Image Classification Using Spatial Difference Descriptor Under Spatial Pyramid Matching Framework

Yuhui Li[1,2(✉)], Jiucheng Xu[1], Yifan Zhang[2], Chunjie Zhang[3], Hongsheng Yin[4], and Hanqing Lu[2]

[1] School of Computer and Information Engineering, Henan Normal University, Xinxiang, China
liyuhui0224@126.com, xjch3701@sina.com
[2] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Science, Beijing, China
{yfzhang,luhq}@nlpr.ia.ac.cn
[3] School of Computer and Control Engineering, University of China Academy of Sciences, Beijing, China
cjzhang@jdl.ac.cn
[4] China University of Mining and Technology, Xuzhou, China
xuzhouyhs@sina.com

Abstract. Spatial pyramid matching (SPM) model is an extension of the bag-of-visual words (BoW) model for local feature encoding. It firstly partitions the image into increasingly fine sub-regions, and then concatenates the histograms within each sub-region. However, the SPM model does not consider the spatial information differences between sub-regions explicitly. To make use of this information, we exploit a novel descriptor called spatial difference. In the process of promoting the performance of image classification, this descriptor is mainly used to concatenate the histograms of bag-of-visual words model under spatial pyramid matching framework. Finally, we conduct image classification experiments on several public datasets to demonstrate the effectiveness of the proposed scheme.

Keywords: Image classification · Spatial difference descriptor · Spatial pyramid matching · Bag-of-visual words · Sparse coding

1 Introduction

In recent years, the bag-of-visual words (BoW) model [1] has been very popular in various image applications, especially for image classification. Codebook generation and histogram representation are two important components for generating bag-of-visual words representation. Bag-of-visual words model has been demonstrated that combining codebook and histogram representation together can achieve good performance after being trained to predict the classes of images.

Though bag-of-visual words model has achieved good performance, there exist obvious drawbacks: both the spatial information and correlations among visual contents are neglected. Later, a spatial pyramid matching (SPM) model [2] was proposed to deal

Q. Tian et al. (Eds.): MMM 2016, Part I, LNCS 9516, pp. 527–539, 2016.
DOI: 10.1007/978-3-319-27671-7_44

with this problem by dividing the whole image into hierarchical sub-regions and concatenating the appropriately weighted histograms of each region. Extensive experimental results have shown that spatial pyramid matching model achieved a remarkable success on a wide range of image classification benchmarks as a fundamental model. However, spatial pyramid matching model has its own weaknesses: the finer the division, the more sensitive it is to location and orientation of visual content, as described in [3]. In view of this, many works focused on improving SPM to overcome these weaknesses, such as researchers [3–6] tried to improve the coding procedure to minimize the representation information loss. Teng et al. [3] explored a weakly spatial symmetry descriptor to boost the performance of bag-of-visual words model by combining weakly spatial symmetry (WSS) and BoW together. Despite its success in the scene image classification domain, WSS model has its own limitations. For example, after dividing the image into many sub-regions and generating histograms of bag-of-visual words model, WSS only computes spatial symmetry information inside each sub-region, rather than considering the spatial difference information between sub-regions. While spatial difference information between sub-regions is much more important than inside spatial symmetry information when dividing images into increasingly fine sub-regions.

Hence we propose a novel approach to relieve the above problems under spatial pyramid matching framework. Firstly, we compute spatial difference information in four kinds of orientations. Secondly, we combine the spatial difference descriptors with histograms of bag-of-visual words model together. Finally, experiments are conducted on several datasets, which mainly include estimating different distance measurements, evaluating the performance with different codebook sizes and comparing with other methods to prove the effectiveness of our approach.

2 Related Work

The bag-of-visual words (BoW) [1] model has been widely used for visual applications. Traditional BoW model uses the *k*-means clustering algorithm and considers the cluster centers as visual words. Local features are then quantized to the nearest visual word. However, this solution leads to severe information loss, which limits its discriminative power. In order to reduce the information loss in the local feature encoding process, many works [6–8] have been proposed. For instance, as a classical and typical one, Yang et al. in [6] proposed one scheme to sparse coding with spatial pyramid matching for codebook generation, and trained linear classifier to save computational cost, which was much more effective than non-linear classifier.

Further more, the traditional BoW model lacks the spatial information. Inspired by the work done by Grauman and Darre11 [4], Lazebnik et al. [2] proposed the spatial pyramid matching (SPM) algorithm which was widely used by many researchers. Later, a lot of works [9–13] have been done to combine the spatial information of local features, which are motivated by the SPM algorithm. A hierarchical matching method with side information was proposed by Chen et al. [9] and it was used for image classification. A weighting scheme was used to select discriminative visual words. Randomization and discrimination was combined into a unified frame work by Yao et al. [10], which was used for fine grained

image categorization. Zhang et al. [11] proposed a pose pooling kernel to recognize sub-category birds. Representing images with components and a bilinear model for object recognition was used in [12] proposed by Zhang et al., Bao and He [13] proposed an improved sparse coding model based on linear spatial pyramid matching (SPM) and scale invariant feature transform (SIFT) descriptors. Teng et al. [3] explored a weakly spatial symmetry descriptor to boost the performance of bag-of-visual words model by combining WSS and BoW together. Authors in [14, 15] explored methods of combining spectral and spatial information directly to boost the final performance. Zhu et al. [16] presented a robust semi-supervised kernel-FCM algorithm incorporating local spatial information to solve the original problem of image classification. In another way, Zhang et al. [17] proposed a novel object categorization method by using the sub-semantic space based image representation. Most of the previous works have their own superiority on some datasets. However, none of them consider the spatial difference information between sub regions. To make use of this lost information, we propose a novel descriptor named spatial difference to improve the performance for image classification.

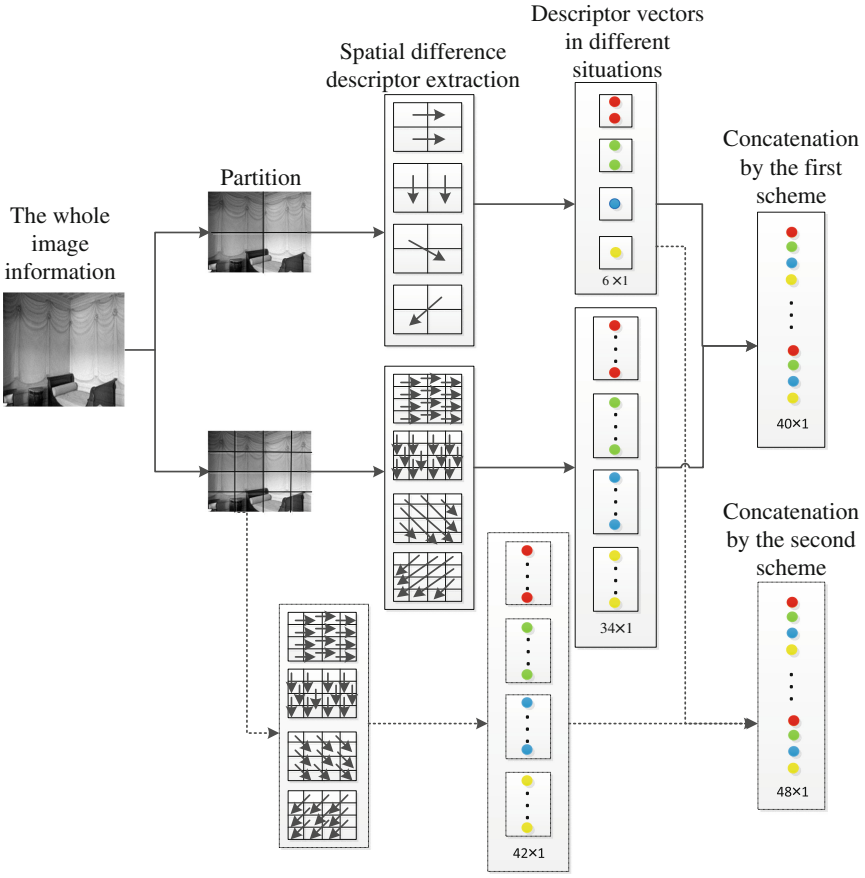

Fig. 1. Flow chart of spatial difference descriptor computation

3 Proposed Framework

Our image classification solution is derived from spatial pyramid matching model [6], which mainly includes five modules: feature extraction, sparse coding, spatial pooling, spatial difference descriptor computation, and finally linear classifier. Firstly, feature extraction accomplishes obtaining the original image representation vectors. Then sparse coding and spatial pooling are used to learn dictionary and encode the local features respectively. After this, the spatial difference descriptor is computed to complete the gist of obtaining discriminative image representation vectors by going one more step, as shown in Fig. 1. In the last module, we use linear SVM classifier to do image classification by training and testing the obtained discriminative feature the same as other common benchmarks.

3.1 Feature Extraction

Local features play a very important role for effective image representation. Choosing the proper local features is helpful to improve the final image classification performance substantially. For better discriminative power, we utilize higher-dimensional "strong features", which are SIFT descriptors of 16×16 pixel patches computed over a grid with 8-pixel spacing. Before extracting features, the image should be all processed into gray scale. At last, These extracted features are then normalized with L_2 norm.

3.2 Sparse Coding

Sparse coding has been widely used for codebook generation in the BoW model. Let $X = [x_1, x_2, \ldots, x_N]$ ($x_i \in \mathbf{R}^{D\times 1}$) be a set of N local image descriptors of each D dimension. Given a codebook with K entries to be learned, $V = [v_1, v_2, \ldots, v_K]$ ($v_i \in \mathbf{R}^{D\times 1}$), each descriptor can be converted into a K-dimensional code to generate the final image representation. Let $U = [u_1, u_2, \ldots, u_N]$ is the set of codes for X. Typically the sparse coding method solves the following optimization problem as:

$$\min_{U,V} \sum_{n=1}^{N} ||x_n - u_n V||^2 + \lambda ||u_n||_1$$
$$s.t. \; ||v_k||^2 \leq 1, \forall k$$

where λ is the regularization parameter. Considering the large amount of local features, we only sample a subset of features to learn the codebook. With the codebook in place, the local features of each image can be encoded.

Yang et al. [6] developed an extension of the SPM method [2] by generating vector quantization to sparse coding followed by multi-scale spatial max pooling, and proposed a linear SPM kernel based on SIFT sparse codes. Their approach, called ScSPM, is naturally derived by relaxing the restrictive cardinality constraint of VQ.

3.3 Spatial Pooling

In the "SPM" layer, we partition an image into $2^l \times 2^l$ spatial sub-regions, where $l = 0$, 1, 2 stands for different scales. The codes of the descriptors are pooled together to get the corresponding pooled features. These pooled features from each sub-region are concatenated and normalized to form the image feature representation. The pooling method used in this paper is max pooling:

$$z_j = \max\left(u_{ij}\right) \quad i = 1, 2, \ldots, N, \; j = K.$$

In our framework, "max pooling" combined with L_2 normalization is used. Max pooling can produce better performance than other pooling methods (i.e. *Sqrt* and *Abs*), as demonstrated by Yang et al. [6], probably due to its robustness to local spatial translation and biological plausibility.

3.4 Spatial Difference Descriptor Computation

After image representation being generated by sparse coding and spatial pooling hierarchically, spatial difference descriptors are computed according to four kinds of spatial difference information. For example, the sub-figure (a) in Fig. 2 describes left to right difference, and the sub-figure (b) describes top to down difference. As for the diagonal differences, we compare two different schemes, as shown in Fig. 1.

Fig. 2. Four kinds of spatial difference information

In the first scheme, diagonal spatial difference information is extracted by computing distances between sub-regions which may not be contiguous. In the second scheme, diagonal spatial difference information is extracted by computing distances between sub-regions which are all contiguous. Both of the two schemes compute spatial difference information between sub-regions rather than in each sub-region.

Because the two sub-regions to be calculated for diagonal spatial difference information may be not contiguous, their correlation is denoted by dotted line, as shown in sub-figure (c) and sub-figure (d) of Fig. 2. Especially, Fig. 1 describes the flow chart of the spatial difference descriptor computation process in the whole image level. In order to differentiate the two computing schemes, we denote the first scheme as spatial difference 1 (SD^1) and the second one as spatial difference 2 (SD^2). It is important to note that, in Figs. 1 and 2, the blocks in which the head and tail of the arrow line locate are the sub-regions we choose to compute difference descriptor.

For two computing sub-regions of the segmented image, two vectors h_1 and h_2 are built based on the size of codebook:

$$h_1 = [h_{11}, h_{12}, \ldots, h_{1K}]$$
$$h_2 = [h_{21}, h_{22}, \ldots, h_{2K}]$$

where K is the size of codebook.

If these two sub-regions are strictly the same, the distance between them would be near 0. Obviously, different distance measurements may lead to different results, and then we can use different methods to compute the distance. Finally, we will adopt *euclidean* distance measurement as the most proper one to calculate the spatial difference information between two vectors. Now we can obtain a feature vector to describe the spatial difference descriptor for the whole image:

$$D_{SD} = [d_1, d_2, \ldots, d_P]$$

where P is the number of sub-region pairs to compute according to Figs. 1 and 2.

To combine histograms of bag-of-visual words model and spatial difference information, we utilize the method as described in [3] by Teng et al. and obtain the final discriminative feature representation:

$$\begin{aligned} W_{ScSPM+SD} &= [H, D_{SD}] \\ &= [h_1, h_2, \ldots, d_m, d_1, d_2, \ldots, d_P] \end{aligned}$$

where H is the histograms of bag-of-visual words, and m is the size of H.

4 Experiments and Results

To evaluate the effectiveness of the proposed method in this paper, we choose to conduct image classification experiments on several public datasets. The datasets are Scene 15 dataset, Caltech 101 dataset and Caltech 256 dataset.

Our approach uses the popular SIFT descriptors. The same as other common benchmarks, we randomly select the training images and use the rest images for testing. This process is repeated for ten times to get reliable results. Mean of per-class classification rates for performance measurement is used and we report the final results by mean and standard deviation of the classification rates. SPM with three pyramid and SD^1 (or SD^2) in which spatial difference information is extracted are used to combine the hierarchical histograms and correlations between sub-regions together. Hence, each image is represented by a vector of $21 \times K$ (size of codebook) + 40 (or 48) for all datasets.

Our experiments mainly include three parts: firstly, fixing the codebook size with 1024, and for Scene 15 dataset and Caltech 101 dataset, we evaluate the performance of different distance measurements. Measuring distance includes six methods, which are *chebychev, jaccard, hamming, cosine, cityblcok* and *euclidean*. Secondly, fixing the distance measurement, we evaluate the performance by changing the codebook size. Finally, we choose to compare with other methods which are closely related with the proposed method by their reported results instead of re-implementing them and test on every datasets for fair comparison mostly.

4.1 Scene 15 Dataset

We firstly try our method on the Scene 15 dataset. There are 4485 images of 15 classes (bedroom, coast, forest, highway, industrial, insidecity, kitchen, living room, mountain, office, opencountry, store, suburb and tallbuiding) in this dataset. Each class has 200 to 400 images. We follow the same experiment setup as Yang et al. [6] did and randomly select 100 images per class for classifier training.

Then we evaluate the performance by using different distance measurements. For fair comparison, we extract WSS descriptor proposed by Teng et al. [3] and use them under the same framework with us. When, taking WSS, SD^2 and SD^1 into consideration, as shown in Fig. 3, we can see that SD^1 can achieve better performance than the other two methods, and the distance measurement of *euclidean* is the best selection to get amazing performance.

We also conduct experiments on different size of codebook to observe the effect for classification on the Scene 15 dataset. As shown in Fig. 4, when the codebook size is 2048, we can achieve the best performance on Scene 15 dataset.

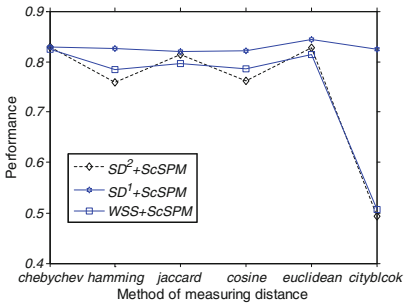

Fig. 3. Performance under different distance measurements on Scene 15

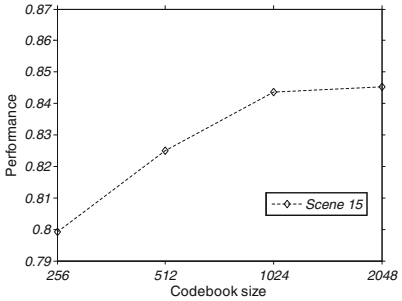

Fig. 4. Performance of codebook size on Scene 15

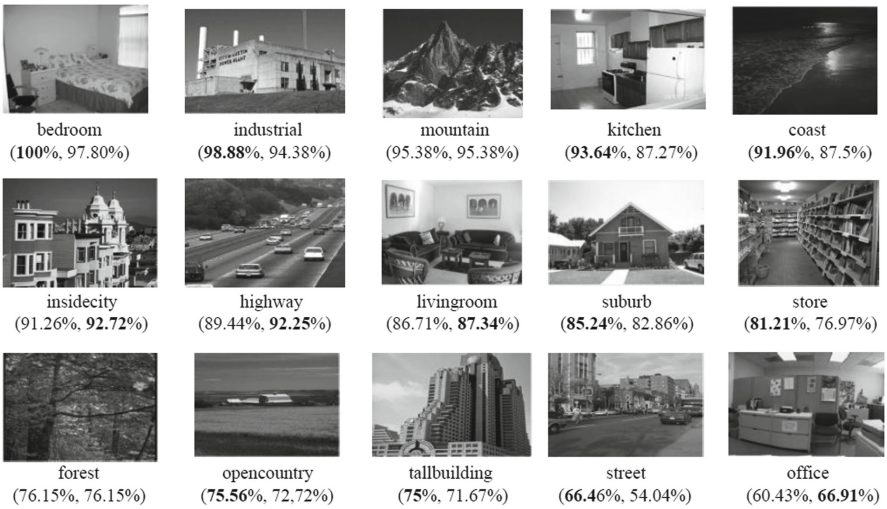

Fig. 5. Example images from classes with classification accuracy comparison on Scene 15

Table 1. Image classification results on Scene 15

Methods	Performance
KSPM [2]	81.40 ± 0.50
WSS + SPM [3]	81.51 ± 0.00
KCSPM [5]	76.70 ± 0.40
LSPM [6]	65.32 ± 1.02
ScSPM [6]	80.28 ± 0.93
LScSPM [7]	**89.75 ± 0.50**
NNScSPM [13]	81.92 ± 0.42
S^3R [17]	83.72 ± 0.78
WSS + ScSPM	81.46 ± 0.00
SD^2 + ScSPM	82.80 ± 0.00
SD^1 + ScSPM	84.52 ± 0.01

Finally, we give the performance of the proposed method and compare with methods proposed by [2, 3, 5–7, 13, 17] in Table 1. As shown in Table 1, LScSPM can achieve high performance on scene classification. The problem reason is that scene images contain plentiful textures in each patch, which results in the unstableness for sparse coding process. By adding Laplacian term, similar patches will be encoded into similar codes, thus the image can be accurately represented [7]. Except LScSPM, we can see that our method SD^1 descriptor under SPM framework achieves comparable results. Figure 5 shows some example images from Scene 15 dataset classes with classification accuracy in brackets. The first number in the bracket is the accuracy obtained by using SD^1 descriptor under spatial pyramid matching framework, and the second number in the bracket is the accuracy obtained by using the classic method proposed by Yang et al. in [6].

4.2 Caltech 101 Dataset

The Caltech 101 dataset contains 8144 images falling in 101 classes including animals, vehicles, flowers, etc., with significant variance in shape. The number of images per class varies from 31 to 800. Most images are medium resolution, i.e. about 300×300 pixel. We follow the common experiment setup as Yang et al. [6] did and randomly select 15 and 30 images per class for classifier training and use the rest images for testing.

On one hand, for fair comparison, as did on Scene 15 dataset, we conduct experiments to evaluate the performance of different distance measurements by taking WSS, SD^2 and SD^1 into consideration under spatial pyramid matching framework. On the other hand, we test the performance with different codebook sizes. We can see from Fig. 6, the method of *euclidean* outperforms the others. When the codebook size is 2048, we can achieve the best performance, as shown in Fig. 7.

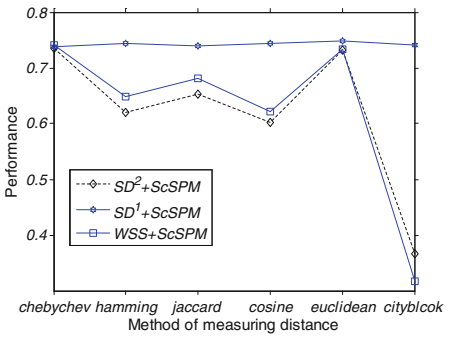

Fig. 6. Performance under different distance measurements on Caltech 101

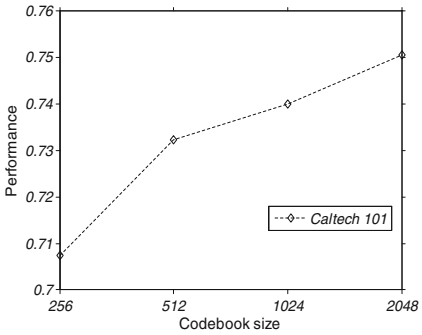

Fig. 7. Performance of codebook size on Caltech 101

At last we give the performance of the proposed method and compare with other methods described [2, 3, 5, 6] in Table 2. Figure 8 shows some typical images owning the top 18th classification accuracy in brackets. The first number in the bracket is the accuracy obtained by using SD^1 descriptors under spatial pyramid matching framework, the second one is the accuracy by using the method proposed by Yang et al. in [6]. From Table 2 and Fig. 8, we can see our method outperforms the other related methods, mainly due to the contribution of spatial difference descriptors.

Table 2. Image classification results on Caltech 101

Methods	15 training	30 training
KSPM [2]	56.40 ± 0.00	64.60 ± 0.80
WSS + SPM [3]	-	67.57 ± 0.00
KCSPM [5]	-	64.14 ± 0.18
LSPM [6]	67.00 ± 0.45	58.81 ± 1.51
ScSPM [6]	67.00 ± 0.45	73.20 ± 0.54
WSS + ScSPM	66.94 ± 0.01	73.39 ± 0.14
SD^2 + ScSPM	67.60 ± 0.00	73.15 ± 0.01
SD^1 + ScSPM	**70.01 ± 0.00**	**74.26 ± 0.01**

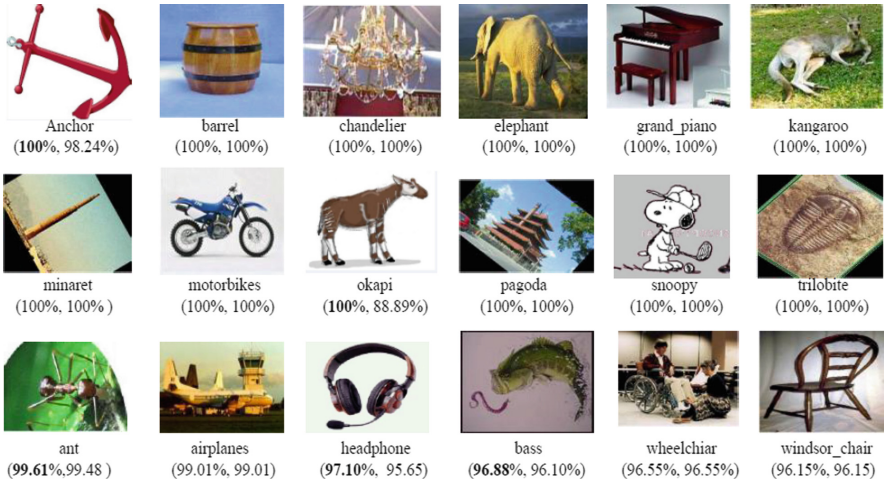

Fig. 8. Example images from classes with the top 18th classification accuracy on the Caltech 101 when using SD^1 descriptor under spatial pyramid matching framework

Table 3. Image classification results on Caltech 256

Methods	15 training	30 training	45 training	60 training
KCSPM [5]	-	27.17 ± 0.46	-	-
LSPM [6]	13.20 ± 0.62	15.45 ± 0.37	16.37 ± 0.47	16.57 ± 1.01
ScSPM [6]	27.73 ± 0.51	34.02 ± 0.35	37.46 ± 0.55	40.14 ± 0.91
LScSPM [7]	30.00 ± 0.14	35.74 ± 0.10	38.54 ± 0.36	40.43 ± 0.38
S^3R [17]	**37.85 ± 0.48**	**43.52 ± 0.44**	**46.86 ± 0.63**	-
KSPM [18]	-	34.10 ± 0.00	-	-
WSS + ScSPM	30.98 ± 0.00	36.90 ± 0.00	39.79 ± 0.00	41.63 ± 0.00
SD^2 + ScSPM	31.25 ± 0.00	36.84 ± 0.00	39.67 ± 0.00	41.63 ± 0.00
SD^1 + ScSPM	31.60 ± 0.00	37.04 ± 0.00	40.25 ± 0.00	**42.66 ± 0.00**

4.3 Caltech 256 Dataset

The Caltech 256 dataset has 256 classes of 29,780 images. Each class contains at least 80 images. Compared with the Caltech 101 dataset, images within the Caltech 256 dataset are more larger intra-class variant. We test our method on 15, 30, 45 and 60 training images randomly chosen in per class respectively. As shown in Table 3, S^3R performs better than our method. This is because it combines the visual similarity and

weak semantic similarity of the training images. Furthermore it is time-costing because of learning extra classifier and space-consuming because of requiring more space for extra sub-semantic feature. Except S^3R, we can see that our approach outperforms all the other related methods, mainly due to the addition of spatial descriptors under spatial pyramid framework. Figure 9 shows some typical images owning the top 18th classification accuracy in brackets. The first number in the bracket is the accuracy obtained by using SD^1 descriptors under spatial pyramid matching framework, the second one is the accuracy by using the method proposed by Yang et al. in [6].

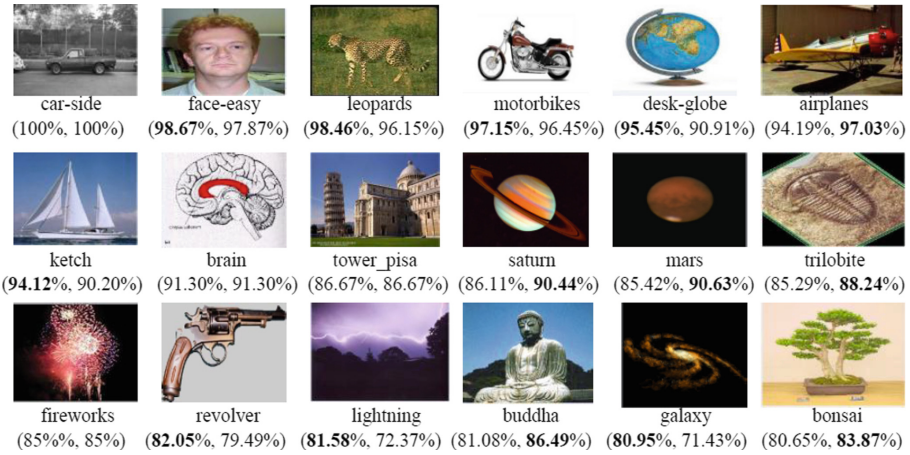

Fig. 9. Example images from classes with the top 18th classification accuracy on the Caltech 256 dataset when using SD^1 descriptors under spatial pyramid matching framework

5 Conclusion

This article focuses on boosting the performance of image classification with spatial difference information. A novel descriptor named spatial difference is proposed to describe the spatial information of differences. And this descriptor is mainly used in the combination with histograms of bag-of-visual words model under spatial pyramid matching framework, which can boost the final performance of image classification. The experimental results on the three public image datasets of the Scene 15 dataset, the Caltech 101 dataset and the Caltech 256 set demonstrate the effectiveness of the proposed method.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (61202325, 61303154, 61379100, 61370169, 60873104).

References

1. Sivic, J., Zisserman, A.: Video google: a text retrieval approach to object matching in videos. In: Proceedings of ICCV, pp. 1470–1477. IEEE (2003)
2. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognition natural scene categories. In: Proceedings of CVPR, pp. 2169–2178 (2006)
3. Teng, K., Wang, J., Tian, Q., Lu, H.: Improving scene classification with weakly spatial information. In: Proceedings of ICIP, pp. 3259–3263 (2013)
4. Grauman, K., Darrell, T.: Pyramid match kernels: discriminative classification with sets of image features. In: Proceedings of ICCV, pp. 725–760 (2005)
5. Smeulders, A., Gemert, J., Veenman, C., Geusebroek, J.: Visual word ambiguity. IEEE Trans. Pattern Anal. Mach. Intell. **32**(7), 1271–1283 (2010)
6. Yang, J., Yu, K., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: Proceedings of CVPR (2009)
7. Gao, S., Tsang, I., Chia, L.: Local features are not lonely-Laplacian sparse coding for image classification. In: Proceedings of CVPR (2010)
8. Wang, J., Yang, J., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In Proceedings of CVPR (2010)
9. Chen, Q., Song, Z., Hua, Y., Huang, Z., Yan, S.: Hierarchical matching with side information for image classification. In: Proceedings of CVPR, pp. 3426–3433 (2012)
10. Yao, B., Khosla, A., Fei-Fei, L.: Combining randomization and discriminative for fine-grained image categorization. In: Proceedings of CVPR, pp. 1577–1584 (2012)
11. Zhang, N., Farrell R., Darrell, T.: Pose pooling kernels for sub-category recognition. In: Proceedings of CVPR, pp. 3665–3672 (2012)
12. Zhang, C., Liu, J., Tian, Q., Han, Y., Lu, H., Ma, S.: A boosting, sparsity-constrained bilinear model for object recognition. IEEE Multimedia **2**, 58–68 (2012)
13. Bao, C., He, L.: Linear spatial pyramid matching using non-convex and non-negative sparse coding for image classification (2015). arXiv:1504.06897v1 [cs. CV]
14. Pasolli, E., Melgoni, F., Tuia, D., Pacifici, F., Emery, W.J.: SVM active learning approach for image classification using spatial information. IEEE Trans. Geosci. Remote Sens. **52**(4), 2217–2233 (2014)
15. Jia, S., Xie, Y., Zhu, Z.: Integration of spatial and spectral information by means of sparse representation-based classification for hyper spectral imagery. In: Proceedings of the 18th Asia Pacific Symposium of Intelligent and Evolutionary Systems, Proceedings in Adaption, Learning and Optimization (2015). doi:10.1007/978-3-319-13356-0_10
16. Zhu, C., Yang, S., Zhao, Q., Cui, S., Wen, N.: Robust semi-supervised kernel-FCM algorithm incorporating local information for remote sensing image classification. J. Indian Soc. Remote Sens. **42**, 35–49 (2014)
17. Zhang, C., Chen, J., Liu, J.: Object categorization in sub-semantic space. Neurocomputing **142**, 248–255 (2014)
18. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset, Caltech-256 Technical report UCB/CSD-04-1366 (2007)