2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Outdoor Scene Understanding Using SEVI-BOVW Model

Haibing Zhang, Shirong Liu and Chaoliang Zhong

Abstract— **A simple and effective novel approach for scene understanding is addressed in this paper. Based on bag of visual words (BOVW) model, explicit semantics associated with the object image was embedded into visual words, and then various types of visual words integrated, and finally the SEVI-BOVW (semantics embedded and vocabulary integrated bag of visual words) model constructed. Mean Shift algorithm was employed to recognize local region image in scene. Compared with image understanding approaches presented in the literature, the proposed approach here can remove a classification or generative model during model training or testing. Objects category recognition can be determined by the number of class-specific semantic visual words, without complex reasoning. The effectiveness of the proposed approach has been demonstrated by the experimental results of scene understanding in a campus.**

I. Introduction

As low-level image recognition based on computer vision reaches maturity, more and more researchers have started considering high-level scene understanding tasks [1]-[3]. High-level scene understanding contains local and global understanding [4]: the former focuses on analyzing and describing the distributions and categories of the local regions in the image, such as different categories of local objects recognition and localization; the latter focuses on understanding the global properties of a scene, such as scene classification and recognition.

Many methods to local scene understanding have been proposed, such as estimating labels pixel by pixel [5]-[7], aggregating features over segmentation regions [8]-[11], and predicting object bounding boxes [12][13]. Most of these approaches require a discriminative or generative model to be trained in advance for each class. In most cases, processing a test image is also quite slow, as it involves steps like running multiple object detectors over the image, performing graphical model inference, or searching over multiple segmentations [3]. Recently, a few researchers have begun advocating nonparametric, data-driven approaches suitable for open-universe datasets. Such approaches do not do any training at all. Instead, for each new test image, they try to retrieve the most similar training images and transfer the desired information from the training images to the query. Liu et al. [14] proposed a non-parametric label transfer method based on estimating SIFT flow or a dense deformation field between images. The biggest drawback of this method is that the optimization problem for finding the SIFT flow is fairly complex and expensive to solve. Tighe et al. [3] implemented a nonparametric solution to image parsing that is as straightforward and efficient as possible, and that relies only on operations that can easily scale to even larger image collections and sets of labels.

The above methods can achieve better performance of scene understanding, but because of training a model, image's content reasoning processes are complicated that result in a large amount of calculations. Inspired by the above methods, we present a simple and effective approach without complex reasoning. The proposed method is outlined in Fig. 1. It consists of SEVI-BOVW model construction and local image region recognition.

The proposed approach for scene understanding can be used in outdoor mobile robot navigation. Massive image data can be transformed into simple knowledge in line with the concept of human cognition, which is helpful to realize the navigation and feasible region detection tasks. Besides, it is convenient for human-machine interaction, knowledge management and storage.

II. Method

This section presents our outdoor scene understanding method as illustrated in Fig. 1. Sections 2.1 and section 2.2 describe SEVI-BOVW model construction and local image region recognition, respectively.

A. SEVI-BOVW Model

The core part of SEVI-BOVW model is visual vocabulary. The so-called "visual words" in traditional visual vocabulary do not have explicit semantics, which essence is the image element, and the understanding performance of the scene is limited. We propose a semantic embedded and integrated visual vocabulary. Every visual word in the visual vocabulary is given an explicit semantic. Then object recognition becomes easier based on the SEVI-BOVW model.

a. SIFT feature extraction

First step to construct the SEVI-BOVW model is the local feature extraction, and we adopt the popular SIFT feature. It is necessary to determine the key local region of image, the interest point detection method to determine the local region sparsely distributed will lose a lot of information. Therefore, dense SIFT features are extracted after evenly image division in this paper, which can retain the maximum extent of image information. As it is shown in Fig. 2, the input image is divided into multiple image blocks of 16×16 pixels, and interval between image blocks is 8 pixels, then SIFT features are extracted from each image block.

b. Visual vocabulary construction

Haibing Zhang and Shirong Liu are with the Institute of Electrical Engineering and Automation, Hangzhou Dianzi University, Hangzhou (email: zhb20110416@126.com, liushirong@hdu.edu.cn, chaoliangzhong@163.com); Chaoliang Zhong is with the Institute of Automation of East China University of Science and Technology, Shanghai (email: chaoliangzhong@163.com).

This work was supported by National Natural Science Foundation of China (No.61175093)

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

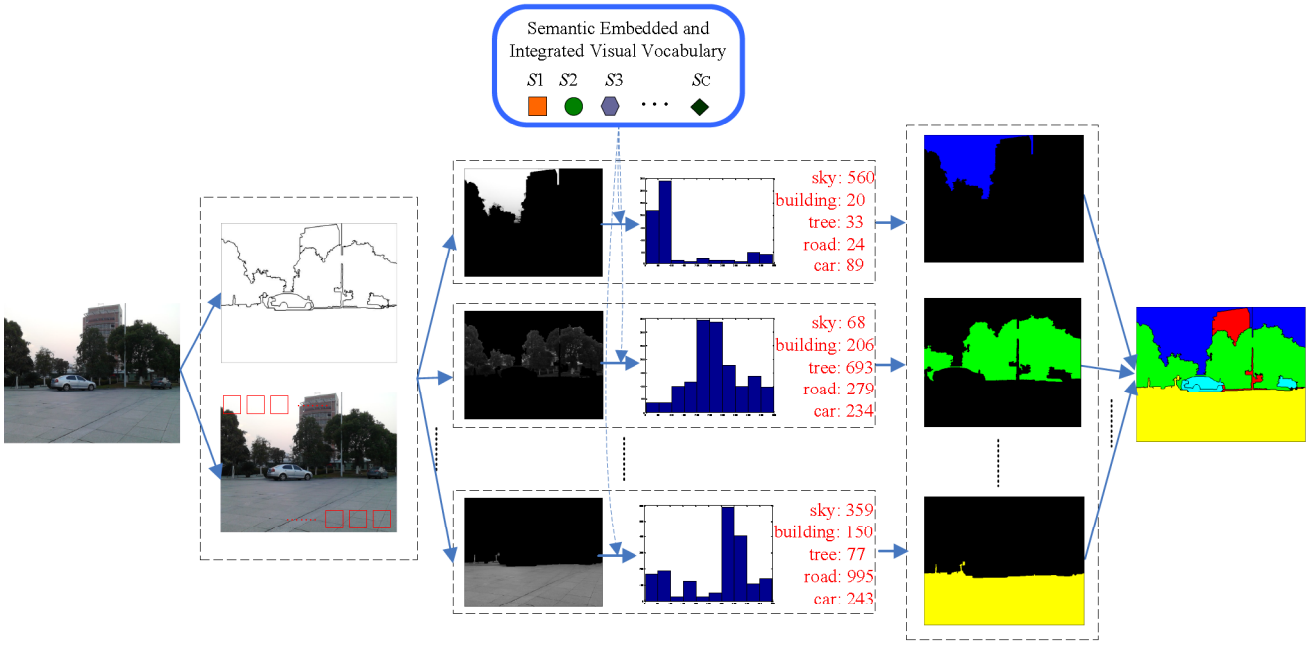

Fig. 1. Overview and sample result of our approach. In the training phase, a semantic embedded and integrated vocabulary is constructed. In the testing phase, given an input image, local regions are obtained by image segmentation based on the mean shift algorithm. To extract the SIFT feature sets of each local region, region histogram representation is formed based on semantic embedded and integrated vocabulary. According to the maximum number of each kind of semantic visual words, the semantic category of the region can be determined.

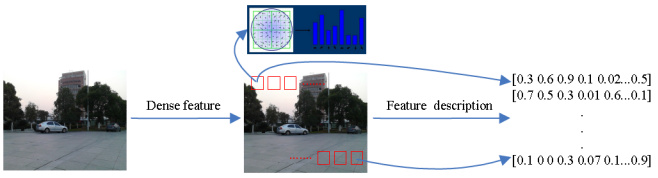

Fig. 2. Dense SIFT feature extraction. We take each image block center as the sampling point, the image block is evenly divided into 4×4 sub regions, and gradient direction histograms are calculated in eight directions in the sub region image of 4×4 pixels. Draw the accumulative value of each gradient direction, so that it can form a seed point. All sub regions of image block generate 4×4=16 seed points, so that each image block can be described by the 16×8=128-d SIFT feature vector.

Universal visual vocabulary is obtained by training on all the local features set of object categories. However, the universal visual vocabulary has follow defects: 1) when there are a large number of features, the clustering algorithm (such as k-means clustering) result in a large burden to cause low computational efficiency; 2) clustering algorithm on the entire feature set clustering, resulting in different categories to be confused with similar features, so the lower distinction between the generated visual words.

The rationale behind building an integrated visual vocabulary is try to find more specific discriminative visual words from each object class in order to avoid interference with other classes. Before the integration of each type of visual vocabulary, the explicit semantic are embedded as it is shown in Fig. 3. Step of semantic embedded and integrated visual vocabulary construction is as follows:

1) If there are C class objects, respectively for all kinds of SIFT feature sets using k-means clustering algorithm training. And the SIFT feature dimension is 128-d.

2) Each cluster center is regarded as a visual word, to generate the C types of visual vocabularies $V_i = \{v_{i1}, v_{i2}, ..., v_{ik}\}$, where $i = 1, 2, ..., C$, k refers to the number of visual words in the i-th category of visual vocabulary.

3) According to the prior knowledge, the C class objects have a total of C semantics. Corresponding semantic labels are embedded to each type of visual dictionary:

$$V_i^{(S_i)} = \{V_{i1}^{(S_i)}, V_{i2}^{(S_i)}, ..., V_{iC}^{(S_i)}\} \quad (1)$$

where S_i is semantic label of the i-th category of visual vocabulary, such as "building", "car", "road" and so on.

4) Finally, we integrate the C categories of visual vocabularies that are embedded semantics:

$$\bar{V}^{(S)} = \{V_1^{(S_1)}, V_2^{(S_2)}, ..., V_C^{(S_C)}\} \quad (2)$$

B. Content Representation and Recognition of the Local Region

Given a test scene image, it contains multiple spatial objects. The research [8]-[11] have shown that, while the image is divided into a number of local regions, the scene understanding problem is simplified to identify each local region. We adopt the mean shift algorithm for test image segmentation in this paper, and recognize each local region.

a. Image segmentation based on mean shift algorithm

Mean shift algorithm [15][16] is a nonparametric method based on the density gradient ascent. Comaniciu et al. [17] applied the mean shift algorithm to the feature space analysis by, and achieved good preference on image smoothing and segmentation.

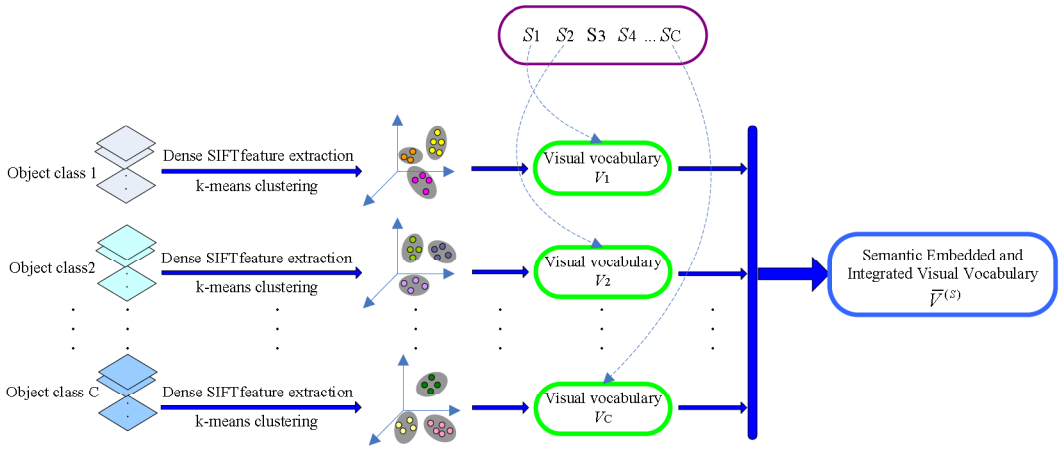

Fig. 3. Visual vocabulary building process. Each class feature represents feature vectors of training images for a specific image category. Before the integration of each type of visual vocabulary, the explicit semantic (such as "sky", "road" and so on) are embedded.

Mean Shift algorithm is applied to image segmentation will be able to get a high-quality edge image. Given an RGB image, first convert it to Luv color space, and in the space of random position set a certain number of search windows. Then for each window using the Mean Shift algorithm to search the high density areas, and record the central position to obtain a feature palette. According to a set of related parameters, feature vectors in feature space are assigned to the feature palette, finally to obtain segmented regions:

$$R = \{r_n | n = 1, 2, ..., N\} \tag{3}$$

where N refers to the number of local regions after image segmentation, r_n represents the n-th local region.

b. Local region recognition

After image segmentation, semantic annotation should be carried out for each region. Each region contains a kind of main object that has to identify and few other types of disturbance.When objects of each class belong to regions are of low similarity, a simple method for object recognition is proposed based on the semantic embedded and integrated visual vocabulary, which without the classifier or generated model training and classification.

After constructing the semantic embedded and integrated visual vocabulary and completing test image mean shift segmentation, steps of image region identification are as follows:

1) The Eqn. 2 can be written as:

$$\bar{V}^{(S)} = \{v_{c1}^{(S_c)}, v_{c2}^{(S_2)}, ..., v_{cl^{(c)}}^{(S_c)}\}_{c=1}^{C} \tag{4}$$

where C is the total number of object categories; $l^{(c)} = 1, 2, ..., |V_c^{(S_c)}|$, $|V_c^{(S_c)}|$ represents the length of visual vocabulary of the c-th class which is embedded the corresponding semantic S_c; $v_{cl^{(c)}}^{(S_c)}$ is the $l^{(c)}$-th visual word of the c-th class visual vocabulary.

2) Extract dense SIFT features, and each local region contains a feature set:

$$f_n = \{u_{n1}, u_{n2}, ..., u_{nm}\} \tag{5}$$

where $n = 1, 2, ..., N$, N represents the total number of regions; m refers to the number of features that belong to the region r_n; u_{nm} refers to the m-th feature of n-th local region.

3) Features in the region r_n are mapped to the semantic embedded and integrated visual vocabulary to obtain the histogram representation:

$$h_{il^{(c)}}^{(S_c)}(r_n) = \sum_{j=1}^{m} f_{r_n^{(q)}}^{(i)}, i = 1, 2, ..., |\bar{V}^{(S)}| \tag{6}$$

$$f_{r_n^{(q)}}^{(i)} = \begin{cases} 1, & ||u_{nj} - v_{il^{(c)}}^{(S_c)}|| \leq ||u_{nj} - v_{ql^{(c)}}^{(S_c)}||, \\ & q = 1, 2, ..., |\bar{V}^{(S)}| and\ i \neq q \\ 0, & otherwise \end{cases} \tag{7}$$

where $h_{il^{(c)}}^{(S_c)}(r_n)$ refers to the feature number of nearest Euclidean of features in region r_n and visual word $v_{il^{(c)}}^{(S_c)}$; If the Euclidean of j-th feature u_{nj} and visual word $v_{il^{(c)}}^{(S_c)}$ is nearest, $f_{r_n^{(q)}}^{(i)} = 1$, else $f_{r_n^{(q)}}^{(i)} = 0$.

4) For each region, its histogram representation can be represented as:

$$\begin{aligned} h^{(S)}(r_n) = \{&h_{11}^{(S_1)}(r_n), h_{12}^{(S_1)}(r_n), ..., h_{1l^{(1)}}^{(S_1)}(r_n), \\ &h_{21}^{(S_2)}(r_n), h_{22}^{(S_2)}(r_n), ..., h_{2l^{(2)}}^{(S_2)}(r_n), \\ &\vdots \\ &h_{c1}^{(S_c)}(r_n), h_{c2}^{(S_c)}(r_n), ..., h_{cl^{(c)}}^{(S_c)}(r_n)\} \end{aligned} \tag{8}$$

where $h_{cl^{(c)}}^{(S_c)}(r_n)$ refers to the frequency of $l^{(c)}$-th visual word of the c-th class visual vocabulary in the region r_n.

After getting the histogram representation of the region r_n, content of the region r_n can be identified. According to the maximum number of each kind of semantic visual words, the semantic category of the region r_n can be determined as is shown in formula (9)-(11):

$$S_n = S_\gamma \tag{9}$$

where

$$\gamma = \arg\max N_\gamma(r_n) \tag{10}$$

$$N_c(r_n) = \sum_{l^{(c)}=1}^{|V_c^{(S_c)}|} h_{cl^{(c)}}^{(S_c)}(r_n) \tag{11}$$

where $\gamma = c = 1, 2, ..., C$, $N_c(r_n)$ refers to the number of visual words of semantic S_c. Visible to above, the proposed method does not need other classification algorithms to perform complex reasoning, and it costs a small amount of calculation.

III. Experiment Result and Discussion

Image data is acquired in the campus environment, and recognize the campus scenes such as sky, tree, building, road and car, a total of 5 objects. The training samples are local regions that intercepted from campus scene images and there is no fixed size. And the training samples number of each type is 40. Fig. 4 shows 5 samples from each training class.

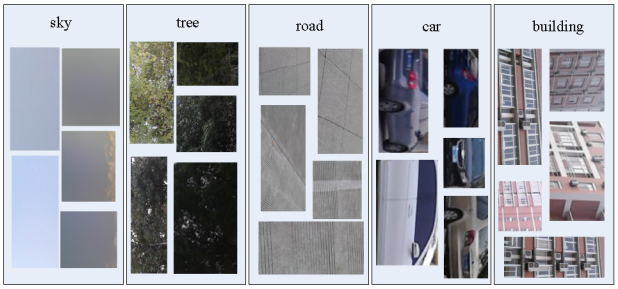

Fig. 4. Example of training samples. Every "local-object" (local region) is intercepted from "intact-object", and each local region contains a kind of main object that has to identify and few other types of disturbances.

Fig. 5 shows the content recognition results of 4 campus scene images (size of each image is 640×480 pixels). Where the first column shows the original 4 images, the second column is the edge image after segmentation using mean shift algorithm, and the next five columns are objects' recognition results that adopt visual vocabularies of different length k. Image segmentation result and length of visual vocabulary have important influence on the result of the object recognition, we only study the latter. Namely, under the condition that the image segmentation result has been established, we study the effect of different visual vocabulary lengths on the recognition results. The 4 image content recognition results show that recognition results are the worst when $k = 40$, as k increases, the image content recognition performance is improved. That is when the number of visual words is too small, the different semantic concept image blocks can be marked as similar visual words, which make the recognition performance is relatively low. As the number of visual word increases, visual words discrimination between classes increased to a maximum value, and it will not be an obvious change later.

As described in section 2.2.1, feature extraction is a method that adopts evenly image partition, and the size of image block is 16×16 pixels, interval between image blocks is 8 pixels. After the image segmentation, there will be some small local regions, when fewer sampling points of feature fall within the small regions and the surrounding regions are different categories, it is easier to make the wrong decision. If there are no sampling points fall within the region, the region has not been identified, so the image recognition results on the presence of "black regions" represented by the unknown regions. Because of the unknown regions are so small, the impact of the "black regions" understanding results for the entire scene can be ignored.

IV. Conclusions

Based on the SEVI-BOVW model, the proposed approach for scene understanding has the following characteristics: 1) each visual word in the integrated visual vocabulary is given an explicit semantic according to the information of categories, so the image primitives are risen to the specific semantic concepts that help further image content recognition; 2) compare with the traditional method, reasoning process of image content is simplified based on the semantic embedded and integrated visual vocabulary, and it can obtain a good understanding performance. This approach can be expanded for the outdoor environment perception and autonomous navigation for mobile robots.

References

[1] M.P. Kuma, D. Koller, "Efficiently Selecting Regions for Scene Understanding," *In: Proceedings of 2010 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, San Francisco, CA, pp. 3217-3224, 2010.

[2] J. Tighe, S. Lazebnik, "Finding Things: Image Parsing with Regions and Per-Exemplar Detectors," *n: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition,CVPR*, Portland, OR, pp. 3001-3008, 2013.

[3] J. Tighe, S. Lazebnik, "Superparsing: Scalable Nonparametric Image Parsing with Superpixels," *International Journal of Computer Vision, IJCV*, vol. 101, no. 2, pp. 329-349, 2013.

[4] A. Bosch, X. Munoz, A. Oliver, R. Marti, "Object and Scene Classification: what does a Supervised Approach Provide us?" *In: Proceedings of the 18th International Conference on Pattern Recognition, ICPR*, Hong Kong, China, vol. 1, pp. 773-777, 2000.

[5] H. Xuming, R.S. Zemel, M.A. Carreira-Perpinan, "Multiscale Conditional Random Fields for Image Labeling," *In: Proceedings of 2004 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, vol. 2, pp. II-695-II-702, 2004.

[6] L. Ladicky, P. Sturgess, K. Alahari, C. Russell, P.H.S Torr, "What, Where and How Many? Combining Object Detectors and CRFs," *In: Proceedings of European Conference Computer Vision, ECCV*, vol. 6314, pp. 424-437, 2010.

[7] J. Shotton, M. Johnson, R. Cipolla, "Semantic Texton Forests for Image Categorization and Segmentation," *In: Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, Anchorage, AK, pp. 1-8, 2008.

[8] C. Galleguillos, B. Mcfee, S. Belongie, G. Lanckriet, "Multi-Class Object Localization by Combining Local Contextual Interactions," *In: Proceedings of 2010 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, San Francisco, CA, pp. 113-120, 2010.

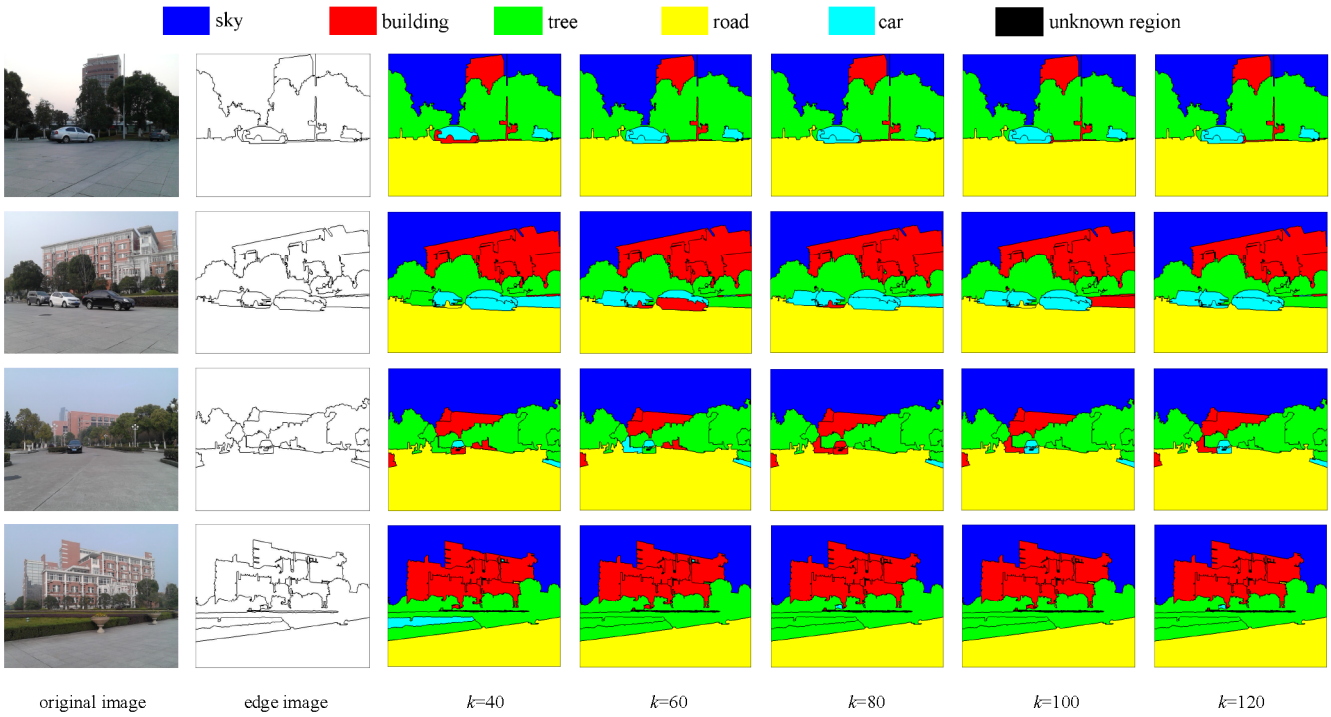

Fig. 5. Results of campus scene understanding. Each image of the first three scene images contains five categories of objects, and the fourth scene image contains four categories of objects. We obtain the best understanding performance of the campus scene when k increases to 100.

[9] S. Gould, R. Fulton, D. Koller, "Decomposing a Scene into Geometric and Semantically Consistent Regions," *In: Proceedings of IEEE International Conference on Computer Vision, ICCV*, Kyoto, Japan, pp. 1-8, 2009.

[10] Z. Min, L. Lixiong, J. Yunde, "An Outdoor Scene Understanding Method Based on Ensemble Classification of Image Regions," *Journal of Image and Graphics*, vol. 9, no. 12, pp. 1143-1148, 2004.

[11] D. Hoiem, A.A. Efros, M. Hebert, "Recovering Surface Layout from an Image," *International Journal of Computer Vision, IJCV*, vol. 75, no. 1, pp. 151-172, 2007.

[12] S. Divvala, D. Hoiem, J. Hays, A. Efros, M. Hebert, "An empirical study of context in object detection," *In: Proceedings of 2009 IEEE Conference Computer Vision and Pat tern Recognition, CVPR*, Miami, FL, pp. 1271-1278, 2009.

[13] G. Heitz, D. Koller, "Learning Spatial Context: Using Stuff to Find Things," *In: Proceedings of European Conference Computer Vision, ECCV*, vol. 5302, pp. 30-43, 2008.

[14] L. Ce , J. Yuen, A. Torralba, "Nonparametric Scene Parsing via Label Transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2368-2382, 2011.

[15] J.J. Hopfield and D.W. Tank, "Computing with neural circuits: A model," *Science*, vol. 233, pp. 625-633, 1986.

[16] L. Xiangru, W. Fuchao, H. Zhanyi, " Convergence of a Mean Shift Algorithm," *Journal of Software*, vol. 16, no. 3, pp. 265-374, 2005.

[17] D. Comaniciu, M. Peter, "Mean Shift: A Robust Approach Toward Feature Space Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603-619, 2002.