Pattern Recognition Letters 34 (2013) 1046–1052

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

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

Image classification using spatial pyramid robust sparse coding

Chunjie Zhang[a], Shuhui Wang[b], Qingming Huang[a], Jing Liu[c], Chao Liang[d,*], Qi Tian[e]

[a] School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
[b] Key Lab of Intell. Info. Process, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[c] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, P.O. Box 2728, Beijing, China
[d] National Engineering Research Center for Multimedia Software, Wuhan University, Wuhan 430072, China
[e] Department of Computer Sciences, University of Texas at San Antonio, TX 78249, USA

ARTICLE INFO

Article history:
Received 13 August 2012
Available online 6 March 2013

Communicated by F. Tortorella

Keywords:
Image classification
Robust sparse coding
Spatial pyramid coding

ABSTRACT

Recently, the sparse coding based codebook learning and local feature encoding have been widely used for image classification. The sparse coding model actually assumes the reconstruction error follows Gaussian or Laplacian distribution, which may not be accurate enough. Besides, the ignorance of spatial information during local feature encoding process also hinders the final image classification performance. To address these obstacles, we propose a new image classification method by spatial pyramid robust sparse coding (SP-RSC). The robust sparse coding tries to find the maximum likelihood estimation solution by alternatively optimizing over the codebook and local feature coding parameters, hence is more robust to outliers than traditional sparse coding based methods. Additionally, we adopt the robust sparse coding technique to encode visual features with the spatial constraint. Local features from the same spatial sub-region of images are collected to generate the visual codebook and encode local features. In this way, we are able to generate more discriminative codebooks and encoding parameters which eventually help to improve the image classification performance. Experiments on the Scene 15 dataset and the Caltech 256 dataset demonstrate the effectiveness of the proposed spatial pyramid robust sparse coding method.

1. Introduction

In recent years, the bag-of-visual-word (BoW) model has become popular in image classification. This model extracts appearance descriptors from local patches and quantizes them into discrete "visual words", and then a compact histogram representation is used to represent images. The descriptive power of the BoW model is severely limited because it discards the spatial information of local descriptors. To overcome this problem, one popular extension method, called the *spatial pyramid matching* (SPM) by Lazebnik et al. (2006), is proposed and has been shown to be effective for image classification. The SPM partitions an image into several segments in different scales, then computes the BoW histogram within each segment and concatenates all the histograms to form a high dimension vector representation of the image.

To obtain good performance, researchers have empirically found that the SPM should be used together with SVM classifier using nonlinear Mercer kernels. However, the computational complexity is $O(n^3)$ and the memory complexity is $O(n^2)$ in the training phase, where n is the size of training dataset. This constrains the scalability of the SPM-based nonlinear SVM method. To reduce the training complexity and improve image classification performance, sparse coding based linear spatial pyramid matching methods (Yang et al., 2009; Serre et al., 2005; Wang et al., 2010) are proposed which help to improve classification performance. In fact, there is another constraint which was neglected in Yang et al. (2009) and Wang et al. (2010), i.e., the spatial locality constraint. For example, 'sky' often lies on the upper side of images, while 'beach' often lies on the lower side of images. When we try to encode an image region about the upper 'sky', it is more semantically meaningful to use the bases which are generated by the local features on the upper side of images. Similarly, it is more meaningful to encode the lower 'beach' with the bases generated from the local features on the lower side of images. We believe this spatial information should be combined with the codebook generation in order to encode local features more efficiently.

Besides, the sparse coding used in Yang et al. (2009) for local feature encoding tried to minimize the reconstruction error of local features by learning the optimal codebook and coding parameters simultaneously with sparsity constraints. After the codebook is

* Corresponding author. Tel./fax: +86 2768777370.
E-mail addresses: cjzhang@jdl.ac.cn (C. Zhang), shwang@jdl.ac.cn (S. Wang), qmhuang@jdl.ac.cn (Q. Huang), jliu@nlpr.ia.ac.cn (J. Liu), liangchao827@gmail.com (C. Liang), qitian@cs.utsa.edu (Q. Tian).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.02.013

learned, the rest local features are encoded by minimizing the reconstruction error with the learnt codebook and sparsity onstraints. To ensure coding parameter's smoothness and reduce encoding information loss, Laplacian sparse coding and non-negative sparse coding are proposed by Gao et al. (2010) and Zhang et al. (2011) respectively. Actually, these sparse coding models assume that the reconstruction error should be Gaussian or Laplacian distribution, which is unable to model real world applications. It would be more effective if we can construct a more robust model than simply assuming the Gaussian or Laplacian distribution of reconstruction error.

In this paper, we present a novel image classification method by using spatial pyramid robust sparse coding (SP-RSC). We give the flowchart in Fig. 1. We first partition images into sub-regions on multiple scales. Then we adopt the robust sparse coding approach to generate the codebook and encode local features of images with the spatial constraint. Different from SPM (Lazebnik et al., 2006), the proposed SP-RSC based visual vocabulary is concatenated with each encoding results from the sub-regions which have the same spatial locality and segmentation scale. For the robust sparse coding, we adopt the maximum likelihood estimation (MLE) approach and try to minimize some function of the coding residuals. This function is associated with the distribution of the coding residuals which robustly encodes the given local feature with sparse regression coefficients. Experimental evaluations on two public datasets demonstrate the effectiveness of the proposed method.

Compared with our previous work (Zhang et al., 2010), we extended the spatial pyramid coding by using robust sparse coding instead of sparse coding both for codebook construction and local feature encoding. The sparse coding assumes the reconstruction error follows the Gaussian or Laplacian distribution while the robust sparse coding has no such constraints, hence helps to encode the local features more efficiently. Besides, more experiments are added to clarify the effectiveness of the proposed spatial pyramid robust sparse coding method.

The rest of the paper is organized as follows. Section 2 gives an overview of some related work. In Section 3, we present the details of the proposed spatial pyramid robust sparse coding method. Experimental results and analysis are given in Section 4. Finally, we give the conclusions in Section 5.

2. Related work

The bag-of-visual-words model (BoW) has been widely used due to its simplicity and good performance. Many works have been done to improve the performance of the traditional bag-of-visual-words model over the past few years. Some literatures devoted to learn discriminative visual vocabulary for object recognition (Perronnin et al., 2006; Jurie and Triggs, 2005; Moosmann et al., 2008). Perronnin et al. (2006) used the Gaussian Mixture Model (GMM) to perform clustering. To alleviate the drawback of k-means clustering, Jurie and Triggs (2005) tried to use a scalable acceptance-radius based clustering method instead. Moosmann et al. (2008) used random forests to construct codebook which helps to improve the classification performance. Others tried to model the co-occurrence of visual words in a generative framework (Boiman et al., 2008; Bosch et al., 2008; Fei-Fei and Perona, 2005; Fei-Fei et al., 2004). Boiman et al. (2008) tried to classify images by nearest-neighbor classification. Bosch et al. (2008) tried to classify scene images using a hybrid generative/discriminative approach. Besides, many researchers also (Lazebnik et al., 2006; Griffin et al., 2007; Oliva and Torralba, 2001; Gemert et al., 2010; Zhang et al., 2006; Sivic and Zisserman, 2003; Grauman and Darrell, 2005) tried to learn more discriminative classifiers by combining the spatial and contextual information of visual words. Oliva and Torralba (2001) modeled the shape of the scene by using a holistic representation. Gemert et al. (2010) proposed to learn visual word ambiguity through soft assignment. Zhang et al. (2006) utilized nearest neighbor classification for visual category recognition. Motivated by Grauman and Darrell's (2005) pyramid matching in feature space, Lazebnik et al. (2006) proposed the spatial pyramid matching (SPM) which has been proven efficient for image classification.

Although the SPM method works well for image classification, it has to be used along with nonlinear Mercer kernels for good performance. However, the computational cost is high ($O(n^3)$) in the training phase. To improve the scalability, Yang et al. (2009) proposed a linear spatial pyramid matching method using sparse coding along with max pooling to classify images and has been shown very effective and efficient. The biological advantage of sparse coding along with max pooling is also proven by Serre et al. (2005). This approach relaxes the restrictive cardinality constraint of vector quantization in traditional BoW model and uses max spatial

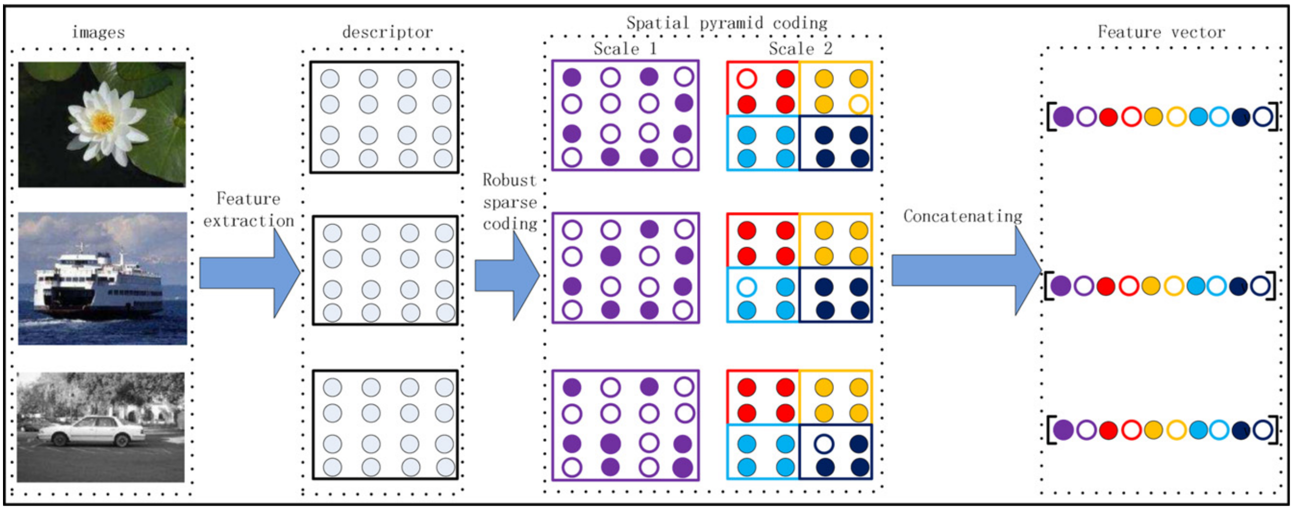

Fig. 1. Flowchart of the proposed spatial pyramid robust sparse coding (with two scales) method. For each image, we densely extract local features. We then do robust sparse coding with spatial pyramid partition of images. The filled and unfilled circles indicate different visual words that local features are encoded while the different color rectangles indicate the spatial pyramid partitions of images. Finally, the histogram representation of each partition is concatenated to form the final image representation. It is best viewed in color.

pooling to compute histogram which reduces the training complexity to $O(n)$. Motivated by this, many researchers (Wang et al., 2010; Yu et al., 2009; Boureau et al., 2010) proposed novel methods to further improve the performance. Wang et al. (2010) proposed to use locality constraints in feature space during the sparse coding phase of Yang et al. (2009) and the theoretical justifications were given by Yu et al. (2009). Boureau et al. (2010) also proposed a novel method to learn a supervised discriminative dictionary for sparse coding.

Sparse representation has been successfully used for visual applications (Yang et al., 2009; Wang et al., 2010; Gao et al., 2010; Zhang et al., 2011; Wright et al., 2009; Huang et al., 2008; Yang and Zhang, 2010; Yang et al., 2011; Kim et al., 2011), such as face classification and image classification. Wright et al. (2009) proposed the sparse representation based classification scheme and applied to face recognition with good performance. To deal with the misalignment and pose variation in face recognition, Huang et al. (2008) proposed a sparse recovery algorithm which is invariant to plane transformation. Yang and Zhang (2010) used the gabor features to reduce the size of dictionary and improved the face recognition accuracy. Yang et al. (2011) proposed to use robust sparse coding to classify faces. A hierarchical image classification method with sparse approximation is also proposed by Kim et al. (2011) which made use of the hierarchical structure information of images. Sparse coding is also used for codebook generation and local feature encoding. Yang et al. (2009) proposed to use sparse coding along with max pooling for image classification and achieved good performance over traditional k-means clustering based method (Sivic and Zisserman, 2003). Wang et al. (2010) used locality constraints during the sparse coding process to speed up computation and coding efficiency. To maintain similarity, Gao et al. (2010) added a Laplacian term to the sparse coding and improved the performance. In order to reduce the information loss caused by sparse coding with max pooling, Zhang et al. (2011) proposed to adopt the non-negative sparse coding instead. A systematic evaluation of recent feature encoding methods is conducted by Chatfield et al. (2011).

3. Spatial pyramid robust sparse coding for image classification

In this section, we give the details of the proposed spatial pyramid robust sparse coding method for image classification. For each image, we first densely extract local image features and then utilize the spatial pyramid principle to encode local features with robust sparse coding. Then we concatenate the BoW representation of different segments as the final image representation. Fig. 1 shows the flowchart of the proposed spatial pyramid robust sparse coding for image classification method.

Recently, Yang et al. (2009) proposed to use sparse coding for image classification and achieved good performance. When the codebook is fixed, learning the coding parameters can be formulated as minimizing the reconstruction error with sparsity constraints. When the reconstruction error follows the Gaussian distribution, the least square solution is the maximum likelihood estimation (MLE) solution. However, assuming the distribution of reconstruction error follows Gaussian may not be discriminative enough for image classification.

To overcome this problem, Yang et al. (2011) tries to find an MLE solution of the coding parameters and codebook. Formally, let $y = [y_1; y_2; \ldots; y_n] \in \mathbb{R}^{n\times 1}$ be the features to be encoded. The codebook is $D = [r_1; r_2; \ldots; r_n]$ where r_i is the i-th row of D. α is the coding parameters of y over D. The reconstruction error is $e = y_i - r_i\alpha$, $i = 1, 2, \ldots, n$. Yang et al. (2011) assumes $e_1, e_2, \ldots, e_n$ are independently and identically distributed. Using Taylor expansion along with necessary simplification, when D is fixed, the optimization problem can be written as:

$$min_{\alpha,D}\|W^{1/2}(y - \alpha D)\|_2^2 \quad s.t.\|\alpha\|_1 \leqslant \sigma \tag{1}$$

Where σ is the sparsity parameter. Yang et al. (2011) chooses to use the logistic function as the weight function

$$w_\theta(e_i) = exp(\mu\delta - \mu e_i^2)/(1 + exp(\mu\delta - \mu e_i^2)) \tag{2}$$

Where μ and δ are the scalars. This robust sparse coding problem can be solved by the iteratively reweighted sparse coding algorithm proposed in Yang et al. (2011) efficiently.

The idea of using spatial pyramid along with the BoW representation of images has been proven very effective for image classification by many researchers. This method partitions an image into increasingly finer spatial sub-regions and computes the histogram of local features from every sub-region (Lazebnik et al., 2006). Usually, $2^l \times 2^l$ subregions, with $l = 0, 1, 2$ are used. Other partition method such as 3×1 is also used to incorporate top and bottom relationships. Take the $2^l \times 2^l$ for example, for L levels and M channels, the resulting concatenated vector for each image has a dimensionality of $M\sum_{l=0}^{L} = M\frac{1}{3}(4^{L+1} - 1)$. However, this spatial information is only used during the image representation process while the spatial information during the codebook generation and feature encoding is lost.

Besides, to preserve the discriminative power of local image features as much as possible, researchers have tried many coding methods, among which the most popular is the k-means model. However, The constraints in the k-means model are very restrictive. To alleviate the discriminative power loss during vector quantization, many works have been done (Yang et al., 2009; Wang et al., 2010; Gao et al., 2010; Zhang et al., 2011, 2010; Yang et al., 2011). Although the effectiveness of these methods have been proven, the spatial information is lost during the coding phase. We believe the spatial information should also be included in the codebook generation and local feature encoding processes.

In this paper, we propose an "orthogonal" approach. We perform spatial pyramid coding in the two-dimensional image space and use robust sparse coding method in the feature space. Specifically, we first partition the image into increasingly finer spatial sub-regions with $2^l \times 2^l$, $l = 0, 1, 2$, as did in Lazebnik et al. (2006). For each sub-region, we first learn the codebook for each sub-region using the local image features within this sub-region. After the codebooks have been learned, we can encode each local feature within one sub-region with the corresponding codebook, as is shown in Fig. 1. In this way, we are able to combine the spatial information into the codebook generation and local feature encoding processes.

Compared with traditional sparse coding methods, the robust sparse coding is more efficient because outliers will be assigned with low weights to reduce their affects on the estimation. For each sub-region, we use the robust sparse coding both for codebook $D = [r_1; r_2; \ldots; r_n]$ learning and local feature encoding $C = [\alpha_1, \alpha_2, \ldots, \alpha_P]$ (where P is the number of local features) by alternatively optimize over D and C while keeping the other fixed. When the codebook D is fixed, finding the local feature encoding parameters can be solved by optimizing over each local feature individually through solving Problem (1).

When the coding parameter C is fixed, Problem (1) reduces to

$$min_{D=[r_1;r_2;\ldots;r_n]}\sum_{i=1}^{n}\|W^{1/2}(y_i - r_iC)\|_2^2 \tag{3}$$

We can solve this problem by optimizing over each r_i, $i = 1, \ldots, n$ individually with the others fixed. This can be solved very efficiently in a similar way as solving Problem (1) without sparsity constraints.

Algorithm 1 gives the details of the proposed spatial pyramid robust sparse coding algorithm (SP-RSC).

Algorithm 1. The proposed spatial pyramid robust sparse coding algorithm.

```
Input:
  The local features $Y, \mu, \delta, \sigma$, threshold $\theta$, max iteration
  number maxiter;
Output:
  The learned codebook $D$ and coding parameters $C$;
1: for $iter = 1, 2, \ldots, maxiter$
2:   Find the optimal encoding parameters $C$ with codebook $D$
     fixed by solving Problem 1, this is achieved by encoding
     each local feature $\alpha_i$ individually while keeping all $\alpha_j, j \neq i$
     fixed;
3:   Find the optimal codebook $D$ with $C$ fixed by solving
     Problem 3, this is achieved by finding each $r_i$ individually
     while keeping all $r_j, j \neq i$ fixed;
4:   Check if $iter > maxiter$ or the decrease of objective
     function of Problem 1 falls below $\theta$.
       If unsatisfied
         go to step 1
       Else
         stop, go to step 6;
5: end for.
6: return $D, C$;
```

After encoding each local feature accordingly, we can extract information from this spatial pyramid robust sparse coding parameters for image representation. In this paper, max pooling is used to generate the BoW representation for each sub-region. This is because the max pooling has been shown very effective when combined with sparse coding or its variants by many researchers (Yang et al., 2009; Serre et al., 2005; Wang et al., 2010; Gao et al., 2010; Zhang et al., 2011). Finally, the BoW representations of all sub-regions are concatenated into a long vector to represent images with L_2 normalization and linear SVM classifier is trained for image category prediction.

4. Experiments

We evaluate the proposed spatial pyramid robust sparse coding method on the fifteen natural scene dataset provided by Lazebnik et al. (2006) and the Caltech 256 dataset by Griffin et al. (2007). We perform all processing in grayscale of images even when sometimes the color images are provided. As to the feature extraction, we follow Lazebnik et al. (2006) and densely compute SIFT descriptors on overlapping 16×16 pixels with an overlap of 8 pixels. Each local feature is normalized with the L_2 norm. The codebook size is set to 1024, as Yang et al. (2009) did for fair comparison. Multi-class classification is done via the one-versus-all rule: a SVM classifier is learned to separate each class from the rest and a test image is assigned the label of the classifier with the highest response. The average of per-class classification rates is used to quantitatively measure the performance.

4.1. Fifteen Scene dataset

The Fifteen Scene dataset composes 4485 images, which vary from natural scenes like forests and mountains to man-made environments like offices and kitchens. Thirteen were provided by Fei-Fei and Perona (2005) (eight of these were originally provided by Oliva and Torralba (2001)) and two were collected by Lazebnik et al. (2006).

We show some example images of the Fifteen Scene dataset in Fig. 2. The major picture sources in this dataset include the COREL collection, personal photographs and Google image search. Each category has 200–400 images, and the average image size is 300×250 pixels. We follow the same experiment procedure of Yang et al. (2009) and randomly choose 100 images per category as the training set and use the remaining images as the test set. This process is repeated for five times.

Table 1 gives the comparison results. We compare the proposed methods with the kernel codebook proposed by Gemert et al. (2010), Wu and Rehg (2011), the ScSPM and the re-implementation of nonlinear kernel SPM by Yang et al. (2009). Spatial pyramid coding (SPC) uses the spatial pyramid partition of images and generate codebook for each partition with sparse coding. Robust sparse coding (RSC) uses robust sparse coding instead of sparse coding Yang et al. (2009) for image classification. Liu and Shah (2007) utilized a maximization of mutual information co-clustering approach for semantic concept clusters discovery. Our re-implementation of ScSPM is not able to reproduce the results reported by Yang et al. (2009) probably due to the feature extraction and normalization process. To make our paper self-complete, we also give the re-implementation of using Laplacian assumption for local feature encoding by solving Problem 2. We use the same set of local features for consistency. This (L_1SPM) achieved 79.85% classification performance. We can see from the results that the proposed robust sparse coding (RSC) outperforms ScSPM, which shows the effectiveness of using robust sparse coding in the coding phase. The RSC is more robust to outliers than traditional methods (Yang et al., 2009; Gemert et al., 2010). Besides, the classification rate can be further improved by using spatial pyramid coding. This demonstrates the effectiveness of the proposed SP-RSC algorithm. The proposed SP-RSC algorithm does not perform as good as the state-of-the-art algorithms, such as CENTRIST (Sohn et al., 2011) which achieved 84.96% accuracy. This is because the CENTRIST descriptor is carefully designed for scene representation. However, we still achieve higher accuracy than other sparse coding based classification methods (Yang et al., 2009; Zhang et al., 2010) which show the effectiveness of the proposed SP-RSC algorithm.

The proposed method is different from methods with large codebook. For each image sub region, our method tries to partition the local feature space separately. The partition number equals the number of image sub regions. However, the large codebook based method tries to do space partition more finely with large codebook size. To analyze the influence of the codebook size, we conduct experiments with sparse coding spatial pyramid matching with different codebook sizes. These classification performances are also given in Table 1. Generally, two conclusions can be made. First, when the codebook size is relatively small, the image classification performance improves with the increase of codebook sizes. However, this performance increase may be hindered when a relatively large codebook size (e.g. 10,000) is used. A too large codebook does not necessarily lead to good results (Oliva and Torralba, 2001). The computational cost is also high to generate a codebook with large size. Second, the proposed method is still able to exceed the performance of traditional methods with large codebook size. This also demonstrates the effectiveness of spatial partition of images.

To analyze the detailed classification performance, we give the classification rate per concept in Table 2. Generally, four conclusions can be made from Table 2. First, we can have similar observation as (Lazebnik et al., 2006) did that the indoor classes (*e.g.* store) are more difficult to classify than the outdoor classes (*e.g.* MITtallbuilding). Second, RSC performs better than ScSPM. This is because the RSC method is more robust to outliers than ScSPM during the local feature encoding process; hence helps to generate more appropriate coding parameters which eventually improve the classification performance. Third, the spatial pyramid based

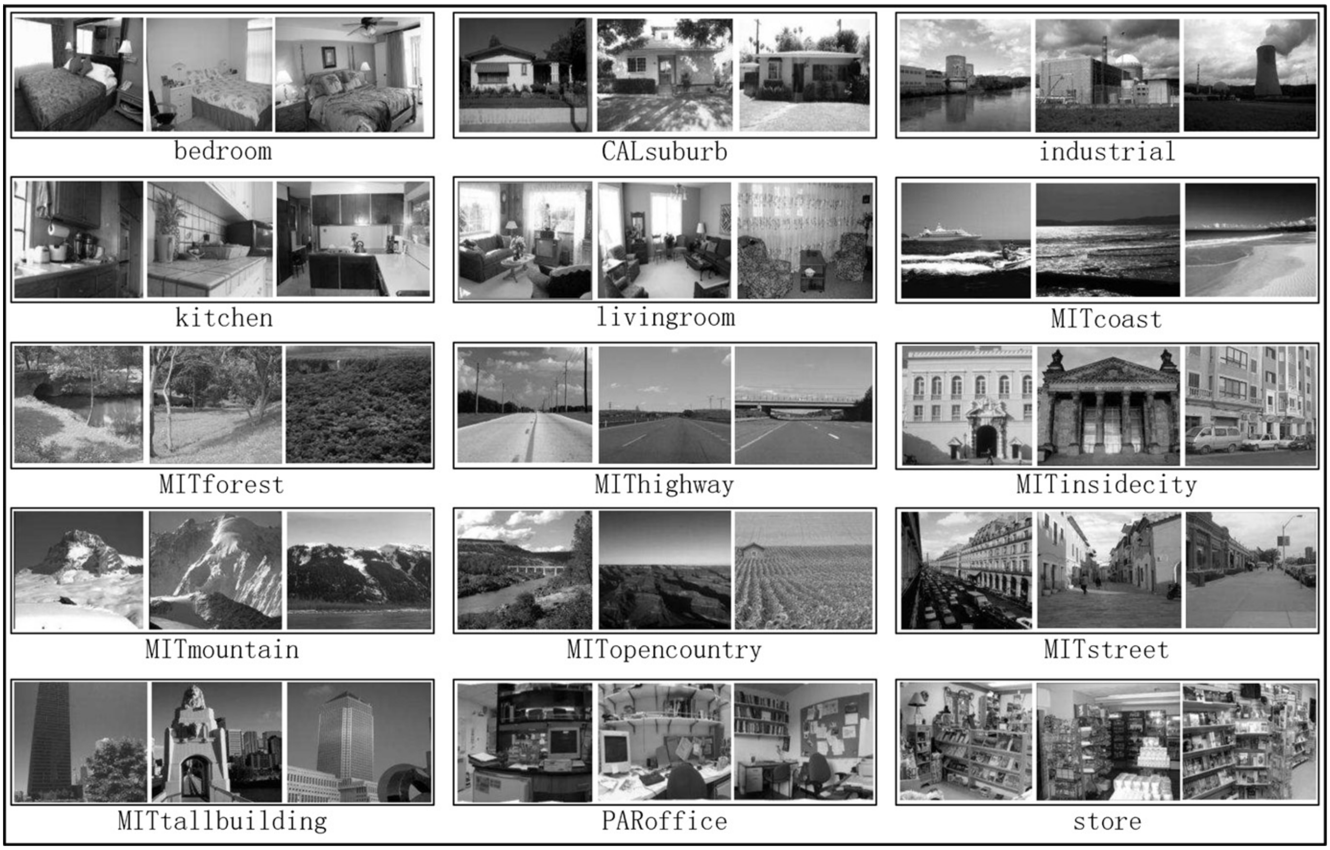

Fig. 2. Example images of the Scene 15 dataset.

Table 1
Classification rate comparison on the Fifteen Scene dataset. Numerical values in the table stand for mean and standard derivation.

Algorithms	Classification rate
KSPM (Yang et al., 2009)	76.73 ± 0.65
KC (Oliva and Torralba, 2001)	76.67 ± 0.39
ScSPM (Yang et al., 2009)	80.28 ± 0.93
Wu and Rehg (2011)	83.25
ScSPM (1024)	78.77 ± 0.52
ScSPM (2048)	80.23 ± 0.62
ScSPM (4096)	80.16 ± 0.58
ScSPM (10,000)	79.84 ± 0.60
L_1SPM	79.85 ± 0.57
SPC	81.14 ± 0.46
RSC	81.59 ± 0.44
SP-RSC	**83.67 ± 0.49**

Table 2
Classification rate per concept for the ScSPM, SPC, RSC and SP-RSC on the Fifteen Scene dataset.

Class	ScSPM	SPC	RSC	SP-RSC
Bedroom	67.24 ± 5.57	78.35 ± 1.03	68.96 ± 1.22	**84.21 ± 2.54**
CALsuburb	85.29 ± 1.42	86.79 ± 0.95	89.41 ± 0.83	**89.55 ± 1.23**
Industrial	56.40 ± 2.00	57.25 ± 2.67	56.25 ± 1.76	**57.34 ± 3.07**
Kitchen	66.36 ± 3.44	68.55 ± 2.54	66.29 ± 2.33	**69.83 ± 3.78**
Livingroom	62.43 ± 2.92	64.02 ± 2.55	64.54 ± 2.40	**65.69 ± 2.38**
Coast	90.53 ± 1.51	92.15 ± 0.61	92.52 ± 0.61	**93.03 ± 1.47**
Forest	84.85 ± 0.91	89.12 ± 1.30	88.37 ± 1.80	**97.67 ± 1.55**
Highway	86.25 ± 2.67	8.12 ± 4.34	87.62 ± 2.34	**88.85 ± 2.18**
Insidecity	88.94 ± 1.16	89.04 ± 1.43	88.94 ± 1.43	**89.50 ± 1.10**
Mountain	84.67 ± 2.70	85.50 ± 2.96	**85.95 ± 1.28**	85.67 ± 2.35
Opencountry	74.19 ± 3.33	79.03 ± 4.55	75.63 ± 2.05	**83.37 ± 0.50**
Street	84.63 ± 2.29	85.79 ± 3.13	92.09 ± 1.31	**93.91 ± 2.07**
Tallbuilding	93.57 ± 0.35	94.05 ± 0.33	97.45 ± 0.33	**98.52 ± 0.28**
PARoffice	86.96 ± 2.25	**87.83 ± 2.84**	87.23 ± 2.74	86.45 ± 1.29
Store	69.77 ± 2.70	71.53 ± 3.50	70.53 ± 1.06	**72.47 ± 1.96**

coding helps to improve the classification performance. This is because the spatial partition is able to combine the spatial information of local features into the coding process. Finally, the proposed SPC, RSC and SP-RSC methods outperform ScSPM for all the fifteen classes.

4.2. Caltech 256 dataset

The Caltech 256 dataset has 256 categories of 29,780 images with high intra-class variability and object location variability. There are at least 80 images for each class of the Caltech 256 dataset. Fig. 3 shows some example images of the Caltech 256 dataset. We randomly choose 15, 30 images per class to train SVM classifiers and use the rest for testing, as Yang et al. (2009) and Griffin et al. (2007) did.

Table 3 shows the performance comparison of the proposed SP-RSC with other sparse coding or soft assignment methods (Lazebnik et al., 2006; Yang et al., 2009; Gemert et al., 2010). Gemert et al. (2010) used kernel codebook for soft assignment of local features and considered the spatial information of local features with spatial pyramid matching. Yang et al. (2009) re-implemented the SPM algorithm for fair comparison with the sparse coding spatial pyramid matching method. We can have similar conclusions as on the Scene 15 dataset that the proposed spatial pyramid robust sparse coding exceeds the performances of SPM, KcSPM and ScSPM. This again demonstrates the effectiveness of the proposed SP-RSC. One thing needs to mention is that the proposed SP-RSC algorithm does not perform as good as LLC (Wang et al., 2010) and CRBM (Sohn et al., 2011) which achieved 41.19% and 42.05% respectively. However, we still achieve higher accuracy than other sparse coding based classification methods. Note that the proposed SP-RSC algorithm can also be combined with the LLC method to speed up computation and further improve image classification performance. We also give the performance of locality constraint spatial pyramid robust sparse coding (LC-SP-RSC) in Table 3 (with 5 nearest neighbors). We can see from Table 3 that the proposed method outperforms LLC (Wang et al., 2010) and is also comparable with CRBM (Sohn et al., 2011). This again demonstrates the effectiveness of the proposed method.

We can see from Tables 1 and 3 that the proposed method performs better on the Fifteen Scene dataset than on the Caltech 256 dataset. There are mainly three reasons. First, the Caltech 256 dataset is more challenging than the Fifteen Scene dataset with large intra and inter class variations. Second, the number of image classes is larger for the Caltech 256 dataset. Basically, the performance decreases with the increase of image classes. Third, the training numbers of the two datasets are also different. We use 100 training images per class for SVM classifier training on the Fifteen Scene dataset while 15, 30 images on the Caltech 256 dataset. To show the relative effectiveness of SP-RSC over RSC, we also give some typical examples where RSC fails but SP-RSC works on the Caltech 256 dataset in Fig. 4. We can see from Fig. 4 that the relative improvement of SP-RSC over RSC mainly occurs when the images are cluttered or have distinctive spatial partition. This consideration of spatial information is helpful to boost image classification.

In the proposed method, the spatial partition plays a very important role. These sub image regions jointly combine the spatial information which can be very effective for image classification. Besides, on analyzing the influence of different spatial regions both on the Fifteen Scene dataset and the Caltech 256 dataset, we found that the Fifteen Scene dataset plays more emphasis on the spatial regions which lie on the central areas of images. For the Caltech 256 dataset, the influences of different spatial regions are more evenly distributed. This is because the Caltech 256 dataset has large inter and intra class variations while most of the objects of the Fifteen Scene dataset lie on the central areas of images.

Table 3
Performance comparison on the Caltech 256 dataset.

Methods	15 training	30 training
KcSPM (Gemert et al., 2010)	–	27.17 ± 0.46
SPM (Lazebnik et al., 2006)	–	34.10
SPM (Yang et al., 2009)	23.34 ± 0.42	29.51 ± 0.52
ScSPM (Yang et al., 2009)	27.73 ± 0.51	34.02 ± 0.35
RSC	29.36 ± 0.35	35.27 ± 0.45
SP-RSC	**30.85 ± 0.49**	**36.73 ± 0.68**
LC-SP-RSC	**35.47 ± 0.63**	**41.87 ± 0.50**

5. Conclusions

This paper proposed a novel image classification method by spatial pyramid robust sparse coding. We first partition images into sub-regions on multiple scales. Then we use robust sparse coding to generate the codebook and encode local features per sub-region. Besides, we use the robust sparse coding technique to encode visual features with the spatial constraint. Local features from the same spatial sub-region of images are collected to generate the visual codebook for this sub-region. We generate the codebook for each sub-region and local features within a particular sub-region are encoded with the corresponding codebook to combine the spatial information. Experiments on the

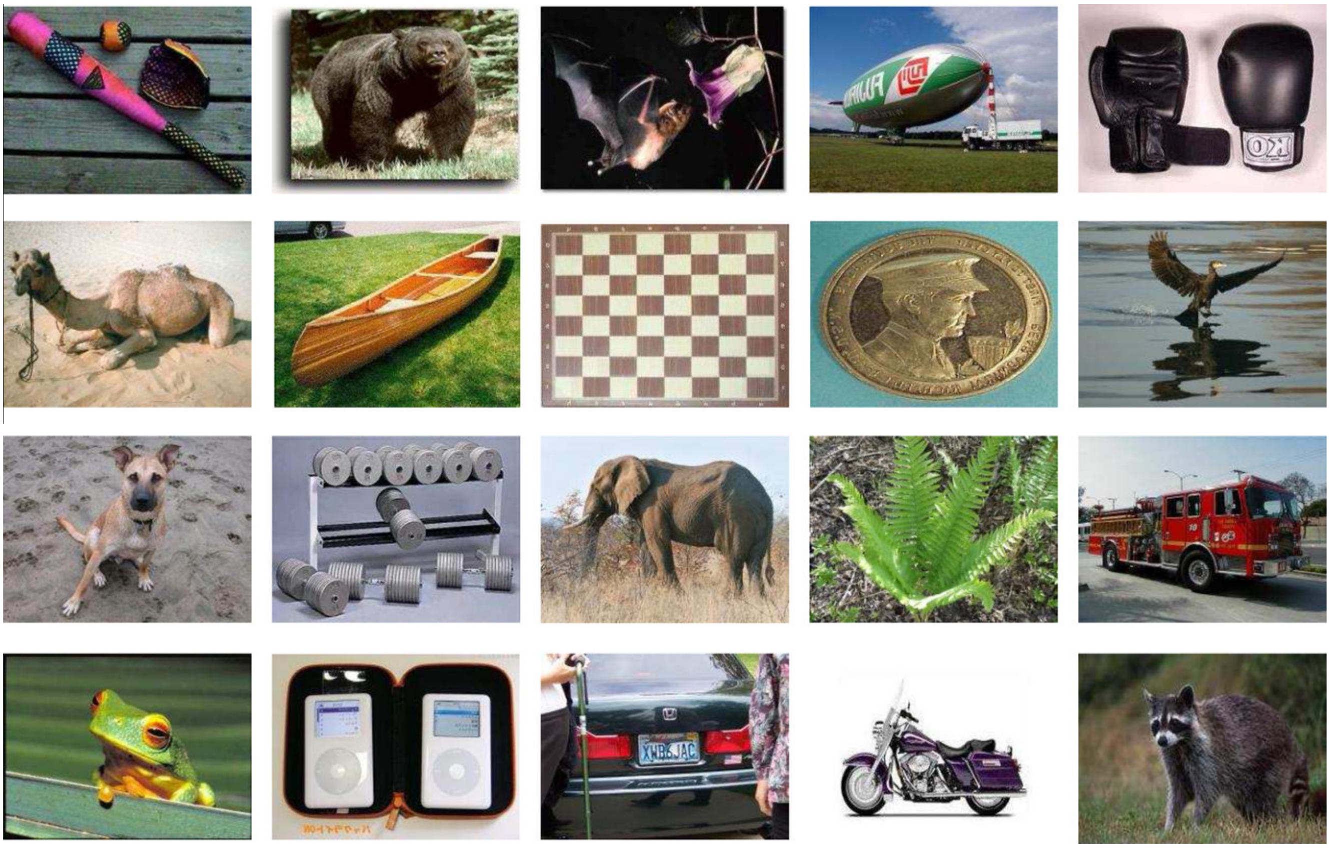

Fig. 3. Example images of the Caltech 256 dataset.

Fig. 4. Typical examples where RSC fails but SP-RSC works on the Caltech 256 dataset.

Fifteen Scene dataset and the Caltech 256 dataset show the effectiveness of the proposed method.

Our future work will concentrate on how to design more efficient algorithms that can adaptively choose spatial image partition to further improve the performance.

Acknowledgement

This work is supported by the Open Project Program of the National Laboratory of Pattern Recognition (NLPR), China Postdoctoral Science Foundation: 2012M520434, National Basic Research Program of China (973 Program): 2012CB316400, National Natural Science Foundation of China: 61025011, 61272329, 61202325.

References

Boiman, O., Shechtman, E., Irani, M., 2008. In defense of nearest-neighbor based image classification. In: Proc. CVPR.

Bosch, A., Zisserman, A., Munoz, X., 2008. Scene classification using a hybrid generative/discriminative approach. IEEE Trans. Pattern Anal Machine Intell. 30 (4), 712–727.

Boureau, Y-Lan, Bach, Francis, LeCun, Yann, Ponce, Jean, 2010. Learning mid-level features for recognition. In: Proc. CVPR.

Chatfield, K., Lempitsky, V., Vedaldi, A., Zisserman, A., 2011. The devil is in the details: an evaluation of recent feature encoding methods. In: Proc. BMVC.

Fei-Fei, L., Perona, P., 2005. A Bayesian hierarchical model for learning natural scene categories. In: Proc. CVPR.

Fei-Fei, L., Fergus, R., Perona, P., 2004. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In: Proc. WGMBV.

Gao, S.H., Tsang, I.W.H., Chia, L., Zhao, P., 2010. Local features are not lonely-Laplacian sparse coding for image classification. In: Proc. CVPR.

Gemert, J., Veenman, C., Smeulders, A., Geusebroek, J., 2010. Visual word ambiguity. IEEE Trans. Pattern Anal Machine Intell. 32 (7), 1271–1283.

Grauman, K., Darrell, T., 2005. The pyramid match kernel: discriminative classification with sets of image features. In: Proc. ICCV, Beijing, pp. 1458–1465.

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

Huang, J., Huang, X., Metaxas, D., 2008. Simultaneous image transformation and sparse representation recovery. In: Proc. CVPR.

Jurie, F., Triggs, B., 2005. Creating efficient codebooks for visual recognition. In: Proc. ICCV, pp. 17–21.

Kim, B., Park, J., Gilbert, A., Savarese, S., 2011. Hierarchical classification of images by sparse approximation. In: Proc. BMVC.

Lazebnik, S., Schmid, C., Ponce, J., 2006. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proc. CVPR.

Liu, J., Shah, M., Scene modeling using co-clustering. In: Proc. ICCV.

Moosmann, F., Nowak, E., Jurie, F., 2008. Randomized clustering forests for image classification. IEEE Trans. Pattern Anal Machine Intell. 30 (9), 1632–1646.

Oliva, A., Torralba, A., 2001. Modeling the shape of the scene: a holistic representation of the spatial envelope. Internat. J. Comput. Vision 42 (3).

Perronnin, F., Dance, C., Csurka, G., Bressan, M., 2006. Adapted vocabularies for generic visual categorization. In: Proc. ECCV, pp. 464–475.

Serre, T., Wolf, L., Poggio, T., 2005. Object recognition with features inspired by visual cortex. In: Proc. CVPR.

Sivic, J.S., Zisserman, A., 2003. Video google: a text retrieval approach to object matching in videos. In: Proc. ICCV, pp. 1470–1477.

Sohn, K., Jung, D., Lee, H., Hero, A., 2011. Efficient learning of sparse, distributed, convolutional feature representations for object recognition. In: Proc. ICCV.

Wang, Jinjun, Yang, Jianchao, Yu, Kai, Lv, Fengjun, Huang, Thomas, Gong, Yihong, 2010. Locality-constrained linear coding for image classification. In: Proc. CVPR.

Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y., 2009. Robust face recognition via sparse representation. IEEE Trans. Pattern Anal Machine Intell. 31 (2), 210–227.

Wu, J., Rehg, J., 2011. CENTRIST: a visual descriptor for scene categorization. IEEE Trans. Pattern Anal Machine Intell. 33 (8), 1489–1501.

Yang, M., Zhang, L., 2010. Gabor feature based sparse representation for face recognition with Gabor occlusion dictionary. In: Proc. CVPR.

Yang, Jianchao, Yu, Kai, Gong, Yihong, Huang, Thomas, 2009. Linear spatial pyramid matching using sparse coding for image classification. In: Proc. CVPR.

Yang, M., Zhang, L., Yang, J., Zhang, D., 2011. Robust sparse coding for face recognition. In: Proc. CVPR.

Yu, Kai, Zhang, Tong, Gong, Yihong, 2009. Nonlinear learning using local coordinate coding. In Proc. NIPS.

Zhang, H., Berg, A., Maire, M., Malik, J., 2006. Svm-knn: discriminative nearest neighbor classification for visual category recognition. In: Proc. CVPR.

Zhang, Chunjie, Liu, Jing, Wang, Jinqiao, Tian, Qi, Xu, Changsheng, Lu, Hanqing, Ma, Songde, 2010. Image classification using spatial pyramid coding and visual word reweighting. In: Proc. ACCV, pp. 239–249.

Zhang, Chunjie, Liu, Jing, Tian, Qi, Xu, Changsheng, Lu, Hanqing, Ma, Songde, 2011. Image classification by non-negative sparse coding, low-rank and sparse decomposition. In: Proc. CVPR, pp. 1673–1680.