Image Classification by Non-Negative Sparse Coding, Low-Rank and Sparse Decomposition

Chunjie Zhang[1], Jing Liu[1], Qi Tian[2], Changsheng Xu[1],Hanqing Lu[1], Songde Ma[1]
[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences,
P.O.Box 2728, Beijing, China
[2]University of Texas at San Antonio, One UTSA Circle, San Antonio Texas, 78249-USA
{cjzhang, jliu, csxu, luhq}@nlpr.ia.ac.cn, qitian@cs.utsa.edu, masd@most.cn

Abstract

We propose an image classification framework by leveraging the non-negative sparse coding, low-rank and sparse matrix decomposition techniques (LR-Sc$^+$SPM). First, we propose a new non-negative sparse coding along with max pooling and spatial pyramid matching method (Sc$^+$SPM) to extract local features' information in order to represent images, where non-negative sparse coding is used to encode local features. Max pooling along with spatial pyramid matching (SPM) is then utilized to get the feature vectors to represent images. Second, motivated by the observation that images of the same class often contain correlated (or common) items and specific (or noisy) items, we propose to leverage the low-rank and sparse matrix recovery technique to decompose the feature vectors of images per class into a low-rank matrix and a sparse error matrix. To incorporate the common and specific attributes into the image representation, we still adopt the idea of sparse coding to recode the Sc$^+$SPM representation of each image. In particular, we collect the columns of the both matrixes as the bases and use the coding parameters as the updated image representation by learning them through the locality-constrained linear coding (LLC). Finally, linear SVM classifier is leveraged for the final classification. Experimental results show that the proposed method achieves or outperforms the state-of-the-art results on several benchmarks.

1. Introduction

As a fundamental problem in computer vision, image classification has attracted a lot of attention in recent years. Among many image representation models, the bag of visual words (BoW) model [1] has been widely used by many researchers [2-4] and shown very good performance. The BoW model contains mainly two modules: (i) codebook generation and quantization of features extracted from local image patches; (ii) histogram based image representation and prediction. Recently, it has been shown that combining the two modules with sparse representation is very effective and can achieve the state-of-the-art performance.

As to the first module of the BoW model, *k*-means is usually used to generate codebook and quantize visual descriptors extracted from local image patches by nearest-neighbor search. A histogram is then computed to represent each image by counting the occurrence number of each visual word within this image. Recently, Yang *et al*. [4] developed an extension by generalizing vector quantization to sparse coding. By using sparse coding instead of *k*-means, they are able to learn the optimal codebook and coding parameters for local features simultaneously, hence are able to reduce the quantization loss. Multi-scale max pooling is then used to get the feature representation of images. However, sparse coding has no constraints on the sign of coding coefficients. To satisfy the objective of sparse coding, negative coefficients are sometimes needed, while large numbers of zero coefficients are inevitable. Since non-zero components typically provide useful information, the encoding process with max pooling will bring the loss in terms of those negative components, and further degrade the classification performance.

Instead of learning sparse representations for local features [4], the use of sparse representation for the final classification has also been widely applied to many visual applications and can achieve the state-of-the-art performances, *e.g.*, image restoration [5] and classification tasks [6-11]. These holistically sparse representations on the whole image ensure robustness to occlusions and image corruptions. Training images are often chosen as the bases for sparse representation and test images are then classified by assigning the class with the lowest reconstruction error. Ideally, a test image can be reconstructed by the training samples and only the coefficients of the samples within the same class may be nonzero. This means a test image can be sufficiently recon-

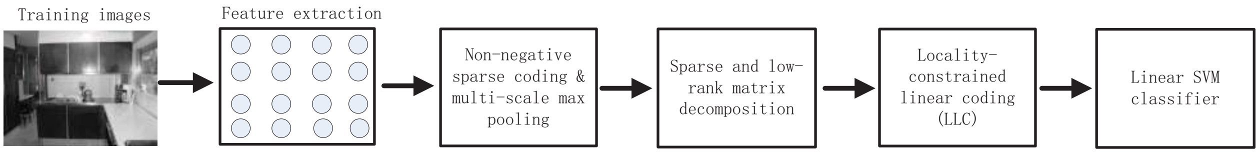
Training images
Feature extraction
Non-negative sparse coding & multi-scale max pooling
Sparse and low-rank matrix decomposition
Locality-constrained linear coding (LLC)
Linear SVM classifier

Figure 1. The flowchart of the proposed method.

structed by the training images of the same class. However, since images are often contaminated with noise; besides, there are often multiple objects in an image with different poses and occlusions. Sometimes using the training images as the bases is not discriminative enough to boost the final classification performance. Moreover, images of the same class often share a lot of similarities and correlate with each other, hence exhibit *degenerated structure* [6]. This semantic information of images can help make correct classification if computed correctly.

In this paper, we propose a new image classification framework by leveraging the non-negative sparse coding, low-rank and sparse matrix decomposition techniques (LR-Sc^{+}SPM). Figure 1 shows the flowchart of the proposed method. Our proposed framework consists of two contributions. First, we extend the recent work on image classification [4] and present to use non-negative sparse coding along with max pooling method to reduce the information loss during the encoding process for image representation.

The second is our main contribution. We propose a new image classification method method by using the low-rank and sparse matrix decomposition technique. Our work is motivated by the observation that: (i) images of the same class often correlate with each other. Ideally, if we stack the BoW representation of images within the same class into a matrix, this matrix will be low-rank; (ii) one image contains only a limited number of objects and a limited type of noise. This results in the characteristics of noise sparsity for the stacked BoW matrix. This low-rank and noise information can be utilized for better image representation than directly using the BoW representation of training images. Specially, to get more discriminative sparse coding bases with the BoW representation of images, we leverage the low-rank and sparse matrix decomposition technique to decompose the BoW representation of images within the same class into a low rank matrix and a sparse error matrix. We then use these bases to encode the BoW representation of images with sparsity and locality constraints. These coding parameters are used to represent images and linear SVM classifier is then utilized to predict the category labels of images. Experimental results on four public datasets demonstrate the effectiveness of the proposed method.

The rest of the paper is organized as follows. Section 2 introduces some related work. Section 3 presents the proposed non-negative sparse coding spatial pyramid matching method (Sc^{+}SPM). Section 4 shows the proposed image classification method by low-rank and sparse matrix decomposition. Experimental results are given in Section 5. Finally we conclude in Section 6.

2. Related Work

The use of the bag-of-visual words (BoW) model [1] has been proven very useful for image classification. Over the past few years, many works have been done to improve the performance of the BoW model. Some tried to learn discriminative visual codebooks for image classification [12, 13]. Co-occurrence information of visual words was also modeled in a generative framework [14, 15]. Others tried to learn discriminative classifiers by considering the spatial information and correlations among visual words [2-4, 7, 10-11]. To overcome the loss of spatial information in the BoW model, motivated by Grauman and Darrell's [3] pyramid matching in feature space, Lazebnik *et al.* [2] proposed the spatial pyramid matching (SPM). Since its introduction, SPM has been widely used and proven very effective.

Recently, Yang *et al.* [4] proposed an extension of the SPM approach by leveraging sparse coding and achieved the state-of-the-art performance for image classification when only one type of local feature (SIFT) is used. This method can automatically learn the optimal codebook and search for the optimal coding weights for each local feature. After this, max pooling along with SPM is used to get the feature representation of images. Inspired by this, Wang *et al.* [16] proposed to use locality to constrain the sparse coding process which can be computed faster and yields better performance. [11, 17] also tried to jointly learn the optimal codebooks and classifiers. However, sparse coding [18] has no constraints on the sign of the coding parameters, negative parameters are sometimes needed to satisfy the sparse coding constrains. For some particular applications [19], non-negative sparse coding [20] is needed.

Not only has sparse coding been used for local features, but also it has been widely used holistically on the entire image. Wright *et al.* [6] tried to do face recognition as finding a sparse representation of the test image by treating the training set as the bases and impressive results were achieved. Bradley and Bagnell [9] tried to train a compact codebook using sparse coding. Yuan and Yan [7] made visual classification with multi-task joint sparse representation by fusing different types of features. Liu *et al.* [19] tried to learn sparse and nonnegative representations of images by solving a set of regression type nonnegative matrix factorization problems. However, because images are often corrupted with noise and there are often multiple objects in one image with different poses and occlusions, sparse coding by directly using the training images as the bases is not discriminative enough to boost the final classification performance.

There has been a lot of work on how to learn good bases for visual applications, *e.g.*, clustering and classification. Some [6-8, 19, 21] tried to use the training samples as the bases directly. To code a new sample, [6-8, 19] used all the training samples while locally linear embedding (LLE) [21] uses the *k* nearest neighbors. Others [18, 20, 22-25] utilized the training data to learn the bases, *e.g.*, *k*-means, Gaussian mixture model (GMM) and sparse coding [18, 20].

Over the past few years, the low-rank matrix recovery problem has been widely studied [22-25] and successfully applied to many applications, such as image processing [22], web data mining [26], and bioinformatic data analysis [27]. It tries to recover a low-rank matrix with an unknown fraction of its entries being arbitrarily corrupted. Under surprisingly broad conditions, this problem can be exactly solved via convex optimization which minimizes a combination of the nuclear norm and the ℓ^1norm [22, 23].

3. Non-negative Sparse Coding Spatial Pyramid Matching (Sc^+SPM)

k-means clustering has been widely used for codebook generation in the BoW model. Let $X = [x_1, x_2, ..., x_N](x_i \in \mathbb{R}^{D\times 1})$ be the set of N local image descriptors of D dimensions. Typically the vector quantization (VQ) by *k*-means clustering method solves the following optimization problem as:

$$\min_{U,V} \sum_{n=1}^{N} \parallel x_n - u_n V \parallel^2 \tag{1}$$

$$s.t. Card(u_n) = 1, |u_n| = 1, u_n \succeq 0, \forall n$$

where $V = [v_1, v_2, ..., v_K]$ $(v_i \in \mathbb{R}^{D\times 1})$ are the K cluster centers to be learned and $U = [u_1, u_2, ..., u_N]$ $(u_i \in \mathbb{R}^{K\times 1})$ are the cluster membership indicators. $Card(u_n) = 1$ is the cardinality constraint. However, this constraint is too strict because each local feature can be assigned to only one visual word, especially for the local features located at the boundary of clusters. To alleviate the quantization loss of VQ, Yang *et al.* [4] relaxed the constraint by using a ℓ^1-norm regularization and turned the VQ into sparse coding as:

$$\min_{U,V} \sum_{n=1}^{N} \parallel x_n - u_n V \parallel^2 + \lambda \parallel u_n \parallel_1 \tag{2}$$

$$s.t. \parallel v_k \parallel^2 \leq 1, \forall k$$

where λ is the regularization parameter. Spatial pyramid matching with max pooling is then used to obtain nonlinear codes to represent images, which achieved the state-of-the-art performances on several datasets when only one type of local feature is used.

However, there is one problem with this sparse coding plus max pooling strategy. To satisfy the objective of sparse coding, negative coefficients are sometimes needed. This coding strategy is suboptimal because max pooling is then used to extract the feature representation of images. Zero (or small positive) coefficients of sparse coding indicate the corresponding bases have no (or very small) influence. However, since zero (or positive value) is always larger than negative values, max pooling strategy will choose zero (or positive value) instead of negative values. Because most of the coefficients in sparse coding are zero, this phenomenon will happen with high probability. That means some useful information is lost which hinders the final classification performance.

To alleviate the information loss of the sparse coding plus max pooling strategy [4], we propose to use non-negative sparse coding instead. The non-negative sparse coding tries to solve the following optimization problem as:

$$\min_{U,V} \sum_{n=1}^{N} \parallel x_n - u_n V \parallel^2 + \lambda \parallel u_n \parallel_1 \tag{3}$$

$$s.t. \parallel v_k \parallel^2 \leq 1, u_n \succeq 0, \forall k, n$$

We follow the same optimization procedure as did in [18] and solve it iteratively by alternatively optimizing over U or V while keeping the other fixed. When U is fixed, this problem is reduced to a least square problem with quadratic constraints as:

$$\min_{V} \parallel X - UV \parallel_F^2 \ s.t. \parallel v_k \parallel^2 \leq 1 \tag{4}$$

Where $\parallel . \parallel_F$ is the Frobenius norm. This can be efficiently solved by using the Lagrange dual method. When V is fixed, we solve the optimization problem (3) by optimizing over each local feature individually as:

$$\min_{u_n} \parallel x_n - u_n V \parallel^2 + \lambda \parallel u_n \parallel_1 \ s.t. u_n \succeq 0 \tag{5}$$

This is a linear regression problem with ℓ^1 norm regularization and non-negative constraints on the coefficients. We adopt the feature-sign search algorithm with projected gradient descent to solve this problem.

Due to the large amount of local features, we only sample some features to learn the codebook. We choose around 45,000 SIFT features randomly chosen to train the codebook by iteratively solving the optimization of problem (4) and problem (5). The iteration number is set to 50. After the codebook has been learned, we can code the local features of each image. Spatial pyramid matching with max pooling is then used to get the BoW representation of images.

4. Image Classification by Low-rank and Sparse Matrix decomposition

In this section, we will first introduce the low-rank and sparse matrix decomposition and then use it for image classification.

4.1. Low-rank and sparse matrix decomposition

Not only has sparse coding been used for local features, but also it has been widely used holistically on the entire image. Training samples are often chosen as the bases for sparse coding or its variants when it is applied on the entire image. However, images are often corrupted with noise and there are often multiple objects in one image with different poses and occlusions, even if they are of the same class. That is, there exist the correlated (or common) items and the specific (or noisy) items among images of the same class. The both parts are more robust and discriminative for image classification than directly using the feature representation of training images because the two parts capture the correlated and specific attributes of images in the same class.

Motivated by these observations, we propose to use the low-rank and sparse matrix decomposition technique to decompose the features of images within each class into a low-rank matrix and a noise matrix. Because images of the same class share a lot of similarities and often correlate with each other, as shown in [4, 6, 10, 16]. Besides, images often undergo gross corruption (such as occlusion or illumination change) which often happens in modern visual applications. This means noises in images may have arbitrarily large magnitude. Here we consider an idealized version and assume the noise is sparse but unknown. Formally, let $H_i = [h_{i,1}, h_{i,2}, ..., h_{i,p_i}]$ be the stacked column vectors of the BoW representations of p_i training images of the i-th class, we try to decompose it as:

$$H_i = L_i + N_i \tag{6}$$

Where L_i and N_i are the low-rank matrix and the noise matrix of the i-th class.$i \in \{1, ..., M\}$ where M is the number of image classes. This problem can be solved by

$$\min_{L_i,N_i} rank(L_i) + \gamma \parallel N_i \parallel_0 \tag{7}$$

$$s.t. H_i = L_i + N_i$$

Here the $\parallel . \parallel_0$ counts the nonzero elements in the error matrix and $\gamma > 0$ is the parameter that balances the rank term and the sparsity error term. However, this problem is non-convex and very hard to solve. Recently, it is shown by [22] that under certain conditions, solving

$$\min_{L_i,N_i} \parallel L_i \parallel_* + \gamma \parallel N_i \parallel_1 \tag{8}$$

$$s.t. H_i = L_i + N_i$$

exactly recovers the low-rank matrix L_i and the sparse matrix N_i. $\parallel . \parallel_*$ is the nuclear norm defined as the sum of all singular values.The augmented lagrange multiplier method (ALM) proposed by Lin *et al* [24] can be adopted to solve this problem.

4.2. Low-rank and sparse matrix decomposition for image classification

After the low-rank matrix L_i and the sparse matrix N_i for each class have been learned, we can use them to encode the histogram information of images. Let $L = [L_1, L_2, ..., L_M]$ and $N = [N_1, N_2, ..., N_M]$, the new bases for sparse coding the BoW representation of images are then defined as $B = [L, N]$. If one image belongs to the i-th class, it will probably be reconstructed by vectors of the i-th low-rank matrix L_i and sparse matrix N_i instead of vectors of other classes. We use the Locality-constrained Linear Coding (LLC) method[16] to reconstruct the pooled feature representation of images by leveraging the bases learned above. As shown by Yu *et al.* [10], locality has been shown to lead to good performance. LLC uses local bases instead of all the bases for reconstruction. Formally, LLC tries to solve:

$$\min_{c_p} \sum_{p=1}^{P} \parallel h_p - Bc_p \parallel^2 + \beta \parallel d_p \bigodot c_p \parallel^2 \tag{9}$$

$$s.t. 1^T c_p = 1, \forall p$$

where $\bigodot$ denotes the element-wise multiplication which favors near-by bases. d_p is defined as:

$$d_p = exp(\frac{dist(h_p, B)}{\sigma}) \tag{10}$$

Where $dist(h_p, B) = [dist(h_p, b_1), ..., dist(h_p, b_T)]$, T is the number of bases and $dist(h_p, b)$ is the Euclidean distance between h_p and b. σ is the weight decay speed for the locality adaptor. We follow the same approximated method in [16] by firstly choose the k-nearest neighbors of h_p and then use its nearest neighbors for reconstruction. This reduces the computation complexity and speeds up the coding phase. In our experiments, we empirically set k to 20. Linear SVM classifier is then used to predict the category of images due to its advantages in speed and good performance for the sparse coding parameters [4, 11, 16, 28].

5. Experiments

In this section, we evaluate the proposed non-negative sparse coding, low-rank and sparse matrix decomposition method (LR-Sc$^+$SPM) on four public datasets: The Scene-15 dataset [2], the UIUC-Sport dataset [29], the Caltech-101

Figure 2. Example images for the Scene-15 dataset.

dataset [30] and the Caltech-256 dataset [31]. The codebook size for non-negative sparse coding is set to 1,024, as in [4, 16, 28]. As to feature extraction, we use the same setup as [4] did because this setup has been proven very effective on these datasets and densely compute SIFT descriptors on overlapping 16×16 pixels with an overlap of 6 pixels. All images are processed in gray scale. These extracted features are then normalized with ℓ^2 norm. For SPM, we follow Lazebnik *et al*. [2] and use the first 3 layers (1×1, 2×2, 4×4) with the same weight for each layer. We use the multi-class linear SVM provided by Yang *et al*. [4] due to its advantages in speed and good performance in max pooling based image classification. Following the common benchmarking procedures, we repeat the experimental process with randomly selecting the training and testing images to obtain reliable results. We record the average per-class classification rates for each run and report our final results by the mean and standard deviation of the classification rates.

5.1. Scene-15 dataset

The major sources of pictures in the Scene-15 dataset include the COREL collection, personal photographs and Google Image Search. Each category has 200 to 400 images with the average image size of 300×250 pixels. The total image number is 4,485. Figure 2 shows some example images of the Scene-15 dataset. We use the same number of training images per category as [2, 4, 28] did and randomly choose 100 image per category and test on the rest. This process is repeated for ten times to obtain reliable results.

Table 1 gives the performance comparison of the proposed method and several other methods [2, 4, 32, 33] on the Scene-15 dataset. The proposed method outperforms the ScSPM by about 10 percent which demonstrates the effectiveness of our method. Since we use non-negative sparse coding instead of sparse coding along with spatial pyramid matching and max pooling, we are able to preserve more information and reduce the quantization loss. Besides, by leveraging the low-rank matrix recovery technique, we are able to learn better bases instead of using the training images as the bases directly. This makes the final image representation more discriminative hence is able to improve the image classification performance.

Table 1. Performance comparison on the Scene-15 dataset. (KSPM: Spatial pyramid matching and kernel SVM classifier; ScSPM: Sparse coding along with spatial pyramid matching; HIK+OCSVM: Histogram intersection kernel and one class SVM for local feature quantization; KCSPM: Kernel codebook and spatial pyramid matching;LScSPM: Laplacian sparse coding and spatial pyramid matching)

Algorithm	Performance
KSPM[2]	81.40 ± 0.50
ScSPM[4]	80.28 ± 0.93
HIK+OCSVM[32]	84.00 ± 0.46
KCSPM[33]	76.70 ± 0.40
LScSPM[28]	89.75 ± 0.50
LR-Sc^+SPM	**90.03± 0.70**

Figure 3. Example images for the UIUC-Sports dataset.

Table 2. Performance comparison on the UIUC-Sport dataset.

Algorithm	Performance
HIK+OCSVM[32]	83.54 ± 1.13
ScSPM[4]	82.74 ± 1.46
LScSPM[28]	85.31 ± 0.51
Sc^+SPM	83.77 ± 0.97
LR-Sc^+SPM	**86.69± 1.66**

5.2. UIUC-Sport dataset

The UIUC-Sport dataset contains eight categories with 1792 images. The eight categories are *badminton, bocce, croquet, polo, rock climbing, rowing, sailing and snow boarding*. The number of images ranges from 137 to 250. Figure 3 shows some example images of this dataset. We randomly select 70 images from each class for training [28, 29] and test on the rest images. We repeat this process for five times.

Table 2 gives the performance comparison of the proposed method and several other methods [4, 28, 32] on the UIUC-Sport dataset. We also give the performance of only using the non-negative sparse coding(Sc^+SPM) for image classification. We can see that the proposed method can

Table 3. Performance comparison on the Caltech-101 dataset (NBNN: Nearest-neighbor in local image feature space; SVM-KNN: A hybrid NN-based and SVM-based method; KMTJSRC: Kernel multi-task joint sparse representation; LLC: Locality-constrained linear coding).

Algorithm	15 training	30 training
KSPM[2]	56.40	64.40 ± 0.80
KCSPM[33]	$-$	64.14 ± 1.18
NBNN[34]	65.00 ± 1.14	70.40
SVM-KNN[35]	59.10 ± 0.60	66.20 ± 0.50
KMTJSRC[7]	65.00 ± 0.70	$-$
ScSPM[4]	67.00 ± 0.45	73.20 ± 0.54
LLC[16]	65.43	73.44
LR-Sc^+SPM	**69.58± 0.97**	**75.68± 0.89**

achieve or outperform the state-of-the-art performance on this dataset. This demonstrates the effectiveness of our method.

5.3. Caltech-101 dataset

The Caltech-101 dataset contains 101 classes with high intra-class appearance shape variability. The number of images per category varies from 31 to 800 images and most of these images are medium resolution, *i.e.* 300×300 pixels. We follow the common experimental setup as did in [4, 7, 17, 19], and randomly choose 15 and 30 images per category for training and up to 30 images for test. This process is repeated for 5 times.

Table 3 gives the performance comparison of the proposed method and several other methods [2, 4, 7, 16, 33-35] on the Caltech-101 dataset. As shown, our method achieves the state-of-the-art performance and outperforms ScSPM by 2.5 percent for 15 training images and 2.4 percent for 30 training images. Besides, our method also outperforms the KMTJSRC [7] which used the BoW representation of images for sparse representation directly. One work worth mentioning is the NBNN [34], where the authors employed nearest neighbor distances in the space of local image features and used the 'Image-to-Class' distance instead of 'Image-to-Image' distance. This scheme improves the image classification performance with heavy computational cost and some approximation algorithm has to be used to speed up the calculation.

Figure 4 shows some example images from classes with highest and lowest classification accuracy from the Caltech-101 dataset with 30 training images per class. Our method performs well on classes which are with little clutter (like watch and motorbikes) or represent coherent natural scenes (like sunflower) and less successful on classes with large intra-class variation (like dolphin and lobster). Besides, images of some classes are manually rotated to face one direction which also influences the classification performance (like accordion and barrel).

Table 4. Performance comparison on the Caltech-256 dataset with 15 training images per class.

Algorithm	Performance
SPM[31]	28.30
ScSPM[4]	27.73 ± 0.51
LScSPM[33]	30.00 ± 0.14
LLC[16]	34.36
LR-Sc^+SPM	**35.31± 0.70**

5.4. Caltech-256 dataset

The Caltech-256 dataset has 29,780 images of 256 classes with higher intra-class variability and object location variability with each image compared with the Caltech-101 dataset. Each class has at least 80 images and images are not manually rotated to face on direction as the Caltech-101 dataset. Figure 5 shows some example images of the Caltech-256 dataset. We randomly choose 15 images per class for training and 15 images per class for test. This process is repeated for 3 times. Table 4 shows the performance comparison of our method and several methods [4, 16, 31, 33] on the Caltech-256 dataset. Our method also achieves the state-of-the-art performance on this dataset which shows the effectiveness of combining non-negative sparse coding with low-rank and sparse matrix decomposition.

5.5. Information loss

Since sparse coding has no constraints on the signs of the coding coefficients, information loss is unavoidable when max pooling is then used. To measure this information loss, we firstly take the absolute value of sparse coding coefficients and then apply max pooling to get a BoW representation (denoted as $p_1 \in \mathbb{R}^{q \times 1}$) of each image and compare with the BoW representation (denoted as $p_2 \in \mathbb{R}^{q \times 1}$) generated by sparse coding plus max pooling. We use the discrepancy percentage to measure the information loss for each image as:

$$discrepancy\ percentage = \frac{q - sum(p_1 == p_2)}{q} \quad (11)$$

Table 5 shows the average discrepancy percentage on the four datasets. This is achieved by taking the average discrepancy percentage of about 150 randomly chosen images per dataset. The exact image number varies depending on the dataset. Since we use the first 3 layers of SPM with the codebook size 1,024, q is 1024×21=21,504 in our experiments. We can see from Table 5 that the sparse coding plus max pooling strategy losses about 30 percent (except Caltech-256 dataset) information which will hinder the final

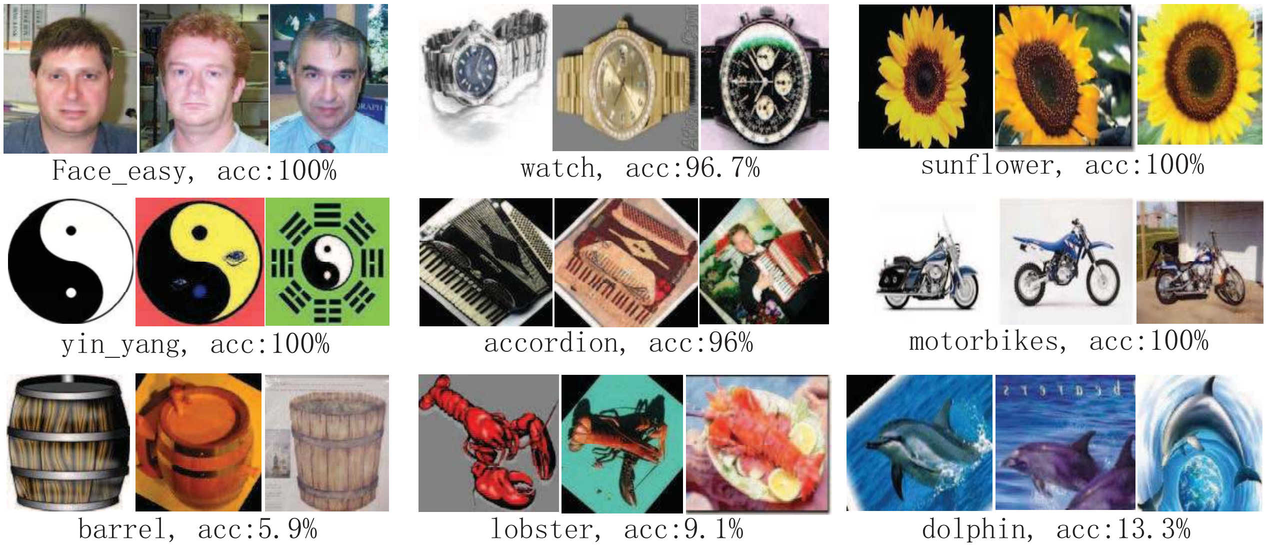

Figure 4. Example images from classes with highest and lowest classification accuracy from the Caltech-101 dataset.

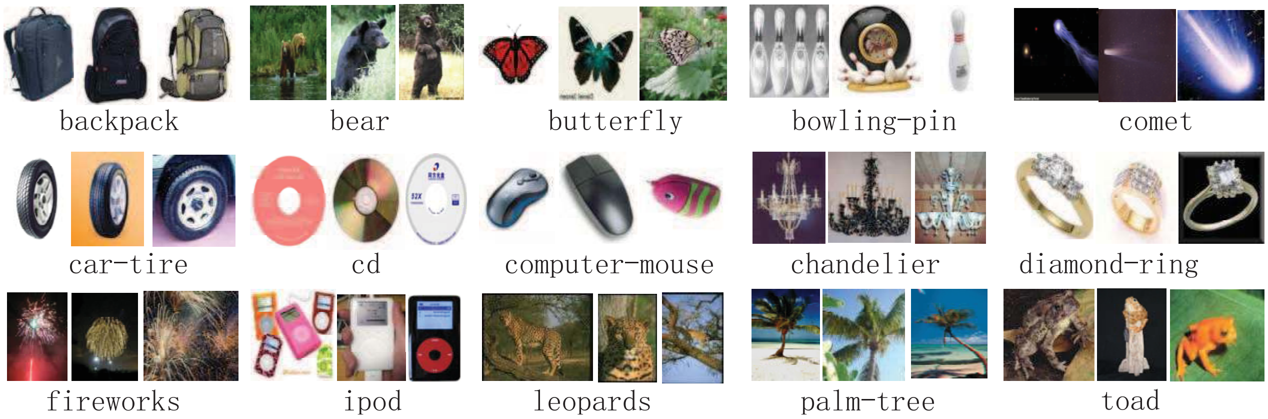

Figure 5. Example images of the Caltech-256 dataset.

Table 5. Discrepancy percentage on the Scene-15 dataset, the UIUC-Sports dataset, the Caltech-101 and 256 datsets.

Dataset	Discrepancy percentage
Scene-15	28.43%
UIUC-Sports	31.96%
Caltech-101	29.64%
Caltech-256	19.61%

classification performance. Besides, the information loss on the Caltech-256 is relatively less than that of the other three datasets. We believe this is because images of the Caltech-256 are relatively more difficult and diverse than other datasets. Local features within each image are encoded more diversely which alleviates the information loss problem.

6. Conclusion

In this paper, we introduced a novel image classification framework by leveraging the the non-negative sparse coding, low-rank and sparse matrix decomposition techniques (LR-Sc^+SPM). Specifically, to reduce the information loss, we propose to use non-negative sparse coding along with max pooling and spatial pyramid matching (Sc^+SPM) to get the BoW representation of images. Besides, we use the low-rank and sparse matrix decomposition technique to get more discriminative bases for sparse representation than directly using the training images as the bases. Experimental results on several public datasets achieve the state-of-the-art performance and demonstrate the effectiveness of the proposed method.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (Grant No. 60903146, 60723005 and 60835002) and 973 Program(Project No. 2010CB327905). This work was supported in part to Dr. Qi Tian by NSF IIS 1052851, Google Faculty Research Award and FXPAL Research Award, respectively.

References

[1] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, Nice, France, 14-17 October 2003, pages 1470-1477.

[2] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, New York, USA, 17-22 June 2006, pages 2169-2178.

[3] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, 2005.

[4] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.

[5] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Sparse representation for color image restoration. *IEEE Transactions on Image Processing*, 2008.

[6] J. Wright, A. Yang, A. Ganesh, S. Satry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, February 2009.

[7] X. Yuan, and S. Yan. Visual classification with multi-task joint sparse representation. In *CVPR*, 2010.

[8] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, and Y. Ma. Towards a practical face recognition system: Robust registration and illumination by sparse representation. In *CVPR*, 2009.

[9] D. M. Bradley and J. A. Bagnell. Differential sparse coding. In *NIPS*, 2008.

[10] K. Yu, T. Zhang, and Y. Gong. Nonlinear learning using local coordinate coding. In *In Advances in Neural Information Processing Systems* 22, 2009.

[11] J. Yang, K. Yu, T. Huang. Supervised translation-invariant sparse coding. In *CVPR*, 2010.

[12] F. Perronnin, C. Dance, G. Csurka, and M. Bressan. Adapted vocabularies for generic visual categorization. In *ECCV*, pages 464-475, 2006.

[13] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *ICCV*, pages 17-21, 2005.

[14] O. Boiman, E. Shechtman and M. Irani. In defense of nearest-neighbor based image classification. In *CVPR*, 2008.

[15] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.

[16] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010.

[17] Y-Lan Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning Mid-level features for recognition. In *CVPR*, 2010.

[18] H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In *NIPS*, 2006.

[19] Y. Liu, F. Wu, Z. Zhang, Y. Zhuang, and S. Yan. Sparse representation using nonnegative curds and Whey. In *CVPR*, 2010.

[20] Patrik O. Hoyer. Non-negative sparse coding. In *IEEE Workshop on Neural Networks for Signal Processing*. Martigny, Switzerland, 2002.

[21] S. Roweis, and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science* vol.290 no.5500, Dec.22, 2000. Pages 2323-2326.

[22] E. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, (submitted). http://watt.csl.illinois.edu/ perceive/matrix-rank/Files/RobustPCA.pdf.

[23] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willskyc, Rank-sparsity incoherence for matrix decomposition, manuscript, http://arxiv.org/abs/0906.2220.

[24] Z. Lin, M. Chen, and L. Wu. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. UIUC Technical report UILU-ENG-09-2215, November 2009.

[25] D. L. Donoho and X. Huo, Uncertainty principles and ideal atomic decomposition, *IEEE transactions on Information Theory*, 47(7), pages 2845-2862, 2001.

[26] G. Zhu, S. Yan, Y. Ma. Image tag refinement towards low-rank, content-tag prior and error sparsity. In *ACM MM*, 2010.

[27] Y. Peng, A. Ganesh, J. Wright and Y. Ma. RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images. In *CVPR*, 2010.

[28] S. Gao, I. Tsang, L. Chia, P. Zhao. Local features are not lonely-Laplacian sparse coding for image classification. In *CVPR*, 2010.

[29] L. J. Li and L. Fei-Fei. What, where and who? Classifying events by scene and object recognition. In *ICCV*, Rio de Janeiro, Brazil, October 14-20, 2007.

[30] Fei-Fei. Li, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In *CVPR Workshop on Generative-Model Based Vision*, 2004.

[31] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical report, CalTech, 2007.

[32] J. Wu and J. M. Rehg. Beyond the Euclidean distance: Creating effective visual codebooks using the histogram intersection kernel. In *ICCV*, Kyoto, Japan, 2009.

[33] J. C. Gemert, C.J. Veenman, A. Smeulders, and J. Geusebroek. Visual word ambiguity. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009.

[34] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *CVPR*, 2008.

[35] H. zhang, A. Berg, M. Maire, and J. Malik. Svm-knn: Discriminative nearest neighbor classification for visual category recongnition. In *CVPR*, 2006.

[36] X. Lian, Z. Li, B. Lu, and L. Zhang. Max-margin dictionary learning for multiclass image categorization. In *ECCV*, 2010.