Scale-Optimized Textons for Image Categorization and Segmentation

Yousun Kang
Tokyo Polytechnic University
Kanagawa, Japan
Email: yskang@cs.t-kougei.ac.jp

Akihiro Sugimoto
National Institute of Informatics
Tokyo, Japan
Email: sugimoto@nii.ac.jp

***Abstract*—Texton is a representative dense visual word and it has proven its effectiveness in categorizing materials as well as generic object classes. Despite its success and popularity, no prior work has tackled the problem of its scale optimization for a given image data and associated object category. We propose scale-optimized textons to learn the best scale for each object in a scene, and incorporate them into image categorization and segmentation. Our textonization process produces a scale-optimized codebook of visual words. We approach the scale-optimization problem of textons by using the scene-context scale in each image, which is the effective scale of local context to classify an image pixel in a scene. We perform the textonization process using the randomized decision forest which is a powerful tool with high computational efficiency in vision applications. Our experiments using MSRC and VOC 2007 segmentation dataset show that our scale-optimized textons improve the performance of image categorization and segmentation.**

***Keywords*-scale-optimized textons; image categorization; image segmentation; visual words;**

I. Introduction

For a given large dataset in the web-site such as photo sharing, automatically categorizing images becomes more and more important in image retrieval systems. Current search engines offer meta-tags based on simple characteristics of images. If a set of text labels to an image based on its visual content is automatically provided, however, an image retrieval system will drastically become easy to use. Image categorization is one way in which we can perform image retrieval, and can be helpful in semantic segmentation and object recognition tasks. In addition, it can enhance understanding of visual content for easy browsing in the web-site.

Recently image categorization frameworks have shown that the dense sampling of visual words [14] and their combinations with image cues can improve their performance significantly [15]. Textons [9] are promising representative dense visual words. Though early texton studies were limited to their exclusive focus on artificial texture patterns instead of natural images [24], recent studies have proven the effectiveness of texton in categorizing materials [20], various scenes [1], and generic object classes [22]. With employing the bag-of-features model [5], the framework using textons as visual words has become popular and has demonstrated its success in recent years [23]. Textons, unlike sparse image features such as SIFT [13] or HOG [6], can be utilized in both object segmentation and recognition thanks to their high density [16].

The major drawback of the bag-of-features model is that it discards the scale and the spatial layout of visual words, which causes a crucial problem for segmentation and recognition. Accordingly, when texton is used as a visual word, how to incorporate the scale and the spatial layout is a big issue.

Many works have been presented to overcome the problem for the spatial layout [11], [21]. To learn the model of object classes with incorporating texture, layout, and context information, Shotton *et al.* proposed TextonBoost algorithm [17] using a boosted combination of texton features. In addition, the texture-layout filter is employed to capture textural context between texture and spatial layout. After years, Shotton *et al.* [16] proposed the semantic texton forests, and the texture-layout filter is utilized in the segmentation module. By using texture-layout filters, they significantly improved the accuracy of segmentation and recognition.

On the other hand, little attention has been paid to discarded scale information for a given image data and the associated object category. In a large dataset, there are many different scales in object present in an image. Even though objects fall in the same category such as 'cow' or 'car', they have quite different scales in a scene. Scale information of an object can be a significant cue for recognizing the object in a scene. Nevertheless, no prior work has been reported to incorporate scale information into textons.

To deal with scale information, we propose scale-optimized textons for image categorization and segmentation. By extracting scale-optimized textons in the textonization process, more discriminative textons including scale information can be utilized in textural context. We approach the scale-optimization problem of textons by using the scene-context scale in each image pixel, i.e., the effective scale of local context to classify an image pixel in a scene [10]. Our textonization process is carried out using random forests [3], which have been shown to be computationally highly efficient, to generate semantic textons. We extend the random forests into multi-scale texton forests to generate different textons in scale, and then, using the scene-context

978-0-7695-4589-9/11 $26.00 © 2011 IEEE
DOI 10.1109/ISM.2011.48

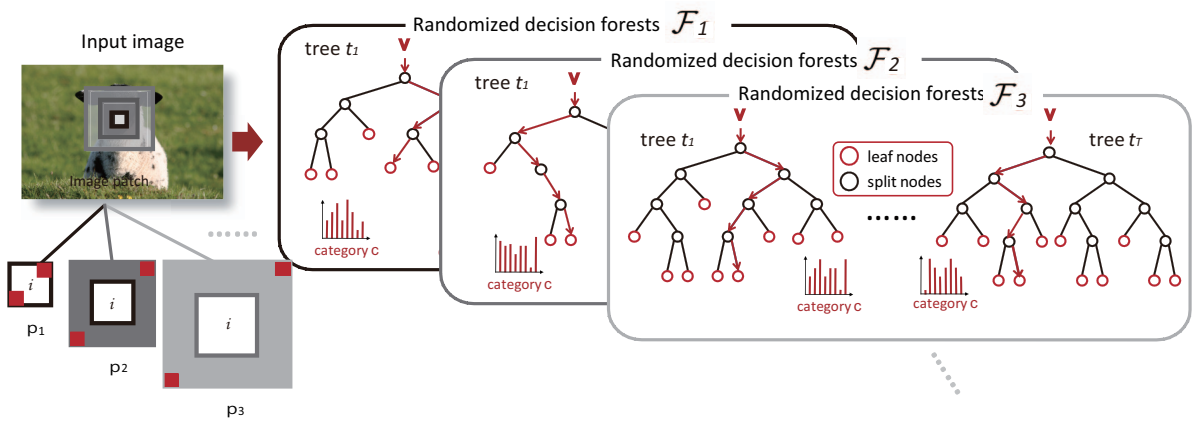


Figure 1. **Dilatation of a region of interest according to scale space k and multi-scale texton forest.** Various sizes of image patches are used for node split function in the multi-scale texton forests (left). The multi-scale texton forest consists of several semantic texton forests [16] with various scale levels and each semantic texton forest consists of randomized decision trees with same scale level (right).

scale, we find the scale-optimized texton, i.e., the texton with the best scale in each image pixel. Accordingly, our textonization process includes semantic and scale information of texture for local textural context. To assess our framework, we compare the accuracy of categorization and segmentation with that of the state-of-the-art [16] using MSRC and VOC 2007 segmentation datasets, showing that our method achieves more accurate categorization and segmentation.

The contribution of this work is the incorporation of scale information into textons as textural context of the object to make them more discriminative. To the best of our knowledge, this is the first work that incorporates scale context into the textonization process. Our scale-optimized textons can be combined with texture-layout filters to improve segmentation accuracy further.

II. Scale-Optimized Textonization

Scale-optimized textons can be obtained by using the scene-context scale in each image pixel. In this section, we explain our textonization process and how to optimize textons to include the best scale using multi-scale texton forests.

A. Multi-scale Texton Forests

We perform textonization process using randomized decision trees to formulate multi-scale texton forests. We employ the semantic texton forests proposed by Shotton *et al.* [16] and generate different scale levels to obtain multi-scale texton forests.

The multi-scale texton forests $\mathcal{F}$ consist of several semantic texton forests with various scale levels $\mathcal{F} = \{\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_s\}$ as shown in Fig. 1, where the scale level is $k = (1, 2, 3, ..., s)$. Each semantic texton forest is a combination of randomized decision trees, each of which has a different set of image patches for its nodes. Split node functions for a randomize decision tree compute the values of raw pixels within an image patch p. By increasing the size of image patches for split node functions, we can expand a semantic texton forest to multi-scale texton forests with different scales.

In the first scale level $k = 1$, an image patch p_1 covers whole pixels within a $(d \times d)$ size on which the split node functions for the first semantic texton forest $\mathcal{F}_1$ act. In the next scale level $k = 2$, the increased image patch p_2 covers the pixels within a $(2d \times 2d)$ size excluding the former image patch p_1. Therefore, the size of image patch p_k is increased to $(kd \times kd)$ pixels excluding the image patch $p_{(k-1)}$ that is for the former scale level $(k-1)$ as shown in Fig. 1.

The combinations of raw pixels within image patches p_k for split node functions are generated randomly, and we also increase the number of the candidates quadratically with respect to the scale level k.

Randomized decision forests have been utilized in classifiers [2], [12] or clustering with the fast and powerful performance. Semantic texton forests [16] are used for both clustering and local classification. To textonize an image, an image patch p_k are passed down the multi-scale texton forest according to their scale level. We can obtain the class distributions $P_k(c|L_k)$ by averaging the local distributions over the leaf nodes $L_k = (l_1, l_2, ..., l_T)$ at scale k as

$$P_k(c|L_k) = \frac{1}{T}\sum_{t=1}^{T} P_k(c|l_t), \quad (1)$$

where c is a category label of a pixel and T is the number of randomized decision trees in $\mathcal{F}_k$. Then, there are several class distributions in multi-scale texton forests as

$$P(c|L) = \{(P_1(c|L_1), P_2(c|L_2), ..., P_s(c|L_s))\}. \quad (2)$$

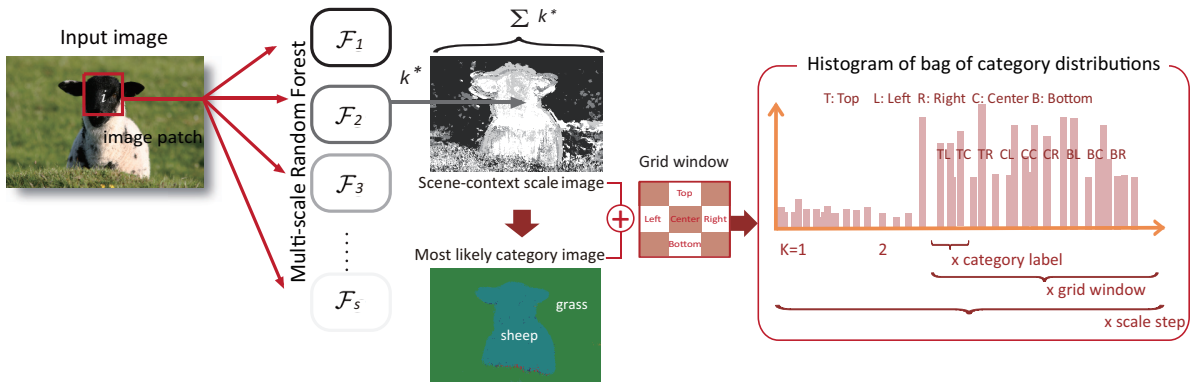


Figure 2. **The scene context scale (left) and the histogram for the bag-of-features model (right).** Left : In the scene-context scale image, darker pixels correspond to smaller scale, so black pixels represent the first scale level $k = 1$ and white pixels represent the largest scale level $k = s$. The most likely category image can be obtain by computing the class distributions of scale-optimized textons. Right : For image categorization and segmentation, we make a histogram using scale-optimized textons. The dimension of a histogram is the number of grid window times number of category times whole scale levels.

B. Scene-Context Scale

Scene-context plays an important role in segmentation and recognition. When the scene-context is used on a per-pixel level, we can capture the local context in which image pixels carry semantic information within a region of interest. Some image pixels, however, have ambiguous features at a very local scale, because the color and texture of the local level do not have capability of identifying the pixel class. Therefore, every image pixel have its available range to search local context in a scene.

The effective region size for local context is called as the scene-context scale [10]. Given object presence and location in a scene, its scale is related to this range and it can be a strong cue for recognizing the objects in the scene. We can estimate the scene-context per each image pixel and use the scene-context scale to find the textons with best scale using multi-scale texton forests.

The scene-context scale of each image pixel is obtained by computing the entropies of an image patch in the leaf nodes of each randomized decision forest. The confidence of each semantic texton forest is thus computed by the entropies of the class distribution over the leaf nodes in $\mathcal{F}_k$ and we regard the confidence as the criterion to find the scene-context scale. Since an object has different scales depending on a scene, and scale of background/foreground appearing together in a scene may be independent of the object, we estimate the scene-context scale per each pixel.

The scale level of the semantic textons forest with minimum entropy of the class distribution is chosen as the scene-context scale at each image pixel i. We compute the entropy $E_k(i)$ of image pixel i from the class distribution $P_k(c|L_k)$ in $\mathcal{F}_k$ as

$$E_k(i) = -P_k(c|L_k) \times \log P_k(c|L_k). \tag{3}$$

Among the all scale levels $k = (1, 2, 3, ..., s)$, the best level k^* is chosen with minimum entropy as

$$k^* = \arg\min_k(\mathcal{F}_k\{E_k(i)\}). \tag{4}$$

The scene-context scale of an image pixel i is the instance k^* of the most likely scale among the whole scale levels.

C. Scale-Optimized Texton

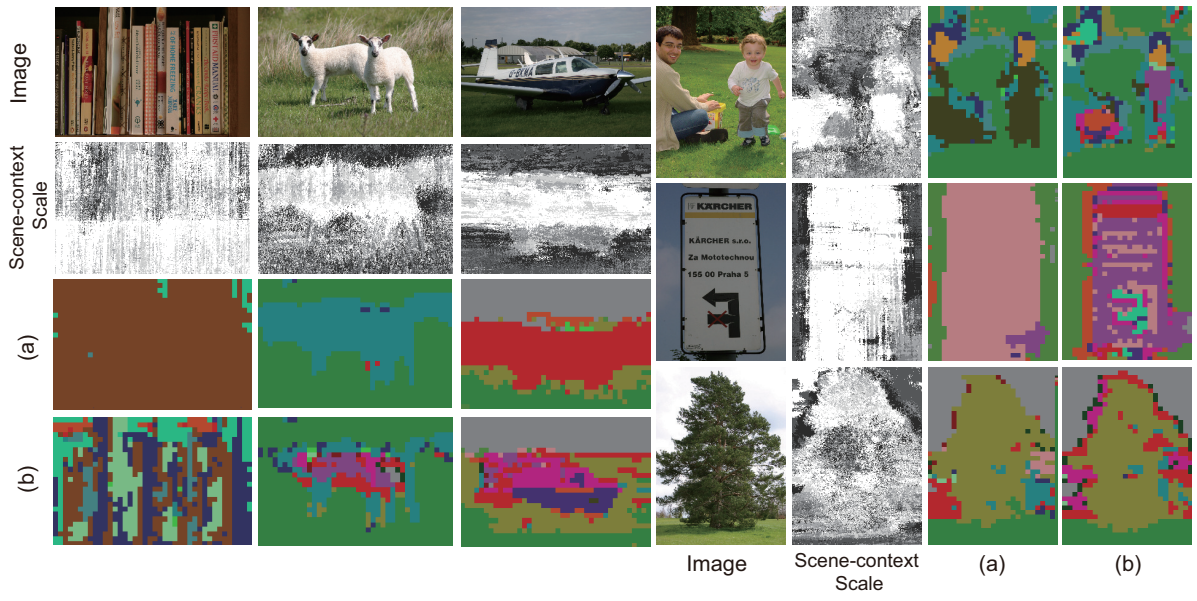
Given an image pixel i, the image patches p centered at the pixel i are classified by descending each randomized decision tree. A randomized decision tree provides both a hierarchical tree structure such as a path from the root to a leaf and the node class distributions at the leaf. From training data, the class distributions can be estimated by averaging the local distributions in a randomized decision trees.

A scale-optimized texton can be generated by computing the scene-context scale of each image pixel from multi-scale texton forests. Among multi-scale texton forests, a semantic texton forest $\mathcal{F}_{k^*}$ is selected in the textonization process. The semantic texton forest $\mathcal{F}_{k^*}$ has the instance k^* of the most likely scene-context scale. We can define the texton generated by the semantic texton forest $\mathcal{F}_{k^*}$ as our scale-optimized texton.

Our scale-optimized textonization process exploits the class distributions $P_{k^*}(c|L_{k^*})$ in the semantic texton forest $\mathcal{F}_{k^*}$ with the scene-context scale k^*. These scale and textural information are utilized in the statistics of scale-optimized textons. By classifying a histogram consisting of the statistics of scale-optimized textons, we can obtain a good performance for pixel-level classification. In addition, we can improve the estimation of class distributions from training data, even the training data do not perform any geometrical transformation in scale and orientation.

III. Categorization and Segmentation

The scale-optimized textons are utilized in the bag-of-features model for image categorization and semantic seg-



	building	grass	tree	cow	sheep	sky	airplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat	class average
(a)	12	90	43	60	79	90	89	36	89	28	34	65	19	6	64	14	46	42	22	36	49	48.3
(b)	2	89	49	43	39	84	36	45	74	30	58	66	28	3	17	6	57	19	14	17	28	38.4

Figure 3. **Clustering and classification results using scale-optimized textons.** Above : (a) Classification results with using scale-optimized textons. (b) Classification results without using scale-optimized textons [16]. Below: Classification accuracies (percent) over the whole dataset, without-(b), and with-(a), the scale-optimized textons. Our new highly efficient scale-optimized textons achieve a significantly improvement on previous work (b) in class average.

mentation. Once a scale-optimized texton is determined, we can calculate the class distributions of each image pixel using the scale-optimized texton. We make a histogram which consists of the class distributions computed across the whole image for image categorization. The histogram contains the scale and textural context by using both the most likely category $c_{i^*} = \arg\max_{c_i} P_{k^*}(c_i|L_{k^*})$, and the most likely scene-context scale $k^* = \arg\min_k(\mathcal{F}_k\{E_k(i)\})$.

However, since the bag-of-features model discards spatial layout, we use a simple grid window to learn the layout of scale and textural context automatically as shown in the middle of Fig. 2. The grid window consists of nine sub-grids such as Top-Left (TL), Top-Center (TC), Top-Right (TR), Center-Left (CL), Center-Center (CC), Center-Right (CR), Bottom-Left (BL), Bottom-Center (BC), and Bottom-Right (BR) as shown in the right of Fig. 2. We concatenated the histograms from TL to BR, and the histogram is used as input to a classifier to recognize object categories.

We adopt the non-linear support vector machine (SVM) to classify each category. Multi-class classification is performed with LibSVM [4] trained using the one-versus-all rule: a classifier is learned to separate each class from the rest, and a test image is assigned the label of the classifier with the highest response.

When a histogram is created over a region of interest for each pixel, it can be utilized in pixel-wise semantic segmentation. To obtain more accurate segmentation performance, it is possible to combine with the texture layout file instead of our simple grid window. However, since the class distributions are extracted from scale-optimized textons, the results of the first clustering and classification guarantee a good performance. We show the performance of clustering and classification in Section IV-A.

IV. Experimental Results

This section presents our experimental results for image categorization and segmentation using scale-optimized textons. We evaluated our algorithm using MSRC [18] and challenging VOC 2007 [7] segmentation datasets that include a variety of objects such as building, cow, sheep, water, face, cat, road, sky, and so on.

In MSRC dataset, there are 256 images for training, 257 images for test, and remaining 59 images for validation. In VOC 2007 segmentation dataset, there are 209 images for training, 210 images for test, and remaining 213 images for

	building	grass	tree	cow	sheep	sky	airplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat	class average
(a) Ours based RBF	97	98	98	100	100	99	100	98	99	98	100	100	100	99	100	100	94	99	100	97	96	***98.7***
(b) Ours based PMK	90	90	73	91	93	94	100	95	77	90	100	96	100	94	96	84	78	98	97	74	93	90.6
(c) Shotton [19]	64	86	75	86	92	90	74	66	64	88	72	84	70	53	90	67	67	57	36	64	77	72.8
(d) Our segmentation	**45**	89	60	62	65	86	80	**50**	**89**	**70**	**58**	**73**	**48**	20	**80**	**44**	68	37	**31**	**57**	43	***59.8***
(e) Shotton [19]	37	86	**62**	**65**	74	83	74	42	87	69	**58**	**73**	47	**24**	77	42	**70**	**45**	28	47	40	58.6

Figure 4. **Image categorization ((a),(b), and (c)) and segmentation ((d) and (e)) results on MSRC dataset**. Categorization and segmentation accuracies (percent) over the whole dataset. The proposed scale-optimized texton achieves a significant improvement of image categorization on previous work.

validation. We used the standard training/validation data for training and used test data for our test.

A. Scale-Optimized Textonization

To access the efficiency of the proposed scale-optimized textons, we compared the class classification accuracy with that of conventional semantic texton forests method [16] that is without using scale-optimized texton.

We separately trained the semantic texton forests in different scale levels. To train the multi-scale texton forest, we prepared six scale levels $k = (1, 2, 3, 4, 5, 6)$ and an initial image patch size was (15×15). Therefore, the size of image patches p for split function is $(15k \times 15k)$ at each scale level k. Each semantic texton forest $\mathcal{F}_k$ had the following parameters, $T = 5$ trees, maximum depth $D = 10$, $400 \times 2k$ feature and $10k$ threshold tests per split function, and 0.25 of the data per tree. Training a semantic texton forest took approximately 30×2^k minutes on MSRC dataset and 60×2^k minutes on VOC 2007 at each scale step.

Fig. 3 shows the several scene-context scale image on the MSRC test dataset. Using the scene-context scale, we can obtain scale-optimized textons, and infer the most likely category for each pixel as shown in Fig. 3(a). On the other hand, Fig. 3(b) shows the results of the state-of-the-art [16] that is based on single-scale semantic texton forests. The single-scale semantic texton forest used the same parameter of the multi-scale texton forests with the first scale level $\mathcal{F}_1$.

Clustering and class classification performance is measured as both the class average accuracy (the average proportion of pixels correct in each category) and the global accuracy (total proportion of pixels correct) as shown in the bottom table of Fig. 3. The global classification accuracy without scale-optimized textons gives 50.2% while that with using scale-optimized textons scale gives 53.0%. In particular, significant improvement can be observed in most of the classes. For some classes such as tree or water, however, we cannot see the improvement. This may come from the fact that they have not influence on scale-optimized textons due to their strong textural property. Across the whole MSRC dataset, using the scale-optimized textons achieved a class average performance of 48.3%, which advances 38.4% of (b) as shown in the table of Fig. 3.

B. Categorization and Segmentation

As a result of image categorization, we obtained the accuracy of MSRC as shown in the upper side of table in Fig. 4. For non-linear SVM classifier, we compared the class average using radial basis function (RBF) kernel and pyramid match kernel (PMK) [8] to the state-of-the-art [16]. We confirmed that the RBF kernel gives improved results than the PMK. As can be seen, the proposed method using the scale-optimized textons gives significantly better results than the selected state-of-the-art and improved the performance for all categories.

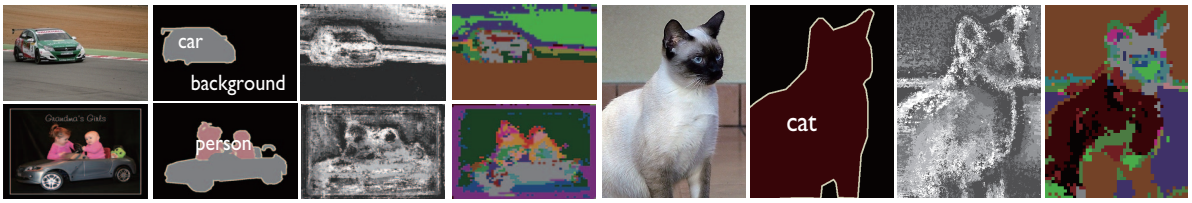
To demonstrate the power of the scale-optimized textons as features for segmentation, we employed the joint boosting algorithm [19] to select discriminative features of the bag-of-features model. The semantic segmentation results on MSRC test data are shown in the bottom side of Fig. 4. As can be seen, the proposed segmentation algorithm improves the accuracy in the local classification process, in particular the classes with the result of noisy clustering such as water, car, bicycle, sign and road, show good performance in this process. We obtained the segmentation results with global 65.2%, class average 59.8% using the bag-of-features model with scale-optimized textons.

We compared the proposed method with the state-of-the-art in the table of Fig. 4. In fact, the results of the state-of-the-art is better than 58.6% in their paper [16], since they augmented the training data with image copies that are artificially transformed geometrically and photometrically. However, in our experiments, we do not use any geometric transformations, and affine photometric transformations such as rotation, scaling, and left-right flipping.

Fig. 5 shows the results of the our scale-optimized textonization using VOC 2007. As shown in the table of Fig. 5, a pixel-level classification based on the class distributions gives a good performance (13.7%) even they do not cooperate with any spatial-layout information.

V. Conclusion

This paper presented a method that incorporates scale information into textons as local textural context of the object to make them more discriminative. Differently from existing methods, our method directly incorporates scale



Segmentation	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	lable	dog	horse	motorbike	person	plant	sheep	sofa	train	monitor	class average
Our scale-optimized textons	50	2	9	15	19	14	3	8	8	9	13	19	13	5	4	28	27	9	7	12	13.7
Brroks 2007 [7]	6	0	0	0	0	9	5	10	1	2	11	0	6	6	29	2	2	0	11	1	8.5
INRIA normal 2007 [7]	1	2	8	2	52	0	12	6	4	0	18	4	0	0	4	29	0	6	0	10	7.7
Image categorization	96	97	98	96	95	98	93	99	91	98	98	98	99	99	77	95	98	99	99	95	95.9

Figure 5. **The result images of clustering and the class classification (above) on VOC 2007.** The VOC 2007 contains 21 challenging categories including background. The bottom table shows the accuracy of the clustering and the class classification and also image categorization (last row).

information into the textonization process. By extending the random forests into multi-scale texton forests, our method generates different textons in scale, and then, using the scene-context scale, finds the scale-optimized texton, i.e., the texton with the best scale in each image pixel. Our experiments showed that using our scale-optimized textons improves the performance of image categorization and segmentation. It is expected that our scale-optimized textons are combined with texture-layout filters [18] to improve segmentation accuracy further.

ACKNOWLEDGMENT

This work was in part supported by JST, CREST and JSPS.

REFERENCES

[1] S. Battiato, G. Farinella, G. Gallo, and D. Ravi. Spatial hierarchy of textons distributions for scene classification. *In LNCS*, 5371:333–343, 2009.

[2] A. Bosch, A. Zisermann, and X. Munoz. Image classification using random forests and ferns. *In ICCV*, pages 1–8, 2007.

[3] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[4] C. Chang and C. Lin. Libsvm: a library for support vector machines., 2001. Software available at http://www.csie.ntu.edu.tw/libsvm.

[5] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. *In ECCV Workshop on Statistical Learning in ComputerVision*, 2004.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *In CVPR*, 1:886–893, 2005.

[7] M. Everingham, L. V. Gool, C. K. I.Williams, J.Winn, and A. Zisserman. *The PASCAL VOC Challenge 2007*. 2007. http://www.pascalnetwork.org/challenges/VOC/voc2007/workshop.

[8] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. *In ICCV*, pages 1458–1465, 2005.

[9] B. Julesz. Textons, the elements of texture perception and their interactions. *Nature*, 290:91–97, 1981.

[10] Y. Kang, H. Nagahashi, and A. Sugimoto. Semantic segmentation and object recognition using scene-context scale. *Proc. of Pacific-Rim Symposium on Image and Video Technology*, pages 39–45, 2010.

[11] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *In CVPR*, pages 2169–2178, 2006.

[12] V. Lepetit, P. Lagger, and P. Fua. Randomized trees for real-time keypoint recognition. *In CVPR*, 2:775–781, 2005.

[13] D. Lowe. Distinctive image features from scale-invariant keypoints. *Int. Journal of Computer Vision*, 60(91–110), 2004.

[14] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *In CVPR*, 2003.

[15] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. *In ECCV*, 2006.

[16] J. Shotton, M. Johnson, , and R. Cipolla. Semantic texton forests for image categorization and segmentation. *In CVPR*, 1:1–8, 2008.

[17] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling formulti-class object recognition and segmentation. *In ECCV*, 3951:1–15, 2006.

[18] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding : Multi-class object recognitionand segmentation by jointly modeling texture, layout, and context. *Int. Journal of Computer Vision*, 81(2-23), 2009.

[19] A. Torralba, P. Murphy, and W. T. Freeman. Sharing visual features for multiclass and multiview object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(5):854–869, 2007.

[20] M. Varma and A. Zisserman. A statistical approach to texture classification from single images. *Int. Journal of Computer Vision*, 62(1):61–81, 2005.

[21] J. Willamowski, D. Arregui, G. Csurka, C. Dance, and L. Fan. Categorizing nine visual classes using local appearance descriptors. *In ICPR Workshop on Learning for Adaptable Visual Systems.*, 2004.

[22] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. *In ICCV*, 2:1800–1807, 2005.

[23] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classificaiton of texture and object categories: A comprehensive study. *Int. Journal of Computer Vision*, 73(2):213–238, 2007.

[24] S. Zhu, C. Guo, Y. Wang, and Z. Xu. What are textons? *Int. Journal of Computer Vision*, 62(1):121–143, 2005.