Optical Engineering

SPIEDigitalLibrary.org/oe

Scene classification based on spatial pyramid representation by superpixel lattices and contextual visual features

Guanghua Gu
Fengcai Li
Yao Zhao
Zhenfeng Zhu

Scene classification based on spatial pyramid representation by superpixel lattices and contextual visual features

Guanghua Gu
Beijing Jiaotong University
Institute of Information Science
and
YanShan University
School of Information Science and Engineering
Qinhuangdao, China
and
Beijing Key Laboratory of Advanced Information
Science and Network Technology
Beijing, China
E-mail: guguanghua@ysu.edu.cn

Fengcai Li
YanShan University
School of Information Science and Engineering
Qinhuangdao, China

Yao Zhao,
Zhenfeng Zhu
Beijing Jiaotong University
Institute of Information Science
and
Beijing Key Laboratory of Advanced Information
Science and Network Technology
Beijing, China

Abstract. Natural scene classification is a challenging open problem in computer vision. We present a novel spatial pyramid representation scheme for recognizing scene category. Initially, each image is partitioned into sub-blocks, applying the technology of superpixel lattices segmentation according to a boosted edge learning boundary map, which makes the objects in each sub-block have the integrity—that is, the features in each sub-block are relatively consistent. Then, we extract the dense scale-invariant feature transform features of the images and form the contextual visual feature description. Finally, the image representations are performed by following the methodology of spatial pyramid. The feature descriptions we present include both local structural information and global spatial structural information; therefore, they are more discriminative for scene classification. Experiments demonstrate that the classification rate can achieve about 87.13% on a set of 15 categories of complex scenes. [DOI: 10.1117/1.OE.51.1.017201]

Subject terms: scene classification; superpixel lattices segmentation; scale-invariant feature transform; contextual visual feature; spatial pyramid.

Paper 110587 received May 27, 2011; revised manuscript received Oct. 31, 2011; accepted for publication Nov. 3, 2011; published online Feb. 7, 2012.

1 Introduction

Scene classification and automatic labeling of an image have drawn increasing attention and been widely applied to various tasks in multiple disciplines. Nonetheless, understanding the meanings or contents of images remains a challenging problem in machine intelligence and statistical learning.

Bosch et al.[1] summarized the scene classification strategies in recent decades. The basic idea is to take the whole image as an entity and represent the characteristics of the scenes relying on low-level features (e.g., color, texture, gradient, etc.).[2–4] Oliva proposed a type of global feature called "Gist"[5–8] that employs a visual attention model to combine global color, intensity, and orientation features to represent a scene. This approach may be sufficient for separating scenes with significant differences in global properties. However, scenes with similar global characteristics (e.g., bedroom vs. living room, forest vs. open country) are not easily differentiated, so the global features may not be discriminative enough. Thus, the strategies of features extracted from local regions have been proposed for scene classification.[9,10]

The bag-of-words (BOW) scheme[11,12] can be interpreted as a sparse sampling of high-level statistics of local features distribution. High-level statistics are suited for detecting the local distinctive patterns in an image, which are thought to be especially important for capturing the characteristics of solid objects. Bosch et al.[1] pointed out that visual words joined with different techniques, such as pLSA and LDA, can achieve better results for scene classification.[13–15] However, these methods disregard the spatial information of the scenes. To overcome this drawback, Lazebnik et al.[16] proposed a successful approach called spatial pyramid matching (SPM). This technique calculates the distribution of visual words at multi-spatial resolutions to form a spatial pyramid representation of an image. It employs the pyramid matching strategy[17] to measure the similarity between pyramids. However, each image is subdivided into rectangular blocks regularly so that an integral object in an image may be split into two sections or more. As a result, the features in the same block have poor consistency.

Recently, contextual information has been employed for object recognition or object detection.[18–20] The methods based on contextual information achieved better recognition performance than those based solely on local features. The idea of using context has therefore been proposed for the image segmentation or labeling task.[21] It provides a novel way of combining contextual information to improve region labeling. There are also some attempts to label regions in the scene images for scene categorization.[9,10,22] However, the region labeling is limited in these methods. In fact, an image of a scene consists of a number of objects, and the

0091-3286/2012/$25.00 © 2012 SPIE

categories of objects in images of the same category vary significantly. Thus, aiming at the problem of scene classification, many unsupervised methods focus on learning the components of images for scene classification.[23–25] Qin and Yung[26] focused on learning the components and corresponding contexts from the scene images in a totally unsupervised manner by extending the traditional visual words learning procedure. The method achieved considerable recognition success.

Unsupervised over-segmentation of an image into superpixels is a common preprocessing step for image parsing algorithms. It has achieved attractive progress in superpixel over-segmentation technology.[27,28] Moore[29,30] proposed a novel algorithm called superpixel lattices segmentation. The segmentation method produces superpixels that are forced to conform to a superpixel lattice. So, all pixels within the same superpixel block will belong to the same real-world object, ideally. The superpixel lattices segmentation algorithm divides one image into a series of superpixels based on the boundary map and boundary cost map.[31]

Although the feature descriptors proposed by Qin and Yung[26] contain local spatial information in some degree, they lack global spatial information. Moreover, it is impossible to meet the computation cost and storage requirement with the scale level increasing. Considering these shortcomings, we just extract features from three scales $(s = 1, 2, 3)$ to learn the visual words and follow the strategy of spatial pyramid matching (SPM) to add the global spatial information. Considering the consistency of features in the same block, in this paper, we apply the technology of the superpixel lattices segmentation to obtain the multi-resolutions instead of the regular rectangular segmentation in Ref. 16. The proposed method merges the pyramid description based on the technology of superpixel lattices segmentation and contextual information so that it contains local gradient information, local structural information, and global spatial information. In addition, our approach ensures the integrity of the subblocks and the consistency of the features in the same block.

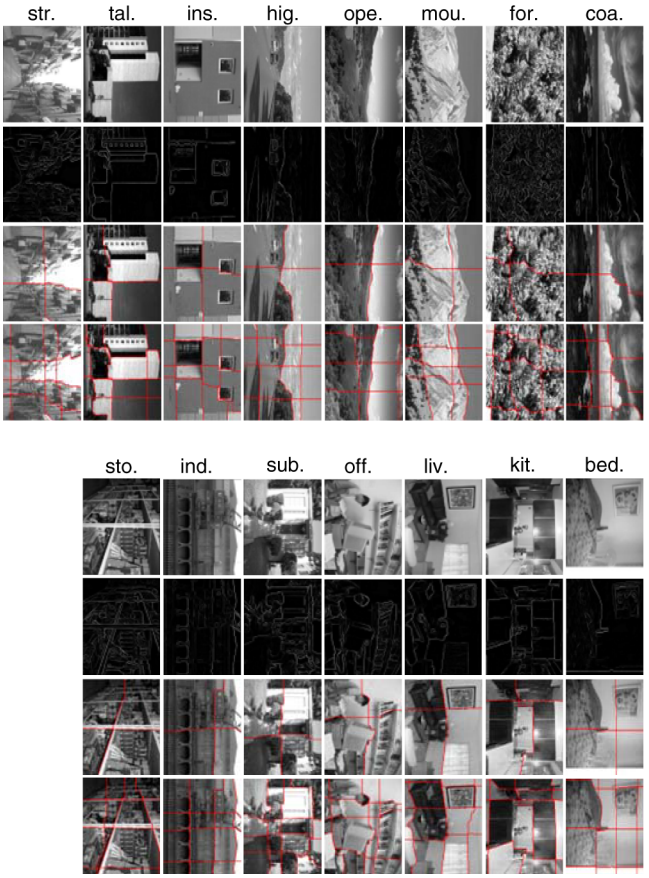


Fig. 1 Examples for superpixel lattices of each category. The first column shows original images, the second column the corresponding BEL maps, the third column 2×2 superpixel lattices, and the last column 4×4 superpixel lattices.

2 Approach

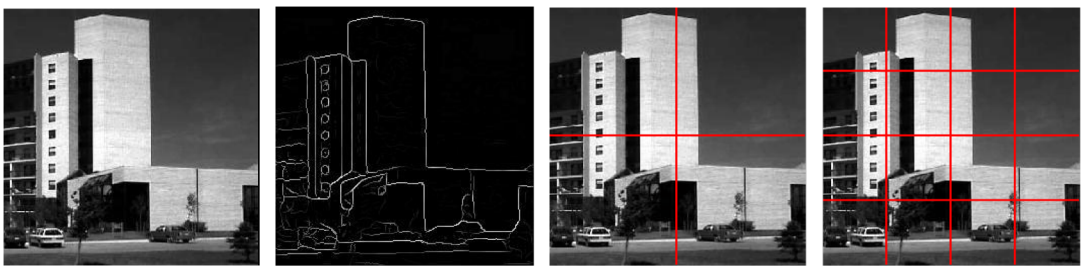
2.1 *Superpixel Spatial Pyramid Blocks*

The input data of superpixel lattices segmentation is a boundary map.[29,30] This is a two-dimensional array containing a measure of the probability that a semantically meaningful boundary is present between two pixels. Among several choices for a boundary map, Dollar et al.[31] provided an efficient algorithm to generate the fast boosted edge learning (BEL) boundary map using a boosted classifier to learn natural boundary. We follow the BEL maps shown in the second column of each group in Fig. 1.

The formation of superpixel lattices is incremental. Initially, the image is split vertically and horizontally based on its corresponding BEL map. Each path (the red line in Fig. 1) splits the image into two parts; thus, two paths (one vertical and one horizontal) cumulatively produce four superpixels (the third column of each group in Fig. 1). Next, we add two vertical paths and two horizontal paths, respectively, to generate 16 superpixels (the last column of each group in Fig. 1). The main problem at each stage is how to form each path and how to ensure these constraints are maintained. The details are mentioned in Ref. 29. We emphasize the attractive property of the technique that each vertical or horizontal path guarantees to follow the boundary. This is the key difference from the algorithm proposed in Ref. 16, which constrainedly divides the image into rectangle blocks, as shown in Fig. 2. In this paper we just take the tall building example, for instance in the following narrative. The superpixel lattices segmentation keeps the integrity of the blocks and the consistency of the features in the same block.

2.2 *Feature Extraction*

The comparative evaluation in Ref. 23 has shown that a dense image representation works better for scene classification. So, in this paper we use a dense regular grid instead of interest points at different scales $s = 1, 2, \cdots S$, which means multiscale gridspacing and patchsizes. Scale-invariant feature transform (SIFT) features are extracted from every 16×16 patch on a grid of step size 8 at scale 1. The corresponding patch sizes are 32×32 on a grid of step size 16 at scale 2 and 64×64 on a grid of step size 32 at scale 3, respectively. To obtain the contextual information, we combine the SIFT features from the coarser scale and neighborhood regions to construct a new description of the region of interest (ROI). Note that the features at the highest scale have no corresponding contextual features. The concept of constructing contextual features is illustrated in Fig. 3.



(a) (b) (c) (d)

Fig. 2 Example of regular lattices of the tall building category.

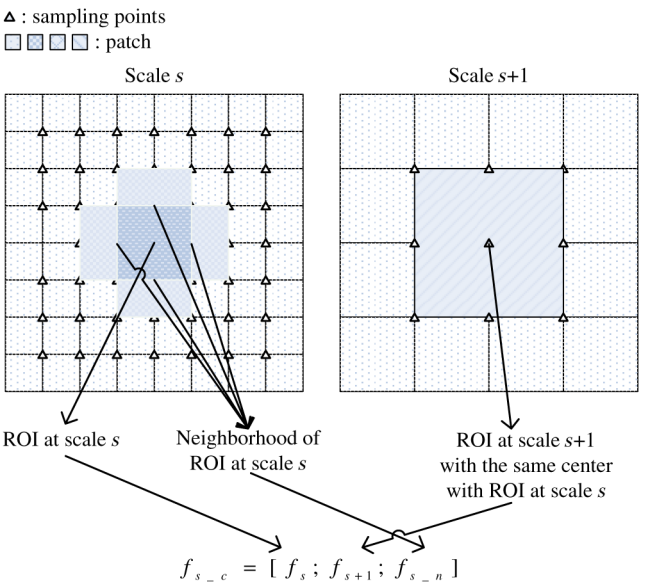


Fig. 3 The schema of the contextual features at scale *s*.

Let f_s denote the SIFT features of ROI at scale s, f_{s+1}, the SIFT features of the region having the same center as the ROI but at a coarser scale level and f_{sn} having the SIFT features of ROI neighbors at scale s. The contextual descriptor of the ROI is built by:

$$f = [f_s; w_C \times f_{s+1}; w_n \times f_{sn}], \tag{1}$$

where w_C and w_N are the weight parameters that control the significance of the features from the coarser level and the neighbor regions, respectively, in order to balance the discriminative power and generalization ability of the contextual information. For all the training images, denote the contextual visual features at scale s as $\{f_1, f_2, \cdots, f_n\}$, where n is the number of features from all the training images. PCA (principle component analysis) is used to reduce the dimension, and the covariance matrix is formed as:

$$E_s = \frac{1}{n}\sum_{k=1}^{n} f_k \cdot (f_k)^T. \tag{2}$$

Then, SVD (singular value decomposition) is performed by:

$$E_s = U_s M_s U_s^T, \tag{3}$$

where M_s is a diagonal matrix with descending singular values, i.e., $m_1 \geq m_2 \geq \cdots \geq m_d$. d represents the dimension of the contextual visual features. In this paper, $d = 768(128 + 128 + 128 \times 4)$. $U_s = \{v_1, v_2, \cdots, v_d\}$ is formed with the eigenvectors $\{v_i\}_{i=1,2,\cdots,d}$ corresponding to singular values $\{m_i\}_{i=1,2,\cdots,d}$. In our experiments Eq. 4 is used to select the reduced dimension d_1 of the features, and T is the decision parameter. $T = 0.95$ means that 0.95% information is preserved in terms of the mean-square error, which controls d_1 to avoid losing the excessive useful information.

$$\min_{d_1} \sum_{k=1}^{d_1} m_k \geq T \cdot \sum_{k=1}^{d} m_k. \tag{4}$$

The transformation matrix P_s for reducing dimension is formulated by taking the first d_1 rows of U_s^T, i.e., $P_s = \{v_1; v_2; \cdots; v_{d_1}\}$. Each lower-dimensional feature vector y is generated from higher-dimensional feature vector f as follows:

$$y = P_s \mathrm{f}. \tag{5}$$

K-means cluster method is applied to create visual words for each scene category at every scale after generating descriptors. The vocabulary at scale s is defined as $V_s = \{v_1^s, v_2^s, \cdots, v_C^s\} = \{v_c^s\}_C$, where C is the number of scene categories and v_c^s is the vocabulary of the category c at scale s. The vocabulary v_c^s is learned by the features $\{y_{1(c)}^s, y_{2(c)}^s, y_{3(c)}^s \cdots\}$ of training images belonging to category c at scale s.

2.3 *Feature Space*

Denote $l = 0, 1, \cdots, L$ as the pyramid levels in this paper. The results in Ref. 16 indicate that the performance of the entire $L = 3$ pyramid remains essentially identical to that of the $L = 2$ pyramid. In this paper we chose $L = 2$. Figure 4 shows an example of a three-level pyramid based on superpixel lattices segmentation. Level 0 denotes the original image, level 1 the 2×2 grid, and level 2 the 4×4 grid. The total number of blocks of an image is 24 $(1 + 4 + 16)$.

The procedure of spatial pyramid representation contains five steps at each scale:

a. Map each contextual feature y_s at scale s to its corresponding visual word in the vocabulary V_s.
b. Count the number of words falling into each block to form the representation of the block expressed by the

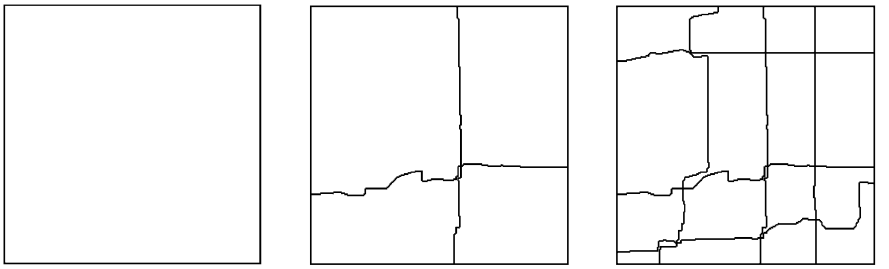


Fig. 4 The three-level superpixel lattices spatial pyramid.

histogram as $\{h_i^l\}$, $l = 0, \cdots, L$; $i = 0, \cdots, 2^{2l}$. The image representation at each level can be constructed by $h_s^l = [h_1^l; \cdots; h_{2^{2l}}^l]$, $l = 0, \cdots, L$.

c. Weight each histogram h^l by Eq. (6) and normalize it with the total number of features of the image:

$$w_{h^l} = \begin{cases} \frac{1}{2^L}, & l = 0 \\ \frac{1}{2^{L-l+1}}, & l \neq 0. \end{cases} \tag{6}$$

d. The image representation at scale s is formed as:

$$h_s^I = [h_s^0; h_s^1; \cdots; h_s^L]. \tag{7}$$

e. Iterate the steps a) to d) until $s = S - 1$The final representation of an image is produced as:

$$H^I = [h_1^I, h_2^I, \cdots, h_{S-1}^I]'. \tag{8}$$

3 Experimental Results

3.1 *Data Sets and Setup*

We evaluate the classification performance on three data sets. Data set 1 contains eight category scenes (coast, forest, mountain, open country, highway, inside city, tall building, and street) provided by Oliva and Torralba.[32] Li Fei-fei et al.[23] extended Data Set 1 by adding five other categories (bedroom, kitchen, living room, office, and suburb) to form Data Set 2. Data Set 3 is a further extension of Data Set 2 by adding two other categories (industrial and store) performed by Lazebnik et al.[16] Each scene category contains 210 to 410 images. The average size of the images is around 300×250 pixels.

All processing is performed by using gray scale images. We follow standard experimental protocols applied in previous tasks. In each experiment 100 images are randomly chosen from each category for training, and the rest for testing. The final result is reported as the average of 10 individual runs, wherein the training and testing samples are replaced randomly.

In our experiments we set the scales $s = 1, 2, 3$ and the size of each category vocabulary $K = 100$ at each scale s. The levels of superpixel spatial pyramid take $l = 0, 1, 2$. All of the experiments are based on the same parameters. Finally, we train the Library for Support Vector Machines (LIBSVM)[33] for the scene recognition.

We perform four different image representations as follows:

SPSLC: Spatial pyramid matching on superpixel lattices with contextual information

SPMC: Spatial pyramid matching with contextual information

SPSL: Spatial pyramid matching on superpixel lattices without contextual information

SPM:[16] Spatial pyramid matching without contextual information

Performance comparisons between SPSLC and SPMC and between SPSL and SPM, respectively, have been made to demonstrate the superiority of superpixel lattices. Also, performance comparisons between SPSLC and SPSL and between SPMC and SPM, respectively, have been made to show the advantage of contextual information. Furthermore, we compared our SPSLC method with three other typical methods, hierarchical Gaussianization (HG),[34] global Gaussian (GG),[35] and contextual visual words(CVW).[26]

3.2 *Results*

We observe the experimental results for two aspects: the effectiveness of superpixel lattices or contextual information. Table 1 illustrates the performance of the four representations (SPSLC, SPMC, SPSL, and SPM) based on Data Set 1. On one hand, SPSLC and SPSL both apply superpixel lattices for spatial information. SPSLC achieves the best classification rate of 94.50% at level 0, 1. The corresponding confusion table is displayed in Fig. 5. In fact, the four methods obtain the best results at level 0, 1 shown in Table 1. SPMC applies the regular lattices scheme and obtains a rate of 91.25% at level 0, 1, lower than SPSLC. Both SPSLC and SPMC integrate the contextual information. To validate the advantage of superpixel lattices, we also

Table 1 Classification result for Data Set 1.

level	SPSLC	SPMC	SPSL	SPM[16]	CVW[26]
0	89.50%	89.50%	86.75%	86.75%	**90.30%**
1	93.75%	90.35%	89.00%	86.90%	
2	91.46%	89.00%	88.00%	84.00%	
0, 1	**94.50%**	**91.25%**	**89.75%**	**87.70%**	
0, 1, 2	92.75%	90.00%	87.25%	87.00%	

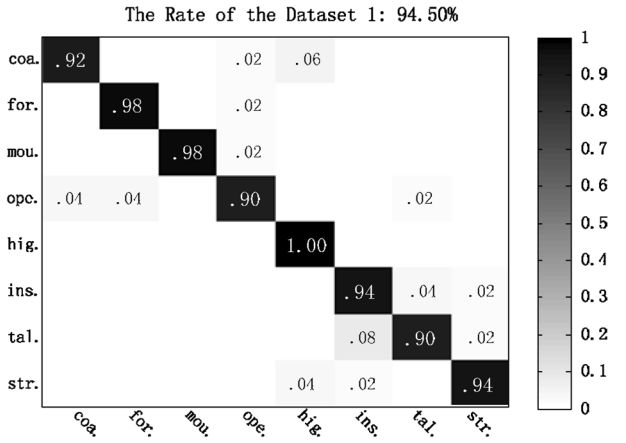


Fig. 5 The confusion table of Data Set 1.

compared SPSL with SPM without the contextual information. Experiments demonstrate that SPSL (89.75%) outperforms SPM (87.70%). On the other hand, SPSLC and SPMC both integrate contextual information for local structural information. With superpixel lattices, SPSLC (94.50%) is better than SPSL (89.75%), and without superpixel lattices SPMC (91.25%) is better than SPM (87.70%). Experiments validate the superiority of contextual information. At last, the comparisons with CVW[26] are performed. We can see from Table 1 that our SPSLC outperforms CVW by 4.2%.

Table 2 and Table 3 display the performance of the four representations based on Data Set 2 and Data Set 3. The conclusions from Table 2 and Table 3 are the same as from Table 1: 1.the methods applying superpixel lattices are superior to ones without superpixel lattices, and 2.the methods using contextual information are superior to ones without contextual information. Table 2 and Table 3 shows that SPSLC is better than SPMC and that SPSL is better than SPM, which demonstrates the superiority of superpixel lattices. Also, SPSLC is better than SPSL, and SPMC is better than SPM, which proves the effectiveness of contextual information. In addition, SPSLC (88.85%) at level 0, 1 outperforms CVW (87.63%) from Table 2, and SPSLC (87.13%) at level 0, 1 outperforms HG (85.20%), GG (86.10%), and CVW (85.16%) from Table 3.

Data Set 2 and Data Set 3 contain some indoor categories, which increases the difficulty of classification. As Quattoni et al. pointed out in Ref. 36, it is more difficult for indoor scene classification than outdoors. The corresponding confusion tables of SPSLC based on Data Set 2 and Data Set 3 are shown in Fig. 6 and Fig. 7, respectively. It is clear that the classification rate of each indoor scene is lower than outdoors, which matches the conclusion proposed by Quattoni.[36]

To learn the influence of our approach for the outdoors, and indoors, respectively, we split Data Set 3 into 10 categories of outdoor scenes (note that we take the industrial category that contains only outdoor scenes) and five categories of indoor scenes. The results are shown in Table 4 and Table 5, respectively. Figure 8 and Fig. 9 show the corresponding confusion rates of SPSLC at level 0, 1. For the outdoors, the rate increases only 2.27% by SPSLC from level 0 to level 1. However, for the indoors, the rate increases 4.44% by SPSLC from level 0 to level 1. The other three methods (SPMC, SPSL, and SPM) reveal similar phenomena.

Table 2 Classification result for Data Set 2.

level	SPSLC	SPMC	SPSL	SPM[16]	CVW[26]
0	83.13%	83.13%	77.90%	77.90%	**87.63%**
1	87.16%	84.06%	83.15%	80.24%	
2	85.54%	83.37%	82.46%	77.20%	
0, 1	**88.85%**	**86.00%**	**84.77%**	**81.46%**	
0, 1, 2	87.07%	84.25%	82.00%	79.00%	

4 Discussions

We discuss our approach for the following four aspects: advantages of superpixel lattices, advantages of contextual information, the selection of the levels in a spatial pyramid scheme, and the performance of SPSLC in the outdoors and indoors.

As mentioned above, to obtain the global spatial information, SPM[16] method is proposed to form a spatial pyramid representation of an image. However, it used regular lattices to divide images into rectangular blocks regularly so that an integral object in one image may be split into two sections or more. As a result, the features in the same sub-block have poor consistency. To overcome this problem, we apply the pyramid representation on superpixel lattices instead of regular lattices. The superpixel lattices scheme keeps the

Table 3 Classification results for Data Set 3.

level	SPSLC	SPMC	SPSL	SPM[16]	HG[34]	GG[35]	CVW[26]
0	82.93%	82.93%	76.28%	76.28%	**85.20%**	**86.10%**	**85.16%**
1	85.90%	83.67%	80.80%	78.13%			
2	83.46%	81.87%	79.07%	76.75%			
0, 1	**87.13%**	**84.26%**	**82.00%**	**79.50%**			
0, 1, 2	85.18%	82.53%	78.40%	76.92%			

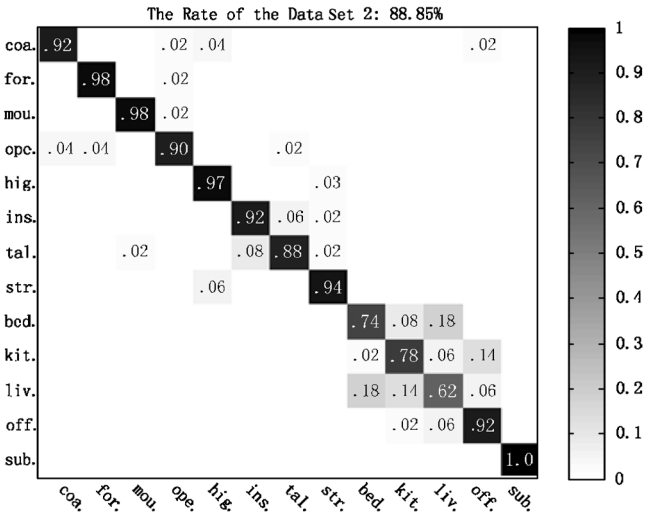


Fig. 6 The confusion table of Data Set 2.

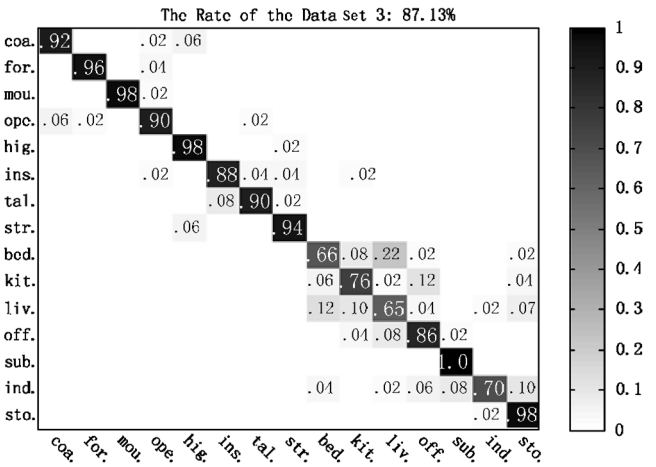


Fig. 7 The confusion table of Data Set 3.

integrity of the sub-blocks and the consistency of the features in the same sub-block. We test our SPSLC by five data sets, and the performances are shown in Tables 1 to 5. The superiority of superpixel lattices is demonstrated by two aspects: with contextual information and without contextual information. Comparing the two methods (SPSLC and SPMC) using

Table 4 Classification results for 10 outdoors categories of 15 scenes.

level	SPSLC	SPMC	SPLS	SPM[16]
0	91.15%	91.15%	86.92%	86.75%
1	93,42%	92.38%	89.00%	88.07%
2	93.00%	91.59%	88.00%	87.00%
0, 1	**94.20%**	**92.94%**	**89.75%**	**88.94%**
0, 1, 2	93.40%	92.00%	87.25%	87.51%

Table 5 Classification results for five indoors categories of 15 scenes.

level	SPSLC	SPMC	SPLS	SPM[16]
0	71.36%	71.36%	67.00%	67.00%
1	76.80%	74.95%	71.20%	70.25%
2	75.27%	72.86%	71.20%	69.75%
0, 1	**78.20%**	**75.64%**	**73.20%**	**71.35%**
0, 1, 2	76.44%	73.37%	69.20%	68.50%

contextual information, we can see from Tables 1 to 5 that SPSLC outperforms SPMC. For the two methods (SPSL and SPM) without contextual information, SPSL is better than SPM. The comparisons indicate the effectiveness of superpixel lattices.

To keep the local structural information, contextual information is integrated in the feature description. We also validate the effectiveness of contextual information by two aspects: with superpixel lattices and without them. First, for the two methods (SPSLC and SPSL) using superpixel

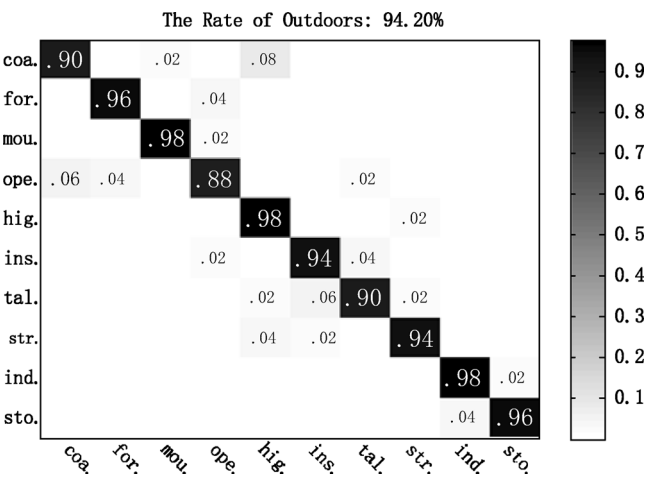


Fig. 8 The confusion table of 10 outdoors categories.

Fig. 9 The confusion table of five indoors categories.

lattices, SPSLC applies contextual information, but SPSL does not. It is clear from Tables 1 to 5 that SPSLC is superior to SPSL. Second, for the two methods (SPMC and SPM) without superpixel lattices, SPMC integrates contextual information, but SPM does not. Experimental results show that SPMC is better than SPM. The comparisons demonstrate the effectiveness of contextual information.

To add the spatial structural information, in this paper spatial pyramid representation strategy is used for image representation. We expect the results can be better at level 2 than either level 0 or level 1 as the results mentioned in Ref. 16. Tables 1 to 5 show that the results improve from level 0 to level 1. We take it for granted that the performance can also increase from level 1 to level 2. However, it is not true. The reason may be that the higher level of the pyramid is oversubdivided with individual bins yielding too few matches. It implies that the higher level of the pyramid restricts the structure of each category scene and only fits the scenes with a single structure. But, in fact, the scenes belonging to the same category with intra-class variations have no fixed structures. In Tables 1 to 5, the performance is best at level 0, 1. This indicates the main advantage of the spatial pyramid representation: it combines multiple resolutions in a principled fashion, so it is resistant to failures at individual levels.

Indoor scenes recognition is a challenging open problem in computer visions. Whereas most outdoor scenes can be well characterized by global properties, indoor scenes cannot. There is a wide range of both local and global discriminative information for most indoor scenes. Moreover, indoor scenes (e.g., kitchen, living room, etc.) exhibit much variability across the different exemplars within each category. It is not true for outdoor scenes (e.g., street, tall building, etc.). Therefore, both outdoor and indoor scenes need a model that can exploit the local and global discriminative information to solve the recognition task. Fortunately, our approach, SPSLC, could meet these requirements. Considering the difference between indoor and outdoor scenes, we tested the advantage of SPSLC on a 10-category outdoors data set and a five-category indoors data set provided by DataSet 3. The results, shown in Table 4 and Table 5, illustrate that SPSLC improves the performance of classification both for outdoor scenes and indoor scenes. In addition, we can conclude from Table 4 and Table 5 that the gain of the performance indoors is higher than outdoors from level 0 to level 1 or from level 0 to level 0, 1, so that our approach improves the performance more for indoors than outdoors.

5 Conclusions

This paper proposed a new image representation method using spatial pyramid representation strategy on superpixel lattices with contextual information. Experimental results demonstrate that it has powerful distinctiveness on the scene classification task due to two reasons: one is the superiority of the superpixel lattices segmentation technology because it guarantees the integrity of the sub-blocks and the consistency of the features in the same sub-block; the other is that the feature descriptions contain both the local and global structural information.

Even so, our scheme has a limitation in the stage of superpixel lattices segmentation because not all the scenes can be split into sub-blocks following our expectation, which affects the performance of scene classification. Despite this limitation, our method outperforms other state-of-the-art methods on scene classification. Our future work will be to study further how to improve the superpixel lattices segmentation and how to improve the classification performance of the indoor scenes.

Acknowledgments

The authors would like to thank the anonymous reviewers for valuable comments. This work was partly supported by Natural Science Foundation of China (No. 61025013 and No. 61172129), Sino-Singapore JRP (No. 2010DFA11010), Beijing Natural Science Foundation, Fundamental Research Funds for the Central Universities (No. 2009JBZ006), and Program for Plan of Science and Technology of Qinhuangdao City (No. 201101A084).

References

1. A. Bosch, X. Munoz, and R. Marti, "A review: which is the best way to organize/classify images by content," *Image Vis. Comput.* **25**(6), 778–791 (2007).
2. M. Szummer and R. W. Picard, "Indoor-outdoor image classification," *Proc. IEEE International Workshop on Content-Based Access of Image and Video Databases (CAIVD) 445*, pp. 42–50, IEEE Computer Society, Bombay, India (1998).
3. A. Vailaya et al., "Image classification for content-based indexing," *IEEE Trans. Image Process.* **10**(1), 117–129 (2001).
4. C. Wallraven, B. Caputo, and A. B. A. Graf, "Recognition with local features: the kernel recipe," *IEEE International Conference on Computer Vision (ICCV)*, 257–264, IEEE Computer Society, Nice, France (2003).
5. A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *Int. J. Comput. Vis.* **42**, 145–175 (2001).
6. A. Oliva and A. Torralba, "Building the gist of a scene: the role of global image features in recognition," *Prog. Brain Res.* **155**, 433–442 (2006).
7. A. Oliva, "Gist of the scene," In *The Encyclopedia of Neurobiology of Attention*, pp. 251–256, Elsevier, San Diego (2005).
8. C. Siagian and L. Itti, "Gist: a mobile robotics application of context-based vision in outdoor environment," in *IEEE CVPR Workshop on Attention and Performance in Computer Vision (CVPR/APCV)*, pp. 88, IEEE Computer Society, San Diego, CA, USA (2005).
9. J. Luo and A. Savakis, "Indoor vs outdoor classification of consumer photographs using low-level and semantic features," in *Proc. of ICIP*, A. Savakis, Ed., pp. 745–748 (2001).
10. J. Vogel and B. Schiele, "A semantic typicality measure for natural scene categorization," in *DAGM*, Springer, Berlin, pp. 195–203 (2004).
11. G. David, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.* **60**(2), 91–110 (2004).
12. G. Csurka et al., "Visual categorization with bags of keypoints," in *Proc. ECCV Workshop on Statistical Learning in Computer Vision*, pp. 59–74, Springer, Prague, Czech (2004).
13. A. Bosch, A. Zisserman, and X. Munoz, "Scene classification via PLSA," in *European Conference on Computer Vision (ECCV)*, Vol. **1**, no. 4, *ACM*, Graz, Austria, pp. 517–530 (2006).
14. D. Yang, B. Li, and H. Zhao, "An adaptive algorithm for robust visual codebook generation and its natural scene categorization application," *J. Electron. Inform. Technol.* **32**(9), 2033–2038 (2010).
15. J. Vogel and B. Schiele, "Semantic modeling of natural scenes for content-based image retrieval," *Int. J. Comput. Vis.* **72**(2), 133–157 (2007).
16. S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2169–2178, IEEE Computer Society, New York, NY, USA (2006).
17. K. Grauman and T. Darrell, "The pyramid match kernel: discriminative classification with sets of image features," in *Tenth IEEE International Conference on Computer Vision*, pp. 1458–1465, IEEE Computer Society, Beijing, China (2005).
18. D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," *Int. J. Comput. Vis.* **80**(1), 3–15 (2008).
19. G. Heitz and D. Koller, "Learning spatial context: using stuff to find things," in *European Conference on Computer Vision (ECCV)*, pp. 30–43, Springer, Marseille, France (2008).
20. C. Wu and H. Aghajan, "Using context with statistical relational models: object recognition from observing user activity in home environment," in *Proceedings of the Workshop on Use of Context in Vision*

Processing (ICMI-MLMI/UCVP), pp. 1–6, ACM, Boston MA, USA (2009).
21. J. Letham, N. M. Robertson, and B. Connor, "Contextual smoothing of image segmentation," in *Proceedings of Computer Vision and Pattern Recognition Workshop on Use of Context in Video Processing (CVPR/UCVP)*, pp. 7–12, IEEE Computer Society, San Francisco, CA, USA (2010).
22. J. Yu and J. Luo, "Leveraging probabilistic season and location context models for scene understanding," in *ACM International Conference on Content-based Image and Video Retrieval (CIVR)*, pp. 169–178, ACM, Niagara Falls, Canada (2008).
23. L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 524–531, IEEE Computer Society, San Diego, CA, USA (2005).
24. P. Quelhas et al., "Modeling scenes with local descriptors and latent aspect," *ICCV* **1**, 883–890 (2005).
25. A. Bosch, A. Zisserman, and X. Muoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE Trans. Pattern Anal. Mach. Intell.* **30**(4), 712–727 (2008).
26. J. Qin and N. H. C. Yung, "Scene categorization via contextual visual words," *Pattern Recogn.* **43**(5), 1874–1888 (2010).
27. J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(8), 888–905 (2000).
28. P. Felzenszwalb and D. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.* **59**(2), 167–181 (2004).
29. A. P. Moore et al., "Superpixel lattices," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, IEEE Computer Society, Anchorage, AK, USA (2008).
30. A. P. Moore et al., "Scene shape priors for superpixel segmentation," in*Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 771–778, IEEE Computer Society, Kyoto, Japan (2009).
31. P. Dollar, Z. Tu, and S. Belongie, "Supervised learning of edges and object boundaries," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 1964–1971, IEEE Computer Society, New York, NY, USA (2006).
32. A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *Int. J. Comput. Vis.* **42**(3), 145–175 (2001).
33. C. Chang and C. Lin, "LIBSVM: a library for support vector machines," http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf, pp. 3 (2010).
34. X. Zhou et al., "Hierarchical Gaussianization for image classification," in *ICCV*, pp. 1971–1977, IEEE Computer Society, Kyoto, Japan (2009).
35. H. Nakayama, T. Harada, and Y. Kuniyoshi, "Global Gaussian approach for scene categorization using information geometry," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2336–2343, IEEE Computer Society, San Francisco, CA, USA (2010).
36. A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, IEEE Computer Society, Miami, FL, USA (2009).

Guanghua Gu received his BSc and MSc degrees from Yanshan University, China, in 2001 and 2004, respectively. He is currently a PhD candidate of Beijing Jiaotong University. He is working in Yanshan University, China. His research interests include image classification, image recognition, and image analysis.

Biographies and photographs of the other authors are not available.