IMPROVING SCENE CLASSIFICATION WITH WEAKLY SPATIAL SYMMETRY INFORMATION

Kezhen Teng[*], *Jinqiao Wang*[*], *Qi Tian*[†], *Hanqing Lu*[*]

[*]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Science
[†]Dept. of Computer Science, University of Texas, San Antonio, USA
{kezhen.teng, jqwang, luhq}@nlpr.ia.ac.cn, qitian@cs.utsa.edu

ABSTRACT

The bag-of-visual-words (BOW) model has been widely used in the field of scene classification. Since it ignores the spatial information, the spatial-pyramid-matching (SPM) model [1] was presented by partitioning the image into increasingly fine blocks and computing histograms of local features in each block. However, the spatial symmetry has never been considered explicitly in scene classification as we known. In this paper, a novel descriptor named weakly spatial symmetry (WSS) is proposed to boost the performance of image classification. After region segmentation, the spatial symmetry is represented by L1 distances of region histograms. Four kinds of spatial symmetry are extracted in blocks of increasing scales as in SPM [1]. The WSS descriptor can be used independently or combined with BOW or SPM for scene classification. Experiments on scene-15 and caltech101 dataset demonstrate the effectiveness of the proposed approach.

Index Terms— scene classification, spatial symmetry

1. INTRODUCTION

Scene classification is one of hot topics in the communities of computer vision and multimedia processing. The last ten years has witnessed a blow up of different models. Beyond all of these is the bag-of-visual-words (BOW) [2], which describes the image content as a histogram of visual words. Though achieves good performance, it exists obvious weakness of neglecting spatial information and correlations among visual content. Then a spatial-pyramid-matching (SPM) model [1] was proposed to tackle this problem by dividing the whole image into hierarchical blocks and concatenating the appropriately weighted histograms of all the blocks. Experimental results demonstrate that SPM is a powerful model in scene classification. The SPM model has been improved in recent years. Some of them[3, 4, 5, 6] tried to improve the coding procedure to minimize the representation information loss. Gemert et al.[3] adopted soft quantization instead of hard quantization, and Yang et al.[4] used sparse coding, Wang et al.[5] found that locality was more important than sparsity. Gao et al.[6] proposed the Laplace-sparse-coding

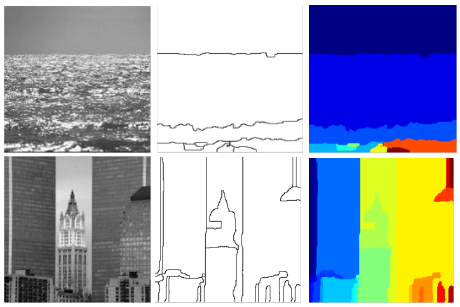

Fig. 1: Examples of spatial symmetry in natural scenes.

SPM (LScSPM) which achieved state-of-the-art performance. Others explored better pooling strategy to extract the most salient visual content. Boureau et al.[7] compared different pooling methods as max pooling and average pooling. Feng et al.[8] explored a general pooling strategy. Jia et al.[9] tried to learn an adaptive pooling partition of images.

Though successfully applied in image classification, there exist some limitations in the SPM model. Firstly, the finer the division, the more sensitive it is to the location and orientation of visual content. As illustrated in [10], the performance of SPM degrades if images were flipped or rotated randomly. Especially, for scene classification task, the intra-class variance of spatial information is more serious than general object image classification where the salient visual objects usually locate around the image center.

Due to these problems in SPM model, we attempt to utilize the spatial symmetry information. As shown in Fig.1, we give two examples from different categories in scene-15 dataset to show the characteristics of spatial symmetry. The source images, dominant edges and segmented regions filled with different colors are showed in three columns respectively. Obviously, the segmentation result in "coast" image contains left to right symmetry, while the result of the "building" image contains top to down symmetry. Therefore, we argue that the symmetry information could serve as a feature to assist distinguishing different scene images, which has not been discussed before to the best of our knowledge.

In this paper, a weakly spatial symmetry (WSS) descrip-

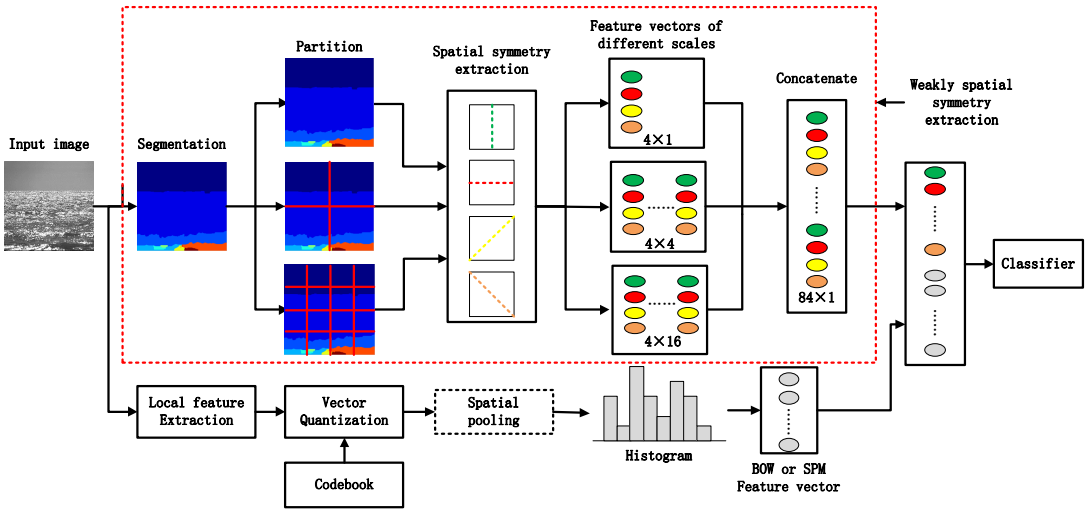

Fig. 2: The proposed scene classification framework.

tor is proposed to boost the performance of BOW model. By combining WSS and BOW, a hybrid descriptor is proposed to model the visual content and spatial symmetry together in a more compact way, which could effectively improve the accuracy of scene classification.

2. THE PROPOSED APPROACH

The framework of our scene classification is illustrated as Fig.2. The input image is first segmented into regions and the WSS features are represented by distances of region histograms. Then we model the spatial symmetry by a spatial pyramid like SPM [1], and further encode them into the BOW model. A nonlinear SVM classifier is used for final classification.

2.1. Weakly spatial symmetry model

2.1.1. Segment image into regions

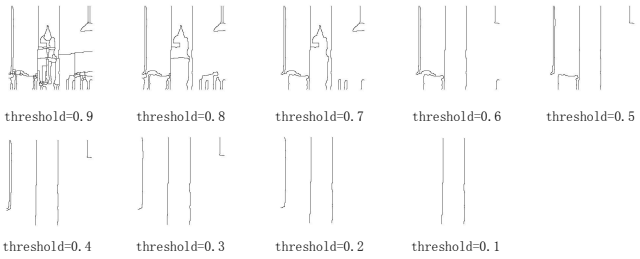

Fig. 3: Examples of segmentation results with different thresholds.

Before extracting the spatial symmetry information, we adopt two different ways to partition images into regions. The first way uses a meticulous segmentation approach in [11]. By adjusting the threshold increasingly, the hierarchical segmentation results can be obtained. This process is shown in Fig.3.

The second way is to obtain a rough segmentation by the visual words distribution. Recall that in BOW model, the local patches are sampled densely and each patch is labeled with the nearest visual word in codebook. Thus each pixel can be given the same label as its nearest local patch. In this way, the region number is equal to size of the codebook.

2.1.2. Description of weakly spatial symmetry

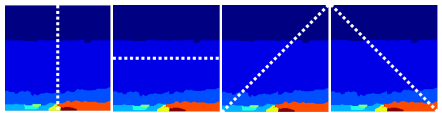

Fig. 4: Four kinds of weakly spatial symmetry.

After region segmentation, four kinds of spatial symmetry are extracted as illustrated in Fig.4. The image is partitioned into two parts in four directions. For example, the first partition describes left to right symmetry and the second describes top to down symmetry. For each part, a histogram is generated to represent pixel distribution for each region. The histogram distance between two parts is used to describe the spatial symmetry.

Specifically, the image is segmented into N regions with labels 1,2,...,N. N is determined by threshold and codebook size respectively under two segmentation methods. And the area of each region is represented by $A_1, A_2, \ldots, A_N$. Obviously $\sum_{i=1}^{N} A_i = A$, where A is the area of the whole image.

For each kind of image partition, two histograms h_1 and

h_2 are built based on the pixel numbers

$$h_1 = [h_{11}, h_{12}, , h_{1N}] \quad h_2 = [h_{21}, h_{22}, , h_{2N}] \tag{1}$$

where h_{ki} $(k = 1, 2, i = 1, 2, \ldots, N)$ represents number of pixels in i th region. Since there exists obvious left to right symmetry in Fig.4, the histograms of the left partition in Fig.4 are similar as illustrated in Fig.5. We use L1 distance to compute the spatial symmetry between two parts. Now we can

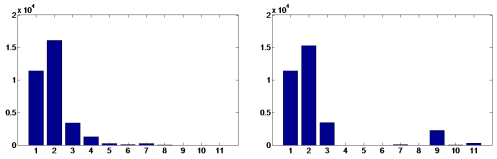

Fig. 5: Histogram comparison of left and right part of "coast" image.

obtain a 4-d feature vector to describe the spatial symmetry of the whole image. Like SPM model, we partition the image into hierarchical coarse-to-fine blocks as shown in Fig. 6, calculate the spatial symmetry for each block, and concatenate into a WSS feature vector,

$$H_{wss} = [h_{w1}, h_{w2}, \ldots, h_{wN}] \tag{2}$$

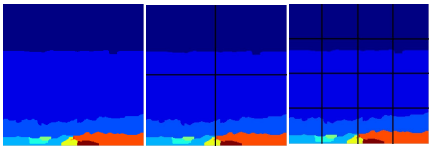

Fig. 6: Illustration of spatial pyramid partition.

2.2. Combing WSS and BOW for scene classification

We encode the spatial symmetry descriptors into the BOW based classification framework. In BOW model, the histogram of visual words is represented as,

$$H_{bow} = [h_{b1}, h_{b2}, \ldots, h_{bM}] \tag{3}$$

where M is the codebook size. Then we combine the WSS features vector with H_{bow} as follow,

$$\begin{aligned} H_{bow+wss} &= [H_{bow}, cH_{wss}] \\ &= [h_{b1}, \ldots, h_{bM}, ch_{w1}, \ldots, ch_{wN}] \end{aligned} \tag{4}$$

where c is the weight of the WSS feature and can be learned by cross validation. Finally, a nonlinear SVM classifier is trained for scene classification.

3. EXPERIMENTS

We verify the proposed approach on two publicly available datasets: scene-15 and caltech101. In our experiments, SIFT features are densely extracted with a step of 8 pixels and a patch size of 16×16. Visual codebook is built with K-means on randomly selected features. Hard quantization and average pooling are used to image encoding. And the intersection kernel SVM is used for classification.

3.1. Experimental results on scene-15

Scene-15 dataset contains 4495 images of fifteen categories of natural and indoor scenes such as coast, tall building, kitchen and so on. The number of images per category varies from 200 to 400. For each category, 100 randomly selected images are used for training and the rest for testing. The experiments are conducted ten times and mean accuracy is used for comparison. Since the image can be segmented by two different ways, we show their results respectively.

3.1.1. Sensitivity to segmentation threshold

As illustrated in Sec.2.1.1, the WSS feature depends on the segmentation results with the approach [11]. Thus an experiment is performed to show the sensitivity to the threshold. As Table 1, the classification performance of WSS model remains relative stable when threshold varies between 0.1 and 0.5. The classification performance falls rapidly if the threshold bigger than 0.5, since the number of salient edges are greatly reduced.

Table 1: Performance of WSS model with different thresholds on scene-15.

Sensitivity to segmentation thresholds					
Threshold	0.1	0.2	0.3	0.4	0.5
Accuracy	50.45%	51.59%	51.16%	50.52%	49.75%
Threshold	0.6	0.7	0.8	0.9	
Accuracy	45.7%	43.38%	36.28%	26.83%	

3.1.2. Sensitivity to different pyramid levels

To show the sensitivity to different pyramid levels, an experiment is performed in scene-15. As shown in Table 2, the first row shows classification accuracy of different levels from coarse to fine. The finer level achieves better results than the coarser ones. This is easy to understand because finer level contains detailed information of the spatial symmetry information for local regions. As can be concluded from Tab.2, the WSS feature achieves an accuracy of 51.59% independently.

3.1.3. Comparison with different approaches

We compare our approach with BOW [2], SPM [1]. Four codebooks are built with different size. For the WSS feature

Table 2: Results on sensitivity to spatial pyramid levels in scene-15.

Sensitivity to spatial pyramid levels			
Spatial Pyramid	1x1	2x2	4x4
Accuracy	26.57%	36.91%	48.54%
Spatial Pyramid	1x1 2x2	2x2 4x4	1x1 2x2 4x4
Accuracy	40.34%	50.92%	51.59%

can be extracted through two different ways of segmentation, we will give the results respectively. For using the segmentation approach [11], as shown in Table 3, different approaches are tested on four different sizes of codebook. WSS alone reaches 51.59% which is lower than BOW since the dimension of WSS is only 84-d. But WSS is more complementary with BOW than SPM. It can be observed that with certain codebook, WSS+BOW always achieve better results than SPM [1] with a shorter feature, and WSS+SPM always achieve the highest accuracy. Fig.7 shows the performance of different approaches under different codebooks. SPM is better than BOW which is consistent with Lazebnik et al. [1], and WSS+BOW is always better and more stable than SPM with much shorter feature vector (84-d). WSS+SPM achieve 81.51%, which is about 3.45% higher than SPM. It can also be seen in Fig. 7 that as the codebook size becomes bigger, the performance of BOW keeps ascending, while SPM peaks at codebook size of 1024 and drops at 2048. Since the two hybrid methods rely respectively on BOW and SPM, their performance also follow the similar trend. For image segmentation by BOW model, as illustrated in Table 4, WSS+BOW also achieves a better performance than BOW and SPM, although this is a relative coarse segmentation. But the results are not as good as WSS with the segmentation approach [11] in Table 3.

Table 3: Comparison with different approaches on scene-15 where images are segmented by algorithm in [11].

Codebook Size	256	512	1024	2048
BOW	71.59% (256-d)	74.04% (512-d)	75.31% (1,024-d)	76.05% (2,048-d)
SPM	77.08% (5,376-d)	77.35% (10,752-d)	78.06% (21,504-d)	77.32% (43,008-d)
WSS+BOW	78.46% (340-d)	78.96% (596-d)	79.56% (1,108-d)	79.53% (2,132-d)
WSS+SPM	80.50% (5,460-d)	80.70% (10,836-d)	81.51% (21,588-d)	80.34% (43,092-d)

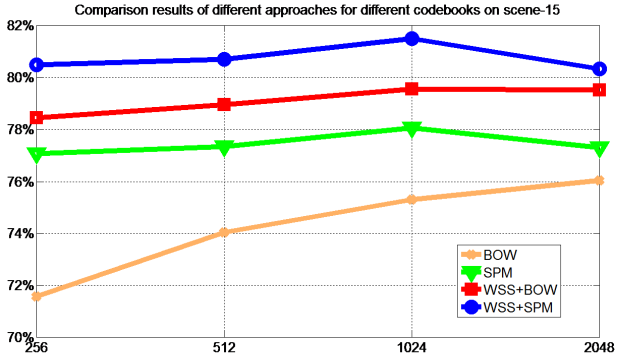

Fig. 7: Comparison results of different approaches for different codebooks on scene-15.

Table 4: Comparison of different approaches on scene-15 dataset where images are segmented incidentally by BOW model.

Accuracy	256	512	1024	2048
BOW+WSS	77.72% (340-d)	78.43% (596-d)	78.63% (1,108-d)	78.66% (2,132-d)
SPM+WSS	79.97% (5,460-d)	80.54% (10,836-d)	80.50% (21,588-d)	79.36% (43,092-d)

3.2. Experimental results on Caltech101

Caltech101 contains 9145 images of 102 categories (including a background category). We randomly select 30 images in each categories for training, and the rest for testing (number of test images not exceeds 50). As [10] mentioned, the images in Caltech101 are all aligned too well to test the true classification ability of different methods. For example, if we randomly flip or rotate some images, the performance of SPM will drop significantly.

The results are shown in Tab.5. Only the codebook of size 1024 is considered for simplicity. All images are flipped left to right with a probability of 0.5. As in Tab.5, performance of SPM dropped 7.81% after flipping, while WSS+BOW only dropped 2.18%. Therefore its performance is lower than SPM because its feature dimension (1108) is almost only 5% of SPM's (21504). However the performance of our WSS+SPM is higher than SPM, which demonstrates that WSS contains extra spatial information.

Table 5: Performance comparison of different models on Caltech101.

Model	BOW	WSS+BOW	SPM	WSS+SPM
Without flip	48.96%	56.20%	66.21%	67.57%
Left-right flip	48.96%	54.02%	58.40%	60.78%
Accuracy drop	0%	2.18%	7.81%	6.79%

4. CONCLUSIONS

This paper focuses on improving the performance of scene classification with spatial symmetry information. A novel feature named weakly spatial symmetry (WSS) is proposed to describe the spatial symmetry. This feature could be effectively encoded in the BOW based classification framework. The comprehensive experimental evaluations on the two public datasets of Scene-15 and Caltech101 demonstrate the effectiveness of the proposed method.

5. ACKNOWLEDGEMENT

This work was supported by 973 Program (2010CB327905) and National Natural Science Foundation of China (61273034, 61070104 and 61272329).

6. REFERENCES

[1] C. Schmid S. Lazebnik and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*. IEEE, 2006.

[2] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *CVPR*. IEEE, 2005, pp. 524–531.

[3] A. Smeulders J. Gemert, C. Veenman and J. Geusebroek, "Visual word ambiguity," in *Transactions on Pattern Recognition and Machine Itelligence*. IEEE, 2010.

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

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

[6] L. Chia P. Zhao S. Gao, I. Tsang, "Local features are not lonely-laplacian sparse coding for image classification," in *CVPR*. IEEE, 2010.

[7] Y. LeCun J. Ponce Y. Boureau, F. Bach, "Learning mid-level features for recognition," in *CVPR*. IEEE, 2010.

[8] Q. Tian S. Yan J. Feng, B. Ni, "Geometric lp-norm feature pooling for image classification," in *CVPR*. IEEE, 2012.

[9] C. Huang Y. Jia and T. Darrell, "Beyond spatial pyramids: Receptive field learning for pooled image features," in *CVPR*. IEEE, 2012.

[10] Y. Lu Q. Tian X. Li, Y. Song, "Spatial pooling for transformation invariant image representation," in *Multimedia*. ACM, 2011.

[11] C. Fowlkes J. Malik P. Arbelaez, M. Maire, "Contour detection and hierarchical image segmentation," in *Transactions on Pattern Recognition and Machine Itelligence*. IEEE, 2011.