content analysis, and machine learning.



Chenyang Xu is pursuing his M.D. degree in School of Electronic and Information Engineering, Xi'an Jiaotong University. His research interests include image and video pattern recognition.

Xueming Qian is a professor at School of Electronic and Information Engineering, Xi'an Jiaotong University. He received the Ph.D. degree in the School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, China, in 2008. His research interests include video/image analysis and retrieval, mobile multimedia search and mining. He published more than 30 papers in international conferences and journals.

Meng Wang is a professor in the Hefei University of Technology, China. He received the B.E. degree and Ph. D. degree in the Special Class for the Gifted Young and the Department of Electronic Engineering and Information Science from the University of Science and Technology of China (USTC), Hefei, China, respectively. He previously worked as an associate researcher at Microsoft Research Asia, and then a core member in a startup in Silicon Valley. After that, he worked in the National University of Singapore as a senior research fellow. His current research interests include multimedia content analysis, search, mining, recommendation, and large-scale computing. He has authored more than 150 book chapters, journal and conference papers in these areas. He received the best paper awards successively from the 17th and 18th ACM International Conference on Multimedia, the best paper award from the 16th International Multimedia Modeling Conference, the best paper award from the 4th International Conference on Internet Multimedia Computing and Service, and the best demo award from the 20th ACM International Conference on Multimedia.

Jianguo Jiang is a professor of School of Computer and Information Science, Hefei University of Technology (HFUT). He is head of the TI-HFUT DSP Laboratory in Engineering Research Center of Safety Critical Industrial Measurement and Control Technology, Ministry of Education. His research interests include image processing and multi-agent system.