Road Image Segmentation and Recognition Using Hierarchical Bag-of-Textons Method

Yousun Kang[1], Koichiro Yamaguchi[2], Takashi Naito[2], and Yoshiki Ninomiya[2]

[1] Tokyo Polytechnic University
yskang@cs.t-kougei.ac.jp
[2] Toyota Central R&D Labs., Inc.
{yamaguchi,naito,ninomiya}@mosk.tytlabs.co.jp

Abstract. While the bag-of-words models are popular and powerful method for generic object recognition, they discard the context information for spatial layout. This paper presents a novel method for road image segmentation and recognition using a hierarchical bag-of-textons method. The histograms of extracted textons are concatenated to regions of interest with multi-scale regular grid windows. This method can learn automatically spatial layout and relative positions between objects in a road image. Experimental results show that the proposed hierarchical bag-of-textons method can effectively classify not only the texture-based objects, e.g. road, sky, sidewalk, building, but also shape-based objects, e.g. car, lane, of a road image comparing the conventional bag-of-textons methods for object recognition. In the future, the proposed system can combine with a road scene understanding system for vehicle environment perception.

Keywords: road image segmentation, hierarchical bag-of-textons, multi-scale.

1 Introduction

Intelligent Transport Systems (ITS) have developed significantly in the last few decades, and vehicle safety has been a particularly active research area [1]. The latest driving assistance systems include many vision-based applications such as lane detection, road detection, and pedestrian detection, which provide drivers with useful information [2]. Current vision-based intelligent vehicles are mostly focused on the detection of obstacles such as cars, bicyclists, and pedestrians.

However, an advanced driving assistance system may in the future be focused on *analysis* or *understanding* of a road scene than the detection of obstacles in road image. The scene understanding system requires integrated and/or advanced vision procedures, which are particularly relevant to image classification, object detection, and semantic segmentation. Among of them, semantic segmentation is a more complete image understanding system.

The role of semantic segmentation is central to visual interpretation and understanding to improve the effectiveness for vehicle environment perception. For example, by segmenting a road image, we can detect hazards or blind spots on the road. Such detection should consider the difference in potential risk for pedestrians standing on the

Y.-S. Ho (Ed.): PSIVT 2011, Part I, LNCS 7087, pp. 248–256, 2011.

road, on a crosswalk, or on the sidewalk. The probability of collision with a particular obstacle and the potential risk associated with a covert hazard can be estimated by segmenting a road image.

Therefore, automatically classifying pixels and parting meaningful regions in a road image is particularly helpful instance in vehicle safety field. This process is referred to as image labeling procedure, since its goal is to associate each pixel in the image with a label denoting a semantically meaningful part. In this paper, we investigate the problem of achieving recognition and segmentation of object classes in road image using hierarchical bag-of-textons method.

The bag-of-features method is one of the most popular and efficient for object recognition and image segmentation. It considers an object in an image as a set of unordered features extracted from local patches. The features are quantized into discrete visual words, with sets of all visual words referred to as a dictionary. Among various features, textons are representative dense visual words and they have been proven effective in categorizing materials as well as generic object classes [3-5]. In addition, textons are utilized in both object segmentation and recognition thanks to their high density and efficient [6].

However, the major drawback of the bag-of-features model is that it discards the spatial layout of visual words, which causes a serious problem for segmentation and recognition. In order to overcome the drawback, many researchers devote to develop the extension of the bag-of-feature model. Lazebnik *et al.* [7] proposed spatial pyramid matching (SPM) that utilizes the aggregated statistics of the local features on fixed sub-regions. SPM embeds a part of the spatial information over the whole image by partitioning an image into a sequence of sub-regions, so that they showed the good performance in scene categorization and object recognition.

In this paper, we propose a hierarchical bag-of-textons method that uses pairs of regular grid windows and neighborhood textons combined with multiple resolutions. Some objects in a road scene have a particular relation with other objects, e.g., cars are on the road, the road is below the sky, lanes surround the road, and so on. It is important to learn spatial layout and relative position between objects from the surrounding image. We uses a sequence of multi-scale grids and then computes a bag-of-textons histogram for each sub-region with different scale. Thus, the representation of the an object is the concatenation vector of all the histograms.

To classify the features of multi-class objects based on localized frequency of textons, we employ the Joint Boosting algorithm. We evaluate on our datasets including the variety road environment scenes and objects, e.g., road, tree, lane, sky, pole, sidewalk, car, and building. To assess how much a hierarchical bag-of-textons method helps with image segmentation and object recognition in road images, we have compared the recognition accuracy of conventional bag-of-textons methods. The experimental results show the proposed method improves the segmentation and recognition accuracy compared with the conventional bag-of-textons methods. As future work, we are interested in integrating the system into motion and semantic segmentation for vehicle environment perception. The proposed method can expand into a dynamic 3D scene analysis system in the near future.

The paper is organized as follows: Section 2 explain the filter bank and textonization process for road input image. Section 3 describes the feature extraction module for the hierarchical bag-of-textons method and the boosted classifier. Experimental results on performance and our conclusions are presented in the final two sections, respectively.

2 Textonization Process

In driving assistance system, infrared images are particularly useful for pedestrian detection of night vision systems and driver monitoring. The infrared is divided into near, far, and mid infrared, however, in this paper, we referred only to the near infrared. Near infrared is defined by water absorption, and the effect is formed by strongly reflecting off a person's exterior layer, and foliage, such as tree leaves and grass. Thus, in this paper, the near infrared and color image are available for road image segmentation and recognition. Input image has four bands consisting of a band of near infrared and three bands of color.

Convolving the four band image with a bank of linear spatial filters provides a good local descriptor of image patches and an effective statistical representation. Textons are typically a compact representation of filter bank responses for texture classification [9], image segmentation [10], and generic object recognition [11]. Kang *et al.* [8] compared the performance of various filter banks for the multiband image segmentation. Among the various filter banks, the 17-D set, which is proposed by Winn *et al.* [11], led to the best performance. The 17-D set consists of three Gaussians, four Laplacian of Gaussians (LoG), and four first-order derivatives of Gaussians. In order to implement the convolution of four bands image, we increase filter responses by adding the infrared intensity as a color intensity. We utilized the CIE Lab color space for three color bands. Fig. 1 shows how to expand the feature vectors of the 17-D set to 20-D set for a multiband image. The multiband images are convolved with a 20-D filter bank, and the cluster centers of the 20-D filter responses are utilized to generate image textons.

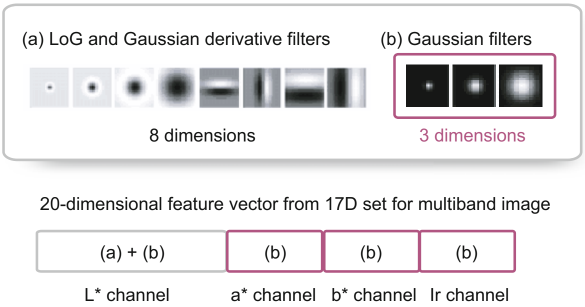


Fig. 1. The 20-D filter bank for a multiband image. The 17-D set filter banks are expanded to 20-D set for a multiband image.

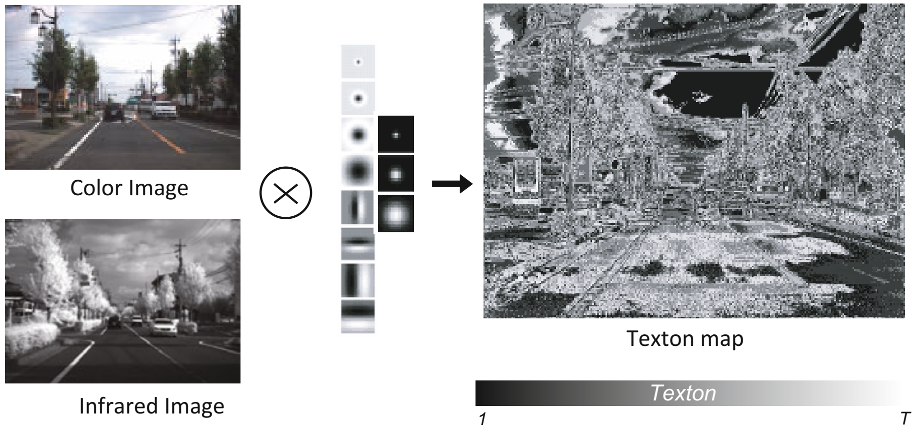


Fig. 2. Textonization image using 20D filter bank. Textons are represented by grayscale from 1 to T.

The road images are convolved with a 20-D filter bank and 20-D responses for all training pixels that are whitened to give zero mean and unit covariance. The K-means clustering is performed to quantize 20-D filter bank responses using a kd-tree algorithm [12]. We accomplished the textonization process using the code of *Calssification.NET* and *TextonBoost* [13] implemented by Shotton *et al.* [14]. Finally, each pixel in each image is textonized in the nearest cluster center, producing the texton map. Fig. 2 shows the texton map which is extracted from color and near-infrared image.

The filter responses are aggregated in the entire training set independently from class labels and clustered using K-means method to generate textons, which represent the visual words in a codebook of images. When a histogram of textons is created over a region of interest, we concatenate the histograms by using regular grids with multi-scale so as to learn automatically spatial relationship.

3 Hierarchical Bag-of-Textons

The bag-of-words models treat an object class as an unordered collection of visual words, sampling a representative set of image patches. However, it is important to extract the spatial configuration of an object and the contextual information from the surrounding image. It allows categorization and image segmentation algorithm to improve the performance by considering the context information of spatial layout. In road environment scene, there are spatial ordering constraints such as a car above a road and lanes are surrounded with road. It is necessary to order structural information between objects from the surrounding image. Therefore, we proposed a hierarchical bag-of-textons method using a spatial layout filter with multi-scale. The spatial layout filter with multi-scale is a pair (R, T) of a pyramid grid window R, and neighbor textons T. Our technique based on a bag-of-textons is capable of coping with spatial

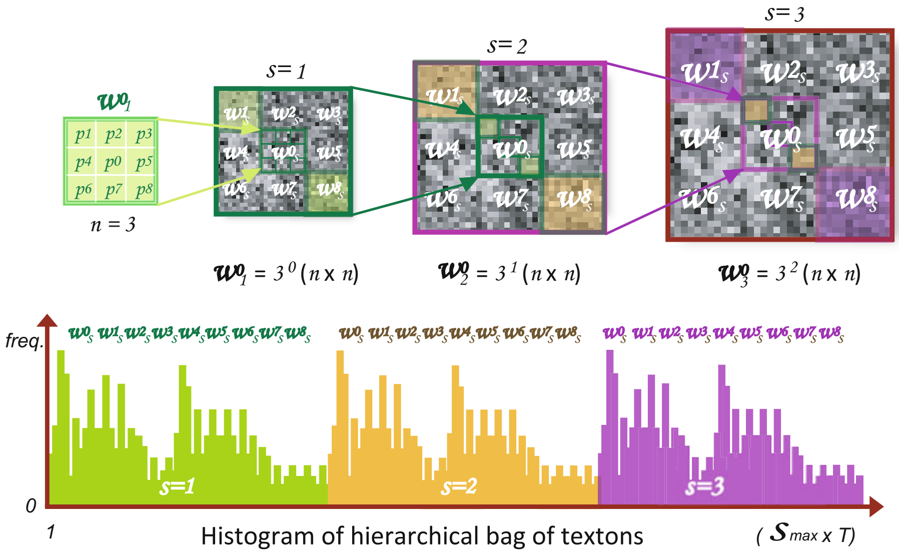


Fig. 3. The histogram of hierarchical bag of textons. Textons are represented by grayscale. The histogram of hierarchical bag of textons are normalized with window size.

ordering constraints of objects. The extracted features are sufficiently general to allow us to automatically learn the context informations for spatial layout and ordering constrain.

Fig. 3 illustrates how the hierarchical bag-of-textons are extracted to features using a multi-scale spatial layout filter. The original bag-of-textons method is computed over local rectangular regions from whole image. As illustrated in Fig. 3, the histogram of hierarchical bag-of-textons is extracted from grid windows increasing its resolution. At first, a set ω_s^0 of a candidate window with a center pixel p_0 are chosen as a $3^{s-1}(n \times n)$ window. The histogram of ω_s^0 concatenate from a top-left (ω_s^1) to bottom-right (ω_s^8) windows covering about $3^s(n \times n)$ the pixel area. The variable s indicates the step of multi-scale. At next, we increase scale step s to expand the features with multi-scale. The multi-scale windows method is effective combined with feature extraction module. We determined the scale step s from 1 to 3 and the initial window n to 3. At last, the size of multi-scale grid windows is normalized to generate a feature vector for object recognition.

A feature vector consists of the grid point's coordinates within the image as a location cue. We concatenated histogram consisting the multi-scale bag-of-texton to the feature vector. Outside the image boundary there is zero contribution to the feature response. We employ the Joint Boost algorithm [15] to select discriminative features of hierarchical bag-of-textons. Random feature selection and sub-sampling improve training time to generate several thousand weak learners. The learned strong classifier is an additive model of the form $H(c, i) = \sum_{m=1}^{M} h_m(c, i)$, summing the classification confidence of M weak classifiers. This confidence value can be reinterpreted as a

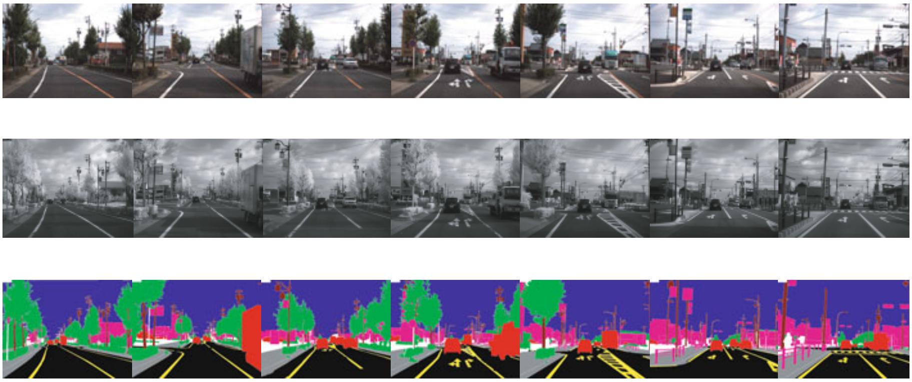


Fig. 4. Multiband Image Dataset Example training images: The first, second, third rows show color images, near-infrared images, and ground truth images, respectively. The assigned classes and colors were: road-black, lane-yellow, sky-blue, tree-green, car-red, trunk and pole-brown, sidewalk-gray, building-magenta, redundancy-white.

probability distribution using the soft-max[17] transformation to give the energy for optimal labeling. Thus, the confidence becomes:

$$P(c|x,i) = log \frac{exp\ H(c,i)}{\sum_c exp\ H(c,i)} \tag{1}$$

At last, the optimal labeling is found by applying the energy minimization algorithm based on the graph cuts [16]. The goal is to find a labeling f which minimizes some energy function. A standard form of the energy function is

$$E(f) = \sum_{p \in P} D_p(f_p) + \sum_{p,q \in N} V_{p,q}(f_p, f_q) \tag{2}$$

where $N \subset P \times P$ is a neighborhood system on pixels. The $D_p(f_p)$ is a data function derived from the probabilities of Joint Boost assigning the label f_p to the pixel p. The $V_{p,q}(f_p, f_q)$ is a smoothness function that measures the cost of assigning the labels f_p, f_q to the adjacent pixels p, q :

$$V_{p,q}(f_p, f_q) = \begin{cases} C & \text{if } f_p \neq f_q \\ 0 & \text{if } f_p = f_q \end{cases} \tag{3}$$

where C is a constant.

4 Experimental Results

This section presents our experimental results for road scene labeling by using the proposed hierarchical bag-of-textons method. We investigated the performance of our system on road image datasets. Input images were captured using a prism-based multiband

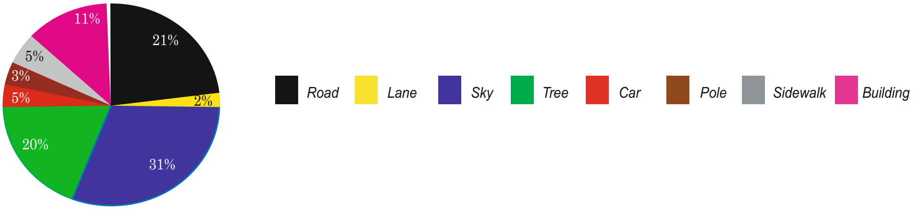


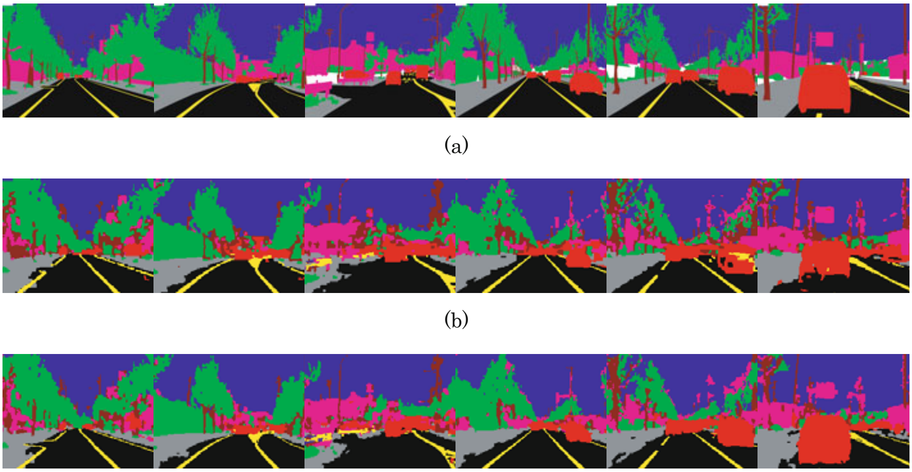
Fig. 5. The proportion of the training pixels in ground truth images

camera (JAI Inc., AD-080CL) mounted on a moving vehicle. The multiband camera can simultaneously obtain both images of color and near-infrared wavelengths. We proceeded to film 3 minutes of daytime footage and made labeled image for each sequence at one fps. In the case of video, each labeled frame could have potentially many other temporally related images associated with it. Each train and test set were captured from 90 video frames at 1.5 minutes. Our dataset contained 8 object classes and assigned a color as shown in Fig. 4. We extracted the features from ahead sequences to get the training patterns and the behind sequences were utilized for the test, which were not used in training image.

We compared the proposed method to conventional bag of texton method, which use single window size (15 × 15 pixels). The amount of training data is biased towards certain classes in our datasets so that we sampled the feature according to proportion of pixels of training set as shown in Fig. 5. We take training and test examples only at pixels lying on a 5 × 5 grid due to exhaustive memory and process time. However, the 20-D filter bank responses and texton map are calculated at full resolution (1024 × 768) for accurate pixel-wise segmentation. The texton number is $T = 297$ for train and test set. At boosting time, we have 10% random feature selection proportion with $M = 6000$ rounding. The constant C of the alpha-expansion algorithm of graph cuts is 0.3 for optimal labeling.

Fig. 6 shows example images for road scene labeling results. In Fig. 7, the table shows the overall recognition rate of the proposed method from total test images. Accuracy is computed by comparing the ground truth pixels to the inferred labeling. Segmentation performance is measured as both the category average accuracy (the average proportion of pixels correct in each category) and the global accuracy (total proportion of pixels correct). The category average is fairer and more rigorous, as it normalizes for category bias in the test set.

The average and the global segmentation accuracy of the proposed method are 80.6% and 84.6%, however, the average and the global of the conventional bag of textons method are 78.3% and 81.6%, respectively. Experimental results showed that the proposed method effectively segments road images and recognize objects in a road environment. The training and test datasets were real video sequences from a multiband camera mounted on a moving vehicle. However, we selected only daytime datasets for the experiments in this paper. If the lighting and weather conditions such as nighttime, snow, and rain are included, our system will struggle. Since robustness is essential for



(a)

(b)

(c)

Fig. 6. Experimental results of test images (a) Ground truth images (b) Labeling results of the conventional bag of textons method (c) Labeling results of the proposed method

Bag of Textons method	*Road*	*Lane*	*Sky*	*Tree*	*Car*	*Pole*	*Sidewalk*	*Building*	Average	Global
Conventinal BoT	89.7	91.1	90.6	74.6	85.6	54.3	83.9	57.2	78.3	81.6
Hierarchical BoT	93.5	93.1	92.7	77.6	86.8	53.9	84.5	62.9	80.6	84.6

Fig. 7. Total results in pixel-wise percentage accuracy on test sequences

ITS, we will attempt to integrate more reasonable features such as appearance features, motion and structural features, and lidar data. However, we have confirmed that the proposed system can play an important role in complex scene understanding for road environment perception. An optimized implementation of our system could be used as an advanced driving assist system.

5 Conclusion

This paper presented a new framework of semantic segmentation scheme for road environment perception using a hierarchical bag of textons method. Experimental results showed that the proposed method can be recognized more accurately than the conventional bag of textons method leading to a considerably better recognition of some objects such as road, tree, and building. Therefore, we can confirm that the proposed system is expected to play an important role in the complex road scene understanding system. In the future, by integrating the algorithm of shape-based objects recognition, we will use the proposed system to expand the road environment perception.

Acknowledgment. This work was supported in part by JSPS.

References

1. Bertozzi, M., Broggi, A., Fascioli, A.: Vision-based Intelligent Vehicles: state of the art and perspectives. Journal of Robotics and Autonomous Systems 32(1), 1–16 (2000)
2. Bertozzi, M., Broggi, A., Cellario, M., Fascioli, A., Lombardi, P., Porta, M.: Artificial vision in road vehicles. Proc. IEEE 90(7), 1258–1271 (2002)
3. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Proc. ECCV Int. Workshop on Statistical Learning in Computer Vision (2004)
4. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (2005)
5. Winn, J., Shotton, J.: The layout consistent random field for recognizing and segmenting partially occluded objects. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (2006)
6. Shotton, J., Johnson, M., Cipolla, R.: Semantic Texton Forests for Image Categorization and Segmentation. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (2008)
7. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (2006)
8. Kang, Y., Kidono, K., Naito, T., Ninomiya, Y.: Multiband image segmentation and object recognition using texture filter banks. In: Proc. Int. Conf. on Pattern Recognition (2008)
9. Varma, M., Zisserman, A.: A Statistical Approach to Texture Classification from Single Images. Int. Journal of Computer Vision 62(1-2), 61–81 (2005)
10. Malik, J., Belongie, S., Leung, T., Shi, J.: Contour and texture analysis for image segmentation. Int. Journal of Computer Vision 43(1), 7–27 (2001)
11. Winn, J., Criminisi, A., Minka, T.: Object Categorization by Learned Universal Visual Dictionary. In: Proc. IEEE Int. Conf. on Computer Vision (2005)
12. Beis, J.S., Lowe, D.G.: Shape indexing using approximate nearest-neighbour search in high-dimensional spaces. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (1997)
13. Jamie Shotton's web site, `http://jamie.shotton.org/work/code.html`
14. Shotton, J., Winn, J.M., Rother, C., Criminisi, A.: TextonBoost: Joint Appearance, Shape and Context Modeling for Multi-Class Object Recognition and Segmentation. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 1–15. Springer, Heidelberg (2006)
15. Torralba, A., Murphy, K.P., Freeman, W.T.: Sharing visual features for multiclass and multiview object detection. IEEE Trans. on Pattern Analysis and Machine Intelligence 19(5), 854–869 (2007)
16. Boykov, Y., Jolly, M.P.: Interactive Graph Cuts for optimal boundary and region segmentation of objects in N-D images. In: Proc. Int. Conf. on Computer Vision (2001)
17. Friedman, J., Hastie, T., Tibshirani, R.: Additive Logistic Regression: a Statistical View of Boosting. Annals of Statistics 28(2), 337–407 (2000)
18. Kumar, S., Hebert, M.: A hierarchical field framework for unified context-based classification. In: Proc. of International Conference on Computer Vision (2005)