DISCRIMINATIVE SPATIAL CODEBOOK GENERATION FOR IMAGE CLASSIFICATION

Chunjie Zhang[1], Wei Xiong[1], Jing Liu[2], Liang Fang[3], Weigang Zhang[4], Qingming Huang[1]
[1]School of Computer and Control Engineering, University of Chinese Academy of Sciences, 100049, Beijing, China
[2]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
P.O. Box 2728, Beijing, China
[3]School of Civil Engineering, Suzhou University of Science and Technology, Suzhou 215011, China
[4]School of Computer Science and Technology, Harbin Institute of Technology at Weihai, Weihai 264209, China
{cjzhang, wxiong, qmhuang}@jdl.ac.cn, jliu@nlpr.ia.ac.cn, fang@mail.usts.edu.cn, wgzhang@hit.edu.cn

ABSTRACT

Codebook plays an important role in the bag-of-visual-words (BoW) model for image classification. However, the traditional codebook generation procedure ignores the spatial information. Although a lot of works have been done to consider the spatial information for codebook generation, most of them rely on fixed region selection or partition of images, hence are not able to cope with the variations of images. To solve this problem, in this paper, we propose a novel discriminative spatial coding algorithm which can automatically generate and select the most representative codebooks for image representation. This is achieved by first generate a number of spatial codebooks through over-complete image partition with overlap. Second, for each local feature to be encoded, the most discriminative codebook is selected by jointly minimizing the encoding error and the codebook's spatial distance. Experimental results on several public image datasets show the effectiveness of the proposed discriminative spatial coding method for efficient image classification.

***Index Terms*—** Spatial codebook, spatial distance, discriminative codebook selection.

1. INTRODUCTION

The bag-of-visual-words (BoW) model is widely used by researchers for image classification. The codebook and the corresponding local feature encoding scheme play a very important role for efficient image classification performances. Traditionally, the codebook is generated by *k*-means clustering [1] or sparse coding [2] which does not consider the spatial information of images. However, the spatial information is useful for classifying images correctly. Motivated by the spatial pyramid matching (SPM) [3], a lot of works have been proposed [4-12] which can be roughly divided into local feature encoding based [4-8] and codebook generated based [9-12].

To consider the spatial information during the local feature encoding process, researchers have tried various strategies [4-8]. Bosch *et al.* [4] proposed to represent shape with a spatial pyramid kernel and achieved good performance. Wu *et al.* [5] proposed to bundle local features together using some region detection algorithm and applied it for near-duplicate image recognition with heavy computation. Wang *et al.* [6] tried to leverage the contextual information for visual applications while Zhang *et al.* [7] used Harr-like transformation of local features for image classification. Kulkarni and Li [8] first conduct affine transformation on images and then extract local features for selection with encouraging results.

Although proven effective, these methods ignore the spatial information during codebook generation process. This means the results are suboptimal because the codebook and corresponding local feature encoding scheme should be jointly considered. To alleviate this problem, researchers tried to generate spatial codebooks instead [9-12]. Zhang *et al.* [9] first partitioned images into sub-regions using the spatial pyramid scheme and generate the corresponding codebook for each image sub-region. Yang *et al.* [10] tried to generate over-complete codebooks using a mixture model while Jia *et al.* [11] went further beyond spatial pyramids and tried to encode the spatial information with receptive fields. The generation of an universal codebook and then adapt it for specific visual applications is also proposed by Perronnin *et al.* [12]. However, most of these methods still used image regions of fixed sized or locations for spatial codebook generation. Besides, the spatial constraint of nearby local features is also ignored which may reduce the

978-0-7695-5050-3/13 $26.00 © 2013 IEEE
DOI 10.1109/ICIG.2013.175

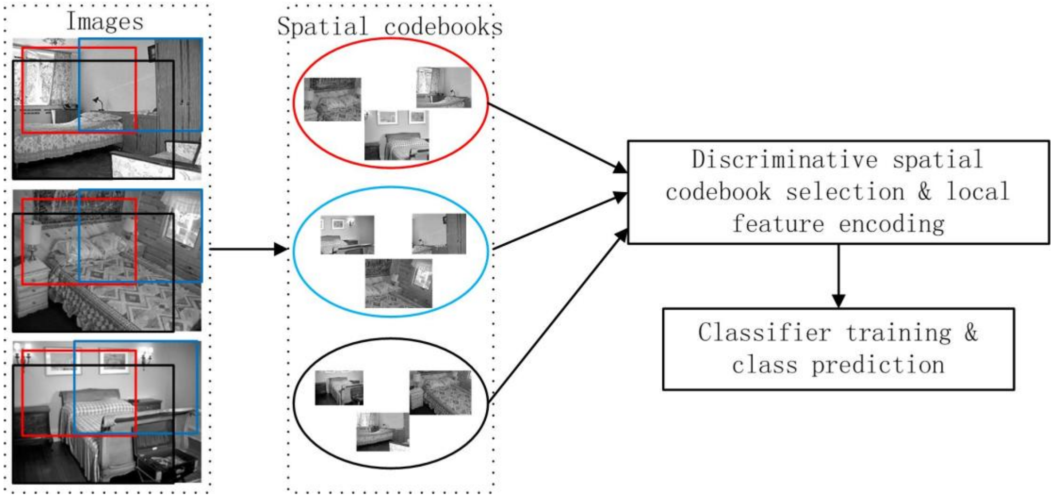
Images
Spatial codebooks
Discriminative spatial codebook selection & local feature encoding
Classifier training & class prediction

Figure 1. Flowchart of the proposed discriminative spatial codebook generation method for image classification. We first select sub-images and then generate the corresponding spatial codebook for local feature encoding and image representation. Finally, classifiers are learned for image class prediction. The rectangular boxes with different colors indicate the selected sub-images for spatial codebook generation. It is best viewed in color.

effectiveness of local feature encoding which eventually hinders the image classification performance.

To solve these above mentioned problems, in this paper, we propose a novel discriminative spatial coding algorithm for image classification. First, we generate a number of over-complete codebooks using multi-scale image regions with overlap. The local features within each corresponding regions of images are used to generate the corresponding codebook. To encode a local feature, the codebook with minimum encoding error and spatial distance is selected for local feature encoding. We then use these encoded parameters for image representation through the max pooling scheme. Finally, SVM classifiers are trained to predict the class of images. Experimental results on several public image datasets show the effectiveness of the proposed discriminative spatial codebook generation method. Figure 1 shows the flowchart of the proposed method.

The rest of this paper is organized as follows. Section 2 gives the details of the proposed discriminative spatial codebook generation method and applies it for image classification. The experimental results and analysis are given in Section 3. Finally, we conclude in Section 4.

2. DISCRIMINATIVE SPATIAL CODEBOOK GENERATION

In this section, we give the details of the proposed discriminative spatial codebook generation method for image classification. We first generate a number of spatial codebooks through over-complete image partition with overlap and then choose the most discriminative codebook by minimizing the encoding error and the spatial distances between local features and codebooks. Finally, we learn one-vs-all SVM classifiers to predict the categories of images.

2.1. Spatial codebooks construction

The consideration of spatial information during codebook generation has been proven very effective by researchers [9-12]. Recently, Zhang et al. [9] proposed to partition images using spatial pyramid and generate the corresponding codebook for each sub-image region which achieved good image classification performance. However, the sub-image regions in [9] are fixed and may not be able to cope with the large image variations. For example, the “bedroom” images have bed with varied poses, positions and scales. If we use the fixed and hard partition as [9] (1×1, 2×2 and 4×4), we are probably not able to model the concept very well.

To alleviate this problem, we propose to use adaptive sub-image regions instead. We densely extract sub-image regions using multi-scales with overlap for each image. The regions are extracted so that the whole image is covered. The sizes of images are normalized to avoid scale problem. The local features within the corresponding regions of images are used to generate the corresponding codebook for this region, as shown in Figure 1. In this way, we can take the spatial information more efficiently into the codebook generation. Besides, this adaptive sub-region selection is more efficient than the spatial pyramid partition in [9]. In fact, the method proposed in [9] is a special case of the proposed method in this paper.

2.2. Discriminative codebook selection by minimizing reconstruction error and spatial distance

Algorithm 1. The proposed discriminative spatial codebook generation algorithm.

```
Input: local feature $x$, the learned codebooks $B_n$, the
location of local feature $p_x$ and codebook $p_{B_n}$, the
maximum number of iteration $MaxIter$.
for each local feature
    for iter=1: $MaxIter$
    a. Search for the optimal encoding parameter $\alpha$
       with $B_{opt}$ fixed.
    b. Search for the optimal $B_{opt}$ with $\alpha$ fixed.
    end
end
Output: the learnt encoding parameters for each local
feature.
```

After generating the spatial codebooks, we are able to encode local features accordingly. Since we have generated a number of codebooks with their corresponding regions, we need to make some selection when encoding local features. Besides, different codebooks have different representation abilities. If we can select the most proper codebook to encode the corresponding local feature, we are able to represent images better than traditional methods.

To encode a particular local feature, in this paper, we propose to choose the most discriminative codebook by choosing the minimum reconstruction error related codebook. Besides, if two local features are spatially near, they should be encoded with codebooks generated by near sub-image regions.

Formally, let $x \in \mathbb{R}^{D\times 1}$ be the local feature to be encoded with D is the dimension of local feature, p_x is the spatial location of local feature x. $B_n \in \mathbb{R}^{D\times Q}$ (Q is the size of codebook) is the n -th learned codebook, with $n = 1, 2, ..., N$, N is the number of spatial codebooks. The spatial position p_{B_n} of codebook B_n is defined as the geometric center of the corresponding sub-image region. In this way, suppose we have a set of M local features, we can choose the most discriminative codebook and the corresponding parameters α to encode local feature x by solving the following optimization problem as:

$$[\alpha_i, B_{opt}] = \arg\min_{(\alpha_i, B_n)} \sum\nolimits_i^M \left\| x_i - B_n \alpha_i \right\|^2 + \lambda \left\| \alpha_i \right\|_1 + \gamma \sum\nolimits_{j,k}^N \theta(p_i, p_j)(p_{B_j} - p_{B_j})^2 \qquad (1)$$

Table 1. Performance comparison on the Scene 15 dataset.

Methods	Performance
SPM [3]	81.40 ± 0.50
SPC [9]	81.14 ± 0.46
ScSPM [15]	80.28 ± 0.93
KC [16]	76.67 ± 0.39
DSC (proposed)	**82.64 ± 0.65**

$$\theta(p_i, p_j) = \begin{cases} 1, \text{ if } p_i \text{ and } p_j \text{ are nearest neighbors} \\ 0, \text{ otherwise} \end{cases}$$

Where λ and γ are the corresponding sparsity and neighbor constraints. Sparsity constraint is also used in this paper as it has been proven more effective than traditional *k*-means based methods for image classification. Problem (1) is hard to solve, hence we take an alternative way and try to encode each local feature separately. When encoding local feature x_i , we iteratively optimize over the encoding parameter α_i and the corresponding codebook B_{opt} while keeping the other fixed. This can be solved efficient by the popular feature-sign-search and the Lagrange dual algorithm proposed in [13]. We give the proposed discriminative spatial codebook generation algorithm in Algorithm 1.

3. EXPERIMENTS

To evaluate the performance of the proposed method, we conduct image classification experiments on the Scene-15 dataset [3] and the Caltech 256 dataset [14], as [9] did. All images are processed in gray scales. We densely extract local features of 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 1,024. To generate the spatial codebook, we densely extract sub-image regions with 64×64 pixels of multi-scales from 1 to 5 with a step of 0.5. Max pooling [15] is used to extract the BoW representation of images. The one-versus-all rule is used for multi-class classification. We use the average of per-class classification rates as the quantitative performance measurement method. This process is repeated for five times to get reliable results.

3.1. Scene 15 Dataset

There are 4,485 images of fifteen classes (*bedroom, suburb, industrial, kitchen, livingroom, coast, forest, highway, insidecity, mountain, opencountry, street, tallbuilding, office* and *store*) in the Scene 15 dataset. Each class has 200 to 400 images with an average of 300×250 pixel size. We follow the same experimental procedure as [9] and randomly choose 100 images per class for classifier training

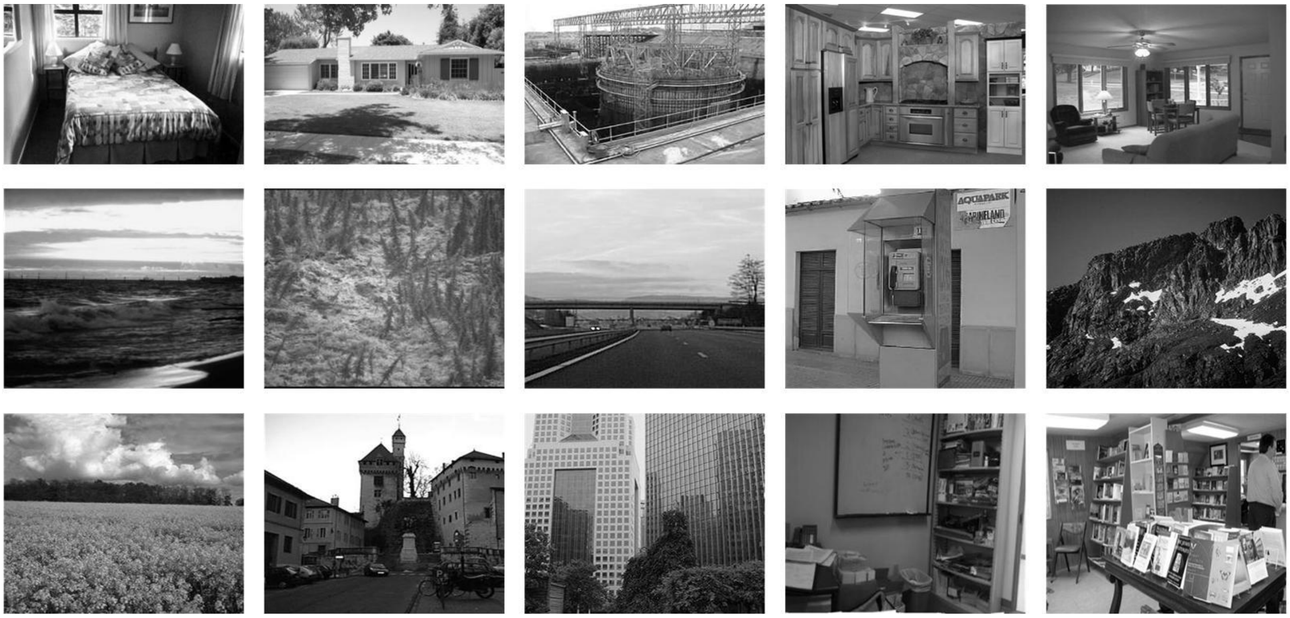

Figure 2. Example images of the Scene 15 dataset.

Table 2. Per class classification performance comparison on the Scene 15 dataset.

Class	ScSPM	SPC	DSC
bedroom	67.2 ± 5.6	78.4 ± 1.0	**81.1 ± 1.3**
suburb	85.3 ± 1.4	86.8 ± 0.9	**88.2 ± 1.1**
industrial	56.4 ± 2.0	57.3 ± 2.7	**59.9 ± 2.1**
kitchen	66.4 ± 3.4	68.6 ± 2.5	**70.4 ± 2.5**
livingroom	62.4 ± 2.9	64.0 ± 2.6	**66.8 ± 2.7**
coast	90.5 ± 1.5	92.2 ± 0.6	**92.9 ± 0.8**
forest	84.9 ± 0.9	89.1 ± 1.3	**90.3 ± 1.0**
highway	86.3 ± 2.7	88.1 ± 4.3	**89.2 ± 2.9**
insidecity	88.9 ± 1.2	89.0 ± 1.4	**90.5 ± 1.2**
mountain	84.7 ± 2.7	**85.5 ± 3.0**	85.3 ± 2.5
opencountry	74.2 ± 3.3	79.0 ± 4.6	**79.8 ± 2.6**
street	84.6 ± 2.3	85.8 ± 3.1	**87.9 ± 1.8**
tallbuilding	93.6 ± 0.4	94.1 ± 0.3	**94.9 ± 0.7**
office	87.0 ± 2.3	87.8 ± 2.8	**89.3 ± 2.2**
store	69.8 ± 2.7	71.5 ± 3.5	**73.2 ± 1.9**

and use the rest of images for performance evaluation. Figure 2 gives some example images of the Scene 15 dataset.

Table 1 gives the performance comparison of the proposed discriminative spatial codebook generation method (DSC) with [3, 9, 15, 16]. We can see from Table 1 that the proposed method achieves good image classification performance. Compared with one codebook based methods [3, 15, 16], the DSC can take advantage of the spatial information into the codebook generation, hence helps to improve the classification accuracy. Besides, compare with fixed sub-region based codebook generation method [9], the adaptive selection of sub-region can cope with image variation more efficiently. On analysis the details of the performance, we find that the improvement of DSC over SPC mainly lies on the indoor classes. This is because the indoor classes have large class variation which cannot be modeled by fixed sub-region selection method [9] well. We also give the per class classification comparison in Table 2.

3.2. Caltech 256 dataset

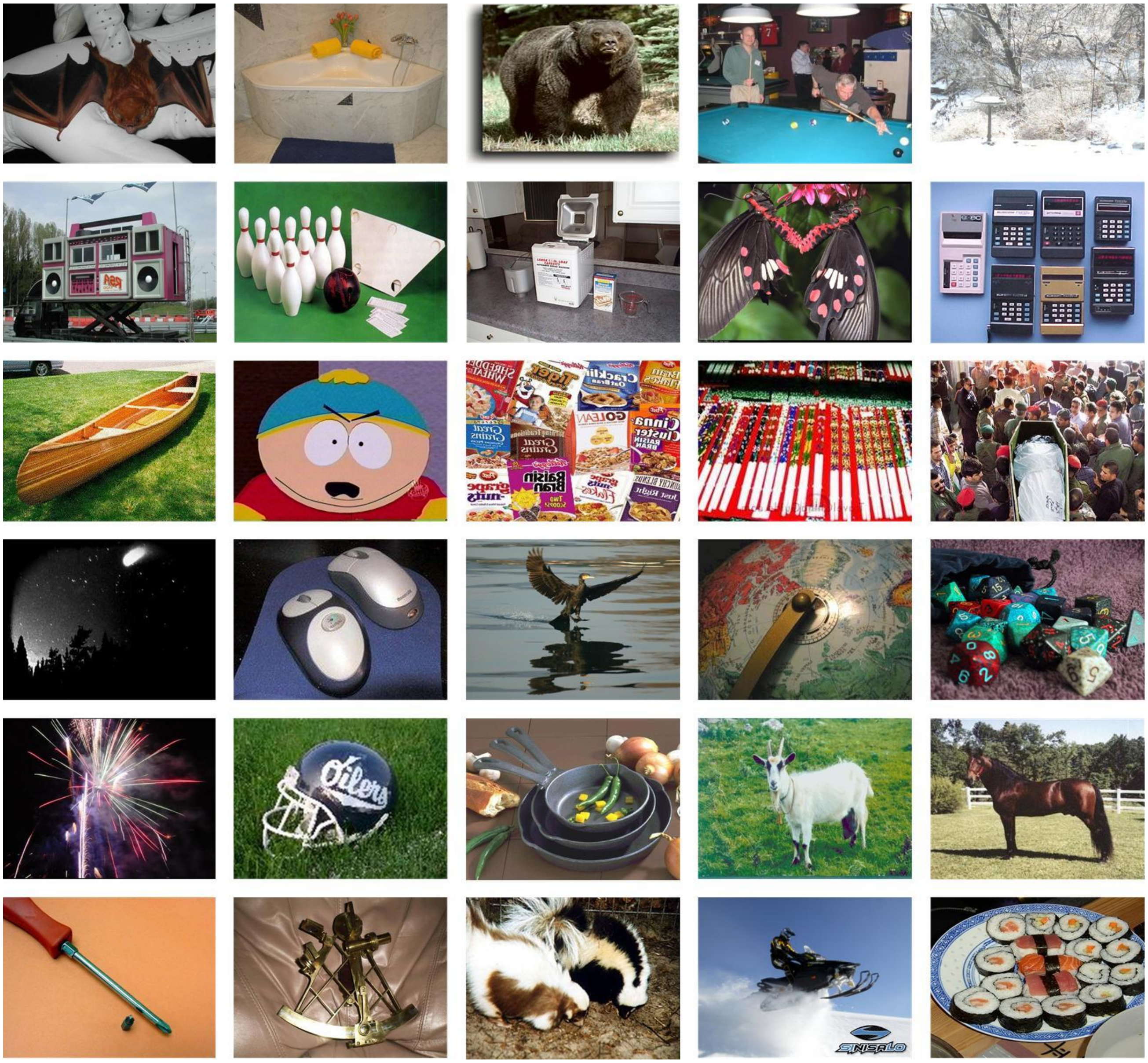

Figure 3. Example images of the Caltech 256 dataset.

The second dataset we consider is the Caltech 256 dataset [14]. This image dataset has 256 classes with 29,780 images. There are at least 80 images per class with high intra and inter class variations. We follow [9, 14, 15] and randomly choose 15, 30 training images per class for model training and the rest of images for testing. Figure 2 gives some example images of the Scene 15 dataset. We can see from Figure 3 that this dataset is more difficult than the Scene 15 dataset.

Table 3 gives the performance comparison of the proposed DSC with [9, 14-17]. We can have similar conclusions as on the Scene 15 dataset that the proposed DSC outperforms the baseline method which demonstrates the effectiveness of the proposed method. Since the Caltech 256 dataset has relatively larger variations, the performance improvement on the Caltech 256 dataset is smaller than on the Scene 15 dataset. However, the proposed DSC can still correctly classify more images than SPC which used fixed spatial codebooks. This proves the usefulness of using spatial adaptive codebooks for image classification.

4. CONCLUSION

Table 3. Performance comparison on the Caltech 256 dataset.

Methods	15 training	30 training
ScSPM [15]	27.73±0.51	34.02±0.35
SPM [15]	23.34±0.42	29.51±0.52
SPC [9]	30.85±0.49	36.73±0.68
SPM [14]	-	34.10
KC [16]	-	27.17±0.46
Classemes [17]	-	36.00
DSC	**31.95±0.50**	**37.54±0.59**

This paper proposes a novel discriminative spatial codebook generation method for image classification. To take the advantage of spatial information during the codebook generation process as well as cope with the image variations, we first densely select sub-image regions with overlapping and then use the local features within each sub-region for the corresponding codebook construction. A discriminative codebook selection method is then proposed by jointly minimizing the sparse encoding error and the spatial distance. Experimental results on several public image datasets show the effectiveness of the proposed method.

Our future work consists of two aspects. First, we will study how to combine the spatial as well as the structure information more efficiently [18]. Second, how to speed up the computation will also be investigated.

5. ACKNOWLEDGEMENT

This work is supported by the President Fund of UCAS; 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, 61202322.

6. REFERENCES

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

[2] C. Zhang, J. Liu, Q. Tian, C. Xu, H. Lu and S. Ma. Image classification by non-negative sparse coding, low-rank and sparse decomposition. In *Proceedings of Computer Vision and Pattern Recognition*, 2011, pages 1673-1680.

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

[4] A. Bosch, A. Zisserman, and X. Munoz, Representing shape with a spatial pyramid kernel, In *Proceedings of the International Conference on Image and Video Retrieval*, 2007, pages 401-408.

[5] Z. Wu, Q. Ke, and J. Sun. Bundling features for large scale partial-duplicate web image search. In *Proceedings of Computer Vision and Pattern Recognition*, 2009, pages 25-32.

[6] X. Wang, X. Bai, W. Liu and L. Latecki, Feature context for image classification and object detection. In *Proceedings of Computer Vision and Pattern Recognition*, June 2011, pages 961-968.

[7] C. Zhang, J. Liu, C. Liang, Q. Huang and Q. Tian, Image classification using Harr-like transformation of local features with coding residuals, Signal Processing, DOI: 10.1016/j.sigpro.2012. 09.007.

[8] N. Kulkarni, and B. Li, Discriminative affine sparse codes for image classification, In *Proceedings of Computer Vision and Pattern Recognition*, 20-25 June 2011, pages 1609-1616.

[9] C. Zhang, S. Wang, Q. Huang, J. Liu, C. Liang and Q. Tian, Image classification using spatial pyramid robust sparse coding, *Pattern Recognition Letters*, DOI:10.1016/j.patrec.2013.02.013.

[10] J. Yang, K. Yu and T. Huang, Efficient highly over-complete sparse coding using a mixture model, In *Proceedings of European Conference on Computer Vision*, 2010.

[11] Y. Jia, C. Huang and T. Darrell, Beyond spatial pyramids: Receptive field learning for pooled image features, In *Proceedings of Computer Vision and Pattern Recognition*, 2012, pages 3370-3377.

[12] F. Perronnin, C. Dance, G. Csurka, and M. Bressan. Adapted vocabularies for generic visual categorization. In *Proceedings of European Conference on Computer Vision*, pages 464-475, 2006, pages 464-475.

[13] H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In *Proceedings of Neural Information Processing Systems*, 2006, pages 801-808.

[14] G. Griffin, A. Holub and P. Perona, Caltech-256 object category dataset, Technical Report, Caltech, 2007.

[15] J. Yang, K. Yu, Y. Gong and T. Huang, Linear spatial pyramid matching using sparse coding for image classification, In *Proceedings of Computer Vision and Pattern Recognition*, 2009.

[16] J. Gemert, C. Veenman, A. Smeulders and J. Geusebroek, Visual word ambiguity, *IEEE Trans. Pattern Anal Machine Intell*. 32(7):1271-1283, 2010.

[17] L. Torresani, M. Szummer, A. Fitzgibbon, Efficient object category recognition using classesmes. In *Proceedings of European Conference of Computer Vision*, Crete, Greece, 2010.

[18] C. Zhang, J. Liu, Q. Tian, Y. Han, H. Lu, and S. Ma. A boosting, sparsity-constrained bilinear model for object recognition, *IEEE Multimedia* 19(2):58-68 (2012).