Superpixel Segmentation based Structural Scene Recognition

Shuhui Bu
Northwestern Polytechnical University
Xi'an, China
bushuhui@nwpu.edu.cn

Zhenbao Liu*
Northwestern Polytechnical University
Xi'an, China
liuzhenbao@nwpu.edu.cn

Junwei Han
Northwestern Polytechnical University
Xi'an, China
jhan@nwpu.edu.cn

Jun Wu
Northwestern Polytechnical University
Xi'an, China
junwu@nwpu.edu.cn

ABSTRACT

This paper presents a novel structural model based scene recognition method. In order to resolve regular grid image division methods which cause low content discriminability for scene recognition in previous methods, we partition an image into a pre-defined set of regions by superpixel segmentation. And then classification is modelled by introducing a structural model which has the capability of organizing unordered features of image patches. In the implementation, CENTRIST which is robust to scene recognition is used as original image feature, and bag-of-words representation is used to capture the local appearances of an image. In addition, we incorporate adjacent superpixel's differences as edge features. Our models are trained using structural SVM. Two state-of-the-art scene datasets are adopted to evaluate the proposed method. The experiment results show that the recognition accuracy is significantly improved by the proposed method.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Object Recognition*; I.5.4 [**Pattern Recognition**]: Application—*Computer Vision*

General Terms

Algorithms, Experimentation, Performance.

Keywords

Scene recognition, Bag of words, Image segmentation, Superpixel, Structural SVM

MM'13, October 21–25, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2502081.2502178.

1. INTRODUCTION

In recent decades, there has been growing interest in scene recognition. How to capture the intrinsic representation of a scene is the key to the success of scene recognition. Research from Oliva *et al.* [1] indicates that recognition of scenes could be accomplished by using global configurations, without detailed object information. They proposed Gist descriptor to represent such spatial structures. Nevertheless, its performance drops dramatically when applied for indoor environments. Lazebnik *et al.* [2] proposed a spatial pyramid matching (SPM) method for scene recognition. The method works by dividing the image into increasingly fine sub-blocks and building a bag of features representation from each sub-block. A major drawback of this method is their high dimensionality. Usually high accuracy is achieved when large numbers of visual words and sub-blocks are used to produce the image representation. Jia *et al.* [3] proposed a receptive field learning method for pooling image features which can overcome drawbacks of SPM. Russakovsky *et al.* [4] proposed object-centric spatial pooling approach for scene recognition which uses inferred location of objects to pool foreground and background features separately. Recently, Wu *et al.* [5] proposed the CENsus TRansform hISTogram (CENTRIST) descriptor which use the Census Transform to capture spatial relations between neighboring pixels. They showed that a spatial hierarchy of such descriptors combined with support vector machine (SVM) classifiers performs well on place recognition. Ehsan *et al.* [6] proposed histogram of oriented uniform patterns (HOUP) descriptor which has the capability invariant to rotations.

In this paper we present a novel framework for scene recognition. From the research results [5], the CENTRIST compared with SIFT has the advantage that it tend to group image regions with similar visual structure into the same code word. However, currently almost all works divide the image into a regular grid of sub-regions which decrease the discriminative power of visual words. Therefore, we propose to use superpixel segmentation method [7] to divide images into meaningful sub-regions and then extract the CENTRIST features of sub-regions as visual descriptors. Next, the original features are converted to bag of words (BoW) features for classification. In addition, differences

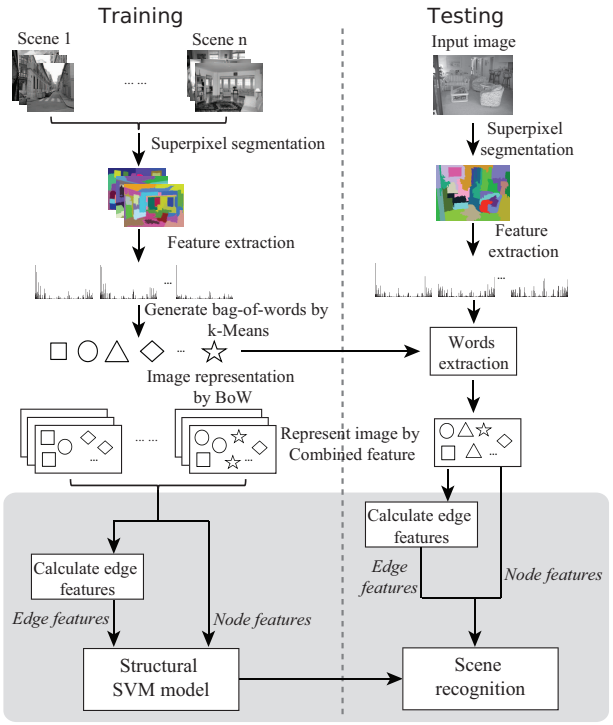

Figure 1: A schematically presentation of proposed framework. The left part represents the training process, and the right part is the test process.

between adjacent superpixels are also extracted as edge features which are adopted to construct a structural model. The structural model provides more semantic discriminability, therefore, they have better reorganization accuracy. We use UIUC dataset [2] to evaluate the proposed method, and it achieved the accuracy of 89.38%. In addition, we applied the proposed method to MIT 67-class scene dataset [8], and the average accuracy of scene reorganization is 48.3%.

2. STRUCTURAL MODEL BASED SCENE RECOGNITION

In this section, we first give a brief processing flow of the proposed method, and then details of each procedure are provided in subsections. Figure 1 shows processing flow of the proposed method for training and recognition. In the model training stage, the fist step is to segment the images into superpixels [7] which have more semantic discriminability. After that CENTRIST descriptor [5] of each superpixel is extracted, bag of features are generated for representing collection of semantic information of a scene. Finally, a structural model is constructed to obtain a discriminative classifier. In the recognition process, the superpixel segmentation and feature extraction are same as that in the training process. Bag of features are extracted using the trained data, and then structural SVM [9] is used to classify the scene category.

2.1 Superpixel Segmentation

In this study we divide the image into superpixels to represent meaningful image regions instead of regular grid based image partition. We have evaluated various superpixel segmentation methods, and found that entropy rate superpixel segmentation method [7] performs better than other methods, moreover it has a relatively fast processing speed.

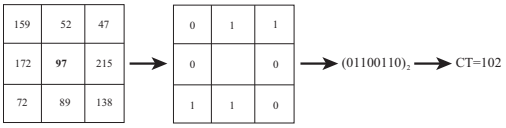

Figure 2: An example of census transform for a pixel which has intensity of 97.

2.2 Feature Extraction

In this study, we use CENRIST feature [5] for representing visual information of scenes, which has the following advantages: superior recognition performance on multiple standard data sets; no parameter to turn; extremely fast processing speed; easy to implement. The CENTRIST is one type of uniform pattern feature, and census transform (CT) [5] of each image pixel is calculated by subtracting the intensity at that pixel from each of its neighboring pixels and converting the results to a base-10 number which is illustrated in Fig. 2. After the CT value of each pixel is calculated, a histogram of CT values is obtained, and then PCA is adopted to reduce the dimension to $V = 42$. The compact descriptor is called as Pca Census Transform (PCT) descriptor.

In this study, the input image is segmented into meaningful S patches by superpixel method, and then each patch's feature is calculated by bag of words method. A bag of words model represents an image $\mathbf{x}$ by an unordered collection of visual words. In the model, a vocabulary $\mathbf{P} = \{\mathbf{p}_1, ..., \mathbf{p}_K\}$ with size K is a set of representative vectors in the descriptor space, obtained by means of unsupervised learning such as k-means. The second step is to quantize the descriptor space in order to obtain a compact representation in a vocabulary of visual words. For each image segment x_i with the descriptor $pct(x_i)$, we define the feature distribution $\theta(x_i) = (\theta_1(x_i), ..., \theta_K(x_i))^T$, a K dimensional vector, and each element is defined as

$$\theta_k(x_i) = ce^{-\frac{||pct(x_i)-\mathbf{p}_k||_2^2}{2\sigma^2}}, \quad (1)$$

where the constant c is selected in such a way that $||\theta_k(x_i)||_1 = 1$, and σ is set to be the maximum distance between any two visual words. In this work, the size of visual vocabulary is set to be K=200 in all of our experiments.

However only using the bag of features will lose the geometric layout of the scene. Thereby, in this study, we model the scene recognition using a model similar to graph matching with node and edge features. The first edge feature is histogram intersection [10], which has the definition of

$$\phi_{t1}(x_i, x_j) = \sum_{k=1}^{V} \min\{\theta_k(x_i), \theta_k(x_j)\}. \quad (2)$$

The second edge feature is the distance between corresponding visual words of adjacent superpixels x_i and x_j. It has the following definition

$$\phi_{t2}(x_i, x_j) = ||\mathbf{p}(x_i) - \mathbf{p}(x_j)||^2. \quad (3)$$

2.3 Structural Model for Scene Recognition

The discriminative approach such as SVM does not rely on explicit probabilistic models for the images in each class, as a consequence it has better performance. Given an image $\mathbf{x} = \{x_1, ..., x_S\}$ consisting of over-segmented image patches $\{x_i\}$, the purpose is to predict a category $\mathbf{y} = (y_1, ..., y_M)$ for the image. The scene category contains M binary values, with each $y_i \in \{0, 1\}$ indicating whether the image belongs to the category i. For an input image $\mathbf{x}$, the prediction $\mathbf{y}^*$ is computed as the argmax of a discriminant function $f_w(\mathbf{x}, \mathbf{y})$ that is parameterized by a vector of weight $\mathbf{w}$.

$$\mathbf{y}^* = \underset{\mathbf{y}}{\operatorname{argmax}} f_w(\mathbf{x}, \mathbf{y}). \tag{4}$$

We model the features by an undirected graph $(\mathcal{V}, \mathcal{E})$, where vertices $\mathcal{V} = \{\theta(x_1), ..., \theta(x_S)\}$ and edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, where S is the superpixel number of the image. This undirected graph is derived from the image by adding a vertex for each superpixel and adding an edge between vertices base on the spatial proximity of the corresponding superpixels. Given $(\mathcal{V}, \mathcal{E})$, the following discriminant function based on node features $\phi_n(x_i) = \theta(x_i)$ and edge features $\phi_t(x_i, x_j)$ is defined as

$$f_{\mathbf{w}}(\mathbf{x}, \mathbf{y}) = \sum_{i \in \mathcal{V}} \sum_{m=1}^{M} y_m [w_n^m \cdot \phi_n(i)] + \sum_{(i,j) \in \mathcal{E}} \sum_{m=1}^{M} y_m [w_t^{i,j} \cdot \phi_t(i, j)]. \tag{5}$$

To simplify the notation, note that Eq. (5) can be equivalently written as $\mathbf{w}^T \Phi(\mathbf{x}, \mathbf{y})$ by appropriately stacking the w_n^m and $w_t^{i,j}$ into $\mathbf{w}$ and the $y^m \phi_n(i)$ and $y^m \phi_t(i, j)$ into $\Phi(\mathbf{x}, \mathbf{y})$. Training can be formulated as the one slack formulation

$$\begin{aligned} min_{w,\xi} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\xi, \\ s.t. \quad & \frac{1}{N}\sum_{i=1}^{N} [\Phi(\mathbf{x}_i, \mathbf{y}_i) - \Phi(\mathbf{x}_i, \overline{\mathbf{y}}_i)] \\ & \geq \frac{1}{N}\sum_{i=1}^{N} \Delta(\mathbf{y}_i, \overline{\mathbf{y}}_i) - \xi, \end{aligned} \tag{6}$$

where $\Delta(\mathbf{y}_i, \overline{\mathbf{y}}_i) = |\mathbf{y}_i - \overline{\mathbf{y}}_i|$ is the zero-one loss function quantifying the deviation between estimated and true scene category, C is a scaling constant, and N is the number of training images. This problem can be solved efficiently using the cutting-plane algorithm [9] for training the structural model.

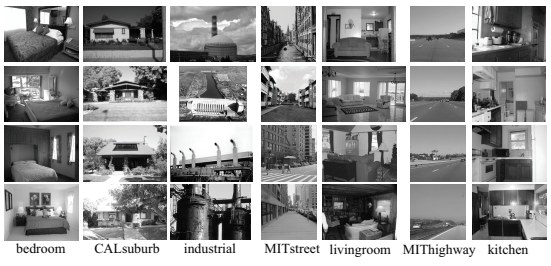

Figure 3: Sample images of the UICU scene dataset.

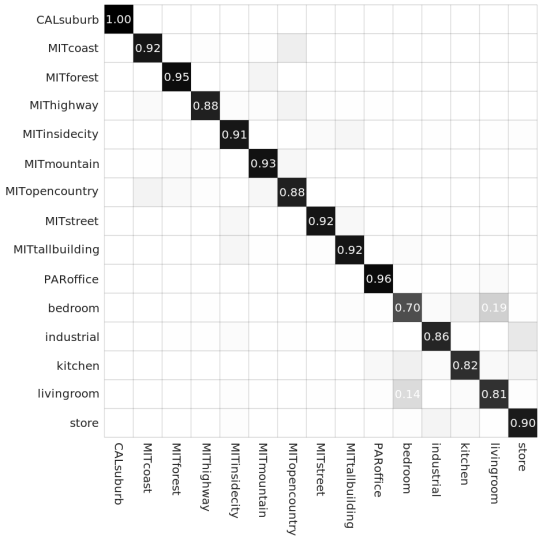

Figure 4: Confusion matrix from one run on UIUC 15-category scene dataset recognition.

3. EXPERIMENTS

To evaluate the proposed method, in the first experiment we use the public available UIUC dataset [2] which contains 4,485 images of 15 different categories. It includes both indoor scenes (office, bedroom, kitchen, living-rooming, store) and outdoor scenes (suburb, coast, forest, highway, inside-city, mountain, open-country, street, tall-building, industrial). Figure 3 shows some of the images from the dataset.

All processing in our experiments was performed in grayscale. The color information was discarded if it is available. In the experiment, the input images were segmented by entropy rate superpixel segmentation method [7] with segment number of $S = 30$. Then each segment's CENTRIST descriptor was calculated to define visual words. The visual vocabulary was created using k-means clustering on all segments of training images. The size of visual vocabulary was set to be K=200. The superpixel segmentation is most time-consuming step in the training, which need about 1 3 seconds. The classification time for a 400x300 image takes about 0.6 seconds.

Figure 4 presents the confusion matrix from one run on UIUC scene dataset, where row and column names are true and classified scene categories, respectively. From the figure, we observed that the highest recognition rate is 100% for 'CALsuburb'. The biggest confusion happens between 'bedroom' and 'livingroom', due to their similar global structure and spatial layout. Averaged recognition accuracy of stat-of-art methods are listed in Table I. Due to the Gist descriptor [6] has poor discriminative power for indoor environments, it output lowest accuracy. Spatial pyramid matching (SPM) method [2] incorporates spatial information, and consequently it outputs better results. The classification result which only uses CENTRIST [5] is 74.2%. Although it not good as SPM, it is still better than Gist. When use PCA to reduce the dimensionality of CENTRIST to 40 (which is called PACT), and uses a level 2 pyramid division [5], the accuracy improved to 83.8%. The proposed method outputs **89.38**% accuracy.

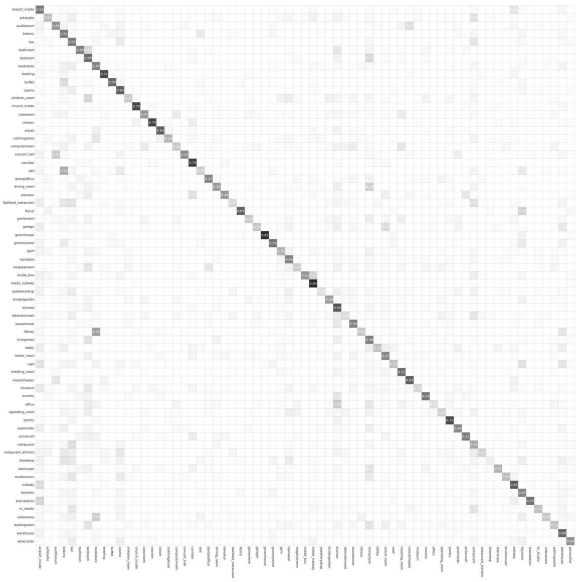

Figure 5: Confusion matrix from one run on MIT 67-class scene dataset.

Method	Accuracy
Gist [1]	65.8%
Spatial Pyramid Matching [2]	81.4 ± 0.50%
CENTRIST [5]	73.29 ± 0.96%
CENTRIST + PACT [5]	83.88 ± 0.76%
Proposed method	**89.38** ± 0.70%

Table 1: Performance comparison on 15 scenes dataset.

In the second experiment, we evaluated the proposed method on a challenging 67-class indoor scene recognition dataset [8]. There are 15,620 images in the dataset. The indoor scenes range from specific categories (e.g., dental office) to generic concepts (e.g., mall). In research [8], the global Gist feature achieved about 21% average recognition accuracy on this challenging dataset. When it was supplemented by local information (in the form of local prototypes based on image segmentation), the accuracy was improved to 25%. By using the PACT descriptor and RBF SVM [5], the average accuracy in five random split of train/test images is 36.9%. By our proposed method, the average accuracy also in five cross-validations improved to **48.3**%, which archives higher accuracy than other methods. Confusion matrix from one run is given in Fig. 5.

4. CONCLUSIONS

In this paper, we presented a novel scene recognition method. Our major contribution is that input image is first segmented into meaningful sub-regions for generating visual words which have higher discriminability. In addition, we constructed a structural model to train and learn scene categories, which incorporates spatial layout to represent the scene image. The UIUC scene dataset is used to evaluate our proposed method compared with other methods. From the results, we can conclude that proposed method has better performance. In addition, the proposed method was also evaluated on MIT 67-class scene dataset. We found that the proposed method has 48.3% average accuracy which is much higher than other methods. Although our proposed method has better performance, there also exist some limitations. First, compared with CENTRIST descriptor and SVM based scene recognition method, our method needs more computation time due to the superpixel segmentation. Second, currently we didn't consider temporal consistency. In the following work, we will improve the classification model which will has higher processing speed and modeling temporal consistency.

5. ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their valuable feedback. This work is supported partly by grants from NSFC (61003137, 61202185), NWPU Basic Research Fund (JC201202, JC201220), Shaanxi Natural Science Fund (2012JQ8037), and Open Project Program of the State Key Lab of CAD&CG (A1306), Zhejiang University.

6. REFERENCES

[1] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[2] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178. IEEE, 2006.

[3] Yangqing Jia, Chang Huang, and Trevor Darrell. Beyond spatial pyramids: Receptive field learning for pooled image features. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3370–3377. IEEE, 2012.

[4] Olga Russakovsky, Yuanqing Lin, Kai Yu, and Li Fei-Fei. Object-centric spatial pooling for image classification. In *Computer Vision–ECCV 2012*, pages 1–15. Springer, 2012.

[5] Jianxin Wu and Jim M Rehg. Centrist: A visual descriptor for scene categorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(8):1489–1501, 2011.

[6] Ehsan Fazl-Ersi and John K Tsotsos. Histogram of oriented uniform patterns for robust place recognition and categorization. *The International Journal of Robotics Research*, 31(4):468–483, 2012.

[7] Ming-Yu Liu, Oncel Tuzel, Srikumar Ramalingam, and Rama Chellappa. Entropy rate superpixel segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2097–2104. IEEE, 2011.

[8] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. 2009.

[9] Thorsten Joachims, Thomas Finley, and Chun-Nam John Yu. Cutting-plane training of structural svms. *Machine Learning*, 77(1):27–59, 2009.

[10] Michael J Swain and Dana H Ballard. Color indexing. *International journal of computer vision*, 7(1):11–32, 1991.