Image Categorization and Semantic Segmentation using Scale-Optimized Textons

Yousun Kang*
Tokyo Polytechnic University
Atsugi 243-0297, Japan
yskang@t-kougei.ac.jp

Akihiro Sugimoto
National Institute of Informatics
Tokyo 101-8430, Japan
sugimoto@nii.ac.jp

Abstract

In computer vision research, a texton is a representative dense visual word for the bag-of-keypoints method. It has proven its effectiveness in categorizing materials and in generic object classes. Despite its success and popularity, no report describes a study that has tackled the problem of its scale optimization for given image data and associated object categories. We propose scale-optimized textons to learn the best scale for each object in a scene. We incorporate them into image categorization and semantic segmentation. Our textonization module produces a scale-optimized codebook of visual words. We approach the scale-optimization problem of textons 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 a randomized decision forest, which is a powerful tool with high computational efficiency in vision applications. Results of our experiments using MSRC and VOC 2007 segmentation datasets demonstrate that our scale-optimized textons improve image categorization and segmentation performance.

Keywords: scale-optimized textons; image categorization; semantic segmentation; decision forest

1 Introduction

Automatically categorizing images has become increasingly important for image retrieval systems such as those for photo-sharing on web-sites. Current search engines offer meta-tags based on simple characteristics of images in a given large dataset. Smart devices and systems for image retrieval would become radically more intelligent and easy to use if a set of text labels to an image based on its visual content could be provided automatically. Image categorization is a means to perform image retrieval and it can be helpful in semantic segmentation and object recognition tasks. Additionally, it can enhance understanding of visual contents for easy browsing of web-sites. Moreover, it can develop convergence technologies for databases, data mining, and artificial intelligence applications.

Recently, image categorization frameworks have shown that dense sampling of visual words [19] and their combinations with image cues [5] can improve their performance significantly [21]. Textons [11] are promising representative dense visual words. Although early texton studies were limited to their exclusive emphasis on artificial texture patterns instead of natural images [35], recent studies have proven the effectiveness of textons for categorizing materials [27], various scenes [1], and generic object classes[32]. Using a bag-of-features model [5, 25], the framework for using textons as visual words has become popular and has demonstrated its success in recent years [34]. Textons, unlike sparse image features such as SIFT [17] or HOG [6], are useful in both object segmentation and recognition because of their high density [22].

IT CoNvergence PRActice (INPRA), volume: 2, number: 1, pp. 2-14

*Corresponding author: 1583 Iiyama, Atsugi, Kanagawa 243-0297 Japan, Tel: +81-46-242-9524, Web: http://researchmap.jp/yskang/?lang=english

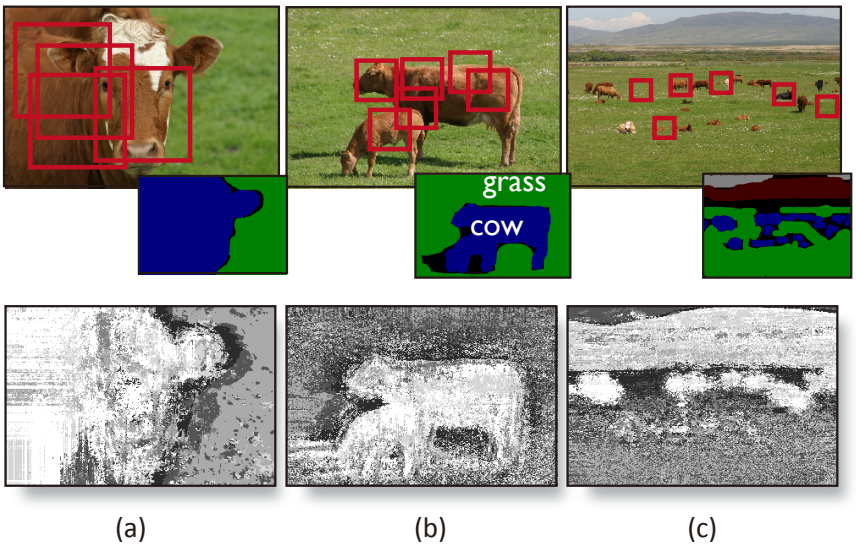


Figure 1: **Example images of the 'cow' category on MSRC dataset.** The objects have different scales in a large dataset such as large scale (a), middle scale (b), and small scale (c). Darker pixels correspond to the smaller scale. Because, scale-optimized textons include both scale and textural context in each image pixel, we can produce more discriminative features to recognize the object in a scene.

The major shortcoming of the bag-of-features model is that it discards the scale and the spatial layout of visual words, which engenders a daunting problem hindering segmentation and recognition. When a texton is used as a visual word, incorporation of the scale and the spatial layout becomes an important issue. First, many works have been presented to overcome the problem of the spatial layout [15, 31, 9, 28, 29]. To learn the model of object classes with incorporating texture, spatial layout, and context information, Shotton *et al.* proposed a texture-layout filter in TextonBoost algorithm [23]. Their filter can capture a textural context between texture and spatial layout using a boosted combination of texton features. The filter markedly improved the accuracy of segmentation and recognition.

Nevertheless, little attention has been devoted to discarded scale information for given image data and associated object categories. A large dataset has numerous scales of objects that are present in an image. As shown in Fig. 1, although objects might fall into the same category such as 'cow', they have different scales in a scene. Scale information of an object can be an important cue for recognizing the object in a scene, but report in the literature describes the incorporation of scale information into textons.

To address scale information and textons, we propose scale-optimized textons for image categorization and semantic segmentation. We use more discriminative bag-of-features by extracting scale-optimized textons in the textonization process. The scale-optimized texton can incorporate a texture-layout filter to capture the scale and textural context of the objects in a scene. We approach the scale-optimization problem of textons 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 [12]. Our textonization process is first conducted using random forests [3], which have been shown to be computationally highly efficient, to generate base textons. We then extend the random forests into multi-scale texton forests to generate various textons with different scales. Furthermore, we estimate a scene-context scale using the proposed multi-scale texton forests[13]. Finally, it is possible to extract the scale-optimized texton. In other words, a scale-optimized texton is a discriminative feature extracted from each image pixel in the best scale for the local

context, accordingly, the scale-optimized texton includes both semantic and scale information for local textural context.

Our scale-optimized textons can be combined with texture-layout filters to improve segmentation accuracy further. For image categorization, a histogram of the class distributions of scale-optimized textons is computed over the whole image. The histogram is combined with texture-layout filters for semantic segmentation. To assess our framework, we compare the accuracy of categorization and segmentation with that of the state-of-the-art [22] using MSRC21 and VOC 2007 segmentation datasets. Our results show that our method achieves better categorization and segmentation accuracy than those of the state-of-the-art using scale-optimized textons. The contribution of this work is the incorporation of scale information into textons as the textural context of the object to make them more discriminative. This report is the first describing method incorporates scale context into the textonization process.

This paper is organized as follows. In Section 2, we review related works on segmentation and recognition related to the spatial layout and scale information. Section 3 explains the textonization process to find the best scale of objects using scene-context scale. Section 4 describes how to combine scale-optimized textons of each category into the image categorization and segmentation module. Section 5 presents experimentally obtained results related to performance. The salient conclusions are presented in the final section.

2 Related work

Texton [11, 18, 27] is an efficient image representation used for both object segmentation and recognition. Densely discriminative textons facilitate pixel-wise segmentation [24] and image labeling [33]. Malik *et al.* [18] analyzed images into texton channels for image segmentation by mapping each pixel to the texton nearest to its vector of bank filter responses. They established a typical textonization process such as computing filter-banks, performing k-means clustering, and nearest-neighbor assignment, but it is quite time-consuming. To avoid time-consuming computations, Shotton *et al.* [22] proposed a fast and efficient textonization process using randomized decision trees.

These textonization processes can produce efficient and powerful textons for segmentation and recognition. When objects in the same category have various scales in a dataset, however, scale becomes an important factor to be considered. The multi-scale framework is commonly used to include scale information.

Grauman and Darrell [10] proposed a fast kernel function called the pyramid match using multi-resolution histograms. The pyramid match hierarchically measures similarity between histograms, which consist of sets of features extracted from the finest resolution to the coarsest one. The proposed kernel approximates the optimal partial matching by computing a weighted intersection over multi-resolution histograms for classification and regression tasks.

Wang and Wang [30] proposed a multiple scale learning framework to learn the best weights for each scale in the spatial pyramid matching [15]. The multiple-scale learning method can ascertain the optimal combination of base kernels constructed in different image scales for visual categorization.

Kivinen *et al.* [14] proposed a multi-scale graphical model for categorization of natural scenes. They developed a nonparametric Bayesian model, which captures an interesting qualitative structure in novel images using a multi-scale representation. The model based on a tree structure engenders a fast and accurate categorization performance.

As explained above, existing multi-scale approaches are developed in kernel-based learning or in a graphical model. Their focus is how to incorporate multi-scale information into the learning process using extracted image features. In contrast, our method directly incorporates scale information into the feature extraction module, i.e., the textonization process. Our scale-optimized textons have scale context

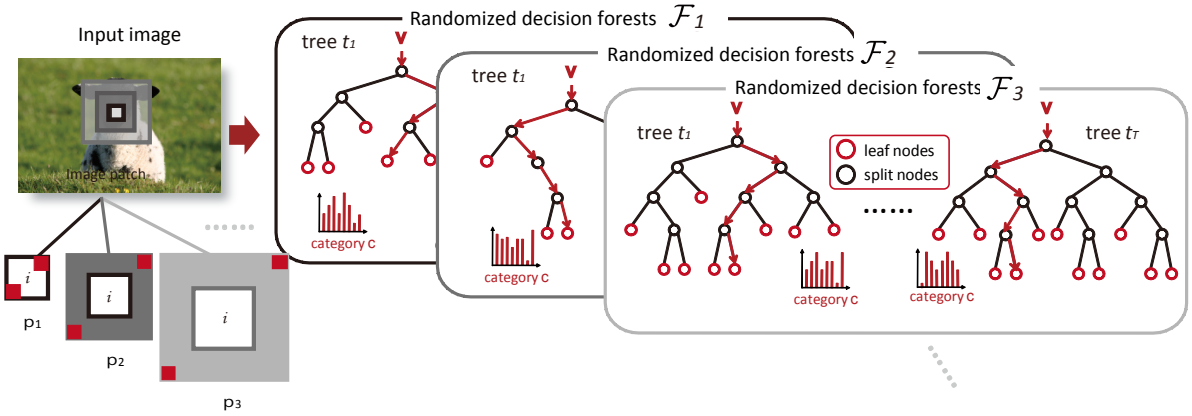


Figure 2: **Dilatation of a region of interest according to scale space k and multi-scale texton forest.** Various sizes of image patches are used for node split function in the multi-scale texton forests (left). The multi-scale texton forest consists of several semantic texton forests [22] with various scale levels. Each semantic texton forest consists of randomized decision trees with the same scale level (right).

in themselves.

3 Scale-Optimized Textonization

Scale-optimized textons are obtainable 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.

3.1 Multi-scale Texton Forests

We perform textonization processing using randomized decision trees to formulate multi-scale texton forests. The semantic texton forests proposed by Shotton *et al.* [22] are used to generate different scale levels to obtain multi-scale texton forests.

The multi-scale texton forests $\mathscr{F}$ consist of several semantic texton forests with various scale levels $\mathscr{F} = \{\mathscr{F}_1, \mathscr{F}_2, ..., \mathscr{F}_s\}$ as shown in Fig. 2, 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 randomized 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 $\mathscr{F}_1$ 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 $p_{(k-1)}$ for the former scale level $(k-1)$ as shown in Fig. 2.

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

Randomized decision forests have been used in classifiers [2, 16] or clustering [20] with fast and powerful performance. Semantic texton forests [22] are used for both clustering and local classification.

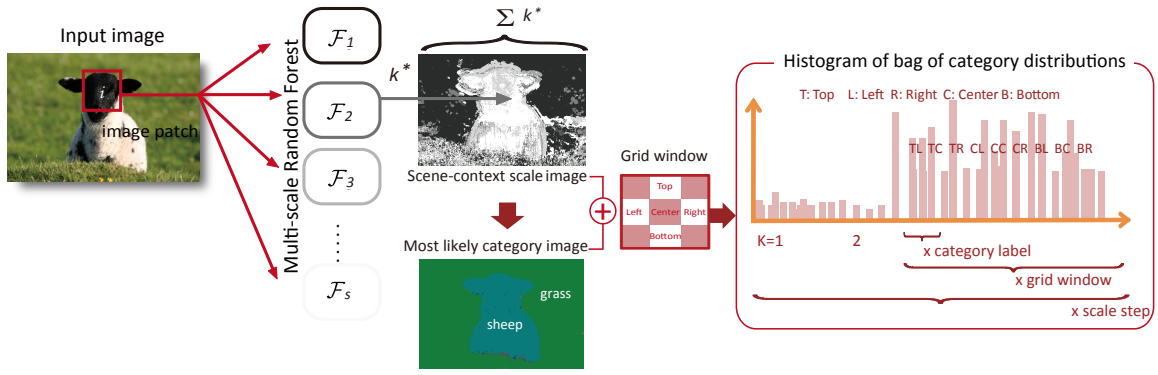


Figure 3: **The scene context scale (left) and the histogram for the bag-of-features model (right).** Left : In the scene-context scale image, darker pixels correspond to a smaller scale, so black pixels represent the first scale level $k = 1$ and white pixels represent the largest scale level $k = s$. The most likely category image can be obtained by computing the class distributions of scale-optimized textons. Right : For image categorization and segmentation, we produce a histogram using scale-optimized textons. The dimension of a histogram is the number of grid windows times the number of categories times all scale levels.

To textonize an image, an image patch p_k is 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

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

where c is a category label of a pixel and T is the number of randomized decision trees in $\mathscr{F}_k$. Several class distributions in multi-scale texton forests exist as

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

3.2 Scene-Context Scale

Scene-context plays an important role in segmentation and recognition [7]. 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 have no capability of identifying the pixel class. Therefore, every image pixel has its available range to search for a local context in a scene.

The effective region size for a local context is designated as the scene-context scale [12]. Given the object presence and location in a scene, its scale is related to this range. It can be a strong cue for recognizing the objects in the scene. We can estimate the scene-context per image pixel and use the scene-context scale to find textons with the 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 therefore computed by the entropies of the class distribution over the leaf nodes in $\mathscr{F}_k$. We regard the confidence as the criterion to find the scene-context scale. Because an object has different

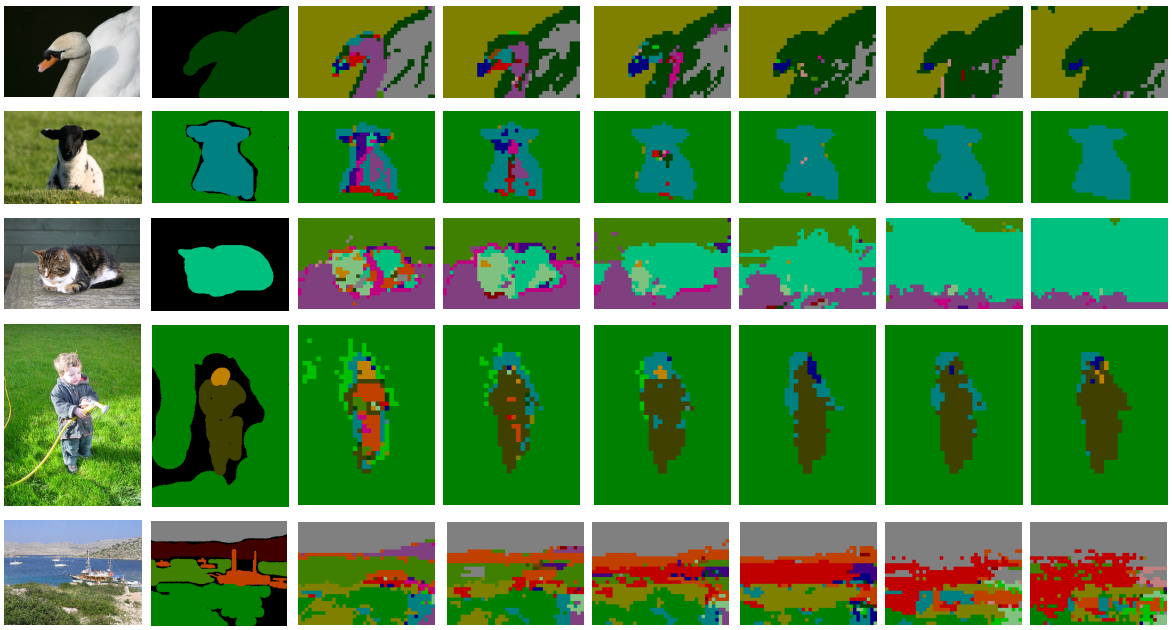


Figure 4: **Clustering and classification results on MSRC segmentation dataset using multi-scale texton forests.** The multi-scale texton forest can generate different textons according to scale levels. (a) Input images. (b) Ground-truth images. (c) – (h) Clustering and classification results according to scale levels $k = (1,2,3,4,5,6)$. The results correspond to each scale level : $k = 1$:(c), $k = 2$:(d), $k = 3$:(e), $k = 4$:(f), $k = 5$:(g), and $k = 6$:(h).

scales depending on a scene, and because the scale of background/foreground appearing together in a scene might be independent of the object, we estimate the scene-context scale per 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 $\mathscr{F}_k$ as

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

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

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

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

3.3 Scale-Optimized Texton

Given an image pixel i, the image patches p centered at 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. Based on training data, the class distributions can be estimated by averaging the local distributions in 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 $\mathscr{F}_{k^*}$ is selected in the textonization process. The semantic texton forest $\mathscr{F}_{k^*}$ has the instance k^* of the most likely scene-context scale. We can define the texton generated by the semantic texton forest $\mathscr{F}_{k^*}$ as our scale-optimized texton.

Our scale-optimized textonization process exploits the class distributions $P_{k^*}(c|L_{k^*})$ in the semantic texton forest $\mathscr{F}_{k^*}$ with the scene-context scale k^*. These scale and textural information are used in the statistics of scale-optimized textons. By classifying a histogram consisting of the statistics of scale-optimized textons, we can obtain good performance for pixel-level classification. Additionally, we can improve the estimation of class distributions from training data, even if the training data perform no geometrical transformation in terms of scale and orientation.

4 Categorization and Segmentation

Scale-optimized textons are used in the bag-of-features model for image categorization and semantic segmentation. Once a scale-optimized texton is determined, we can calculate the class distributions of each image pixel using the scale-optimized texton. We produce a histogram consisting of class distributions computed across the whole image for image categorization. The histogram contains the scale and textural context 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 $k^* = \arg\min_k(\mathscr{F}_k\{E_k(i)\})$.

However, because 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. 3. The grid window consists of nine sub-grids as shown in the right of Fig. 3 : 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). We concatenated the histograms from TL to BR. The histogram is used as input to a classifier to recognize object categories.

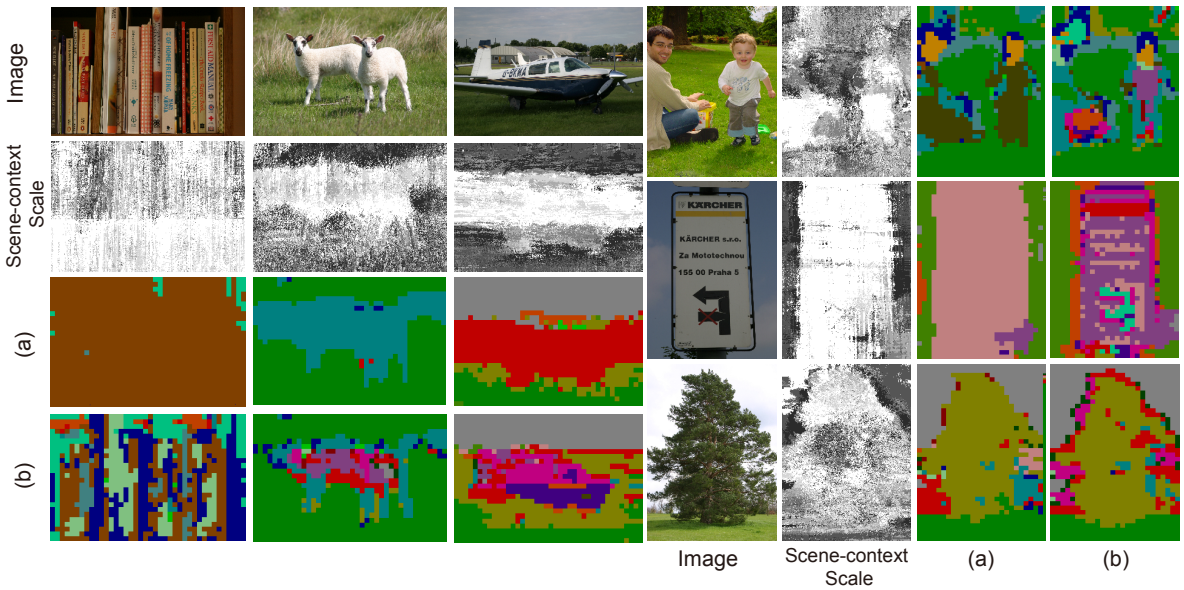
We adopt the non-linear support vector machine (SVM) to classify each category. Multi-class classification is performed with LibSVM [4] trained using the one-versus-all rule: a classifier is learned to separate each class from the rest, and a test image is assigned the label of the classifier with the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

When a histogram is created over a region of interest for each pixel, it is useful 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, because the class distributions are extracted from scale-optimized textons, the results of the first clustering and classification guarantee good performance. We present the performance of clustering and classification in Section 5.1.

5 Experimental Results

This section presents experimentally obtained results for image categorization and segmentation using scale-optimized textons. We evaluated our algorithm using MSRC [24] and challenging VOC 2007 [8] segmentation datasets that include various objects such as building, cow, sheep, water, face, cat, road, and sky.

The MSRC dataset has 256 images for training, 257 images for test, and remaining 59 images for validation. The VOC 2007 segmentation dataset has 209 images for training, 210 images for test, and remaining 213 images for validation. We used standard training/validation data for training and used test data for our test.



	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 5: **Clustering and classification results obtained using scale-optimized textons.** Above: (a) Classification results with using scale-optimized textons. (b) Classification results without using scale-optimized textons [22]. Below: Classification accuracies (percent) over the whole dataset, without-(b), and with-(a), the scale-optimized textons. Our new highly efficient scale-optimized textons achieved marked improvement over that of a previously reported method (b) in terms of the class average.

5.1 Scale-Optimized Textonization

To assess the efficiency of the proposed scale-optimized textons, we compared the class classification accuracy with that achieved by the conventional semantic texton forests method [22] without using the scale-optimized texton.

We separately trained the semantic texton forests at different scale levels. To train the multi-scale texton forest, we prepared six scale levels $k = (1,2,3,4,5,6)$. The initial image patch size was (15×15). Therefore, the size of image patches p for the split function is $(15k \times 15k)$ at each scale level k. Each semantic texton forest $\mathscr{F}_k$ 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 min on the MSRC dataset and 60×2^k min on the VOC 2007 at each scale step.

MSRC Dataset [24] Fig. 4 presents results of clustering and class classification based on multi-scale texton forests. We visualized the most likely categories of each pixel. As shown in Fig. 4, a semantic texton forest has different local class distributions according to its scale. Regarding results of the first and second rows of Fig. 4, we notice that the more the scale level increases, the more the performance also increases. The image of the third row shows the roughest result in the largest scale level. Most

	building	grass	tree	cow	sheep	sky	airplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat	class average
Image categorization																						
(a) Our 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) Our 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
Semantic Segmentation																						
(d) Scale-optimized textons	**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 6: **Image categorization ((a),(b), and (c)) and segmentation ((d) and (e)) results on the MSRC dataset**. Categorization and segmentation accuracies (percent) over the whole dataset. The proposed scale-optimized texton achieves marked improvement of image categorization on that described in previous reports.

intriguingly of all, the fourth row's image has different performance among categories according to the scale level : the 'face' is classified at the smallest scale, but the 'body' is classified at a larger scale. The image of the last row, shows good performance in the smallest scale as we expected. Results show that the scene-context scale is estimated per image pixel.

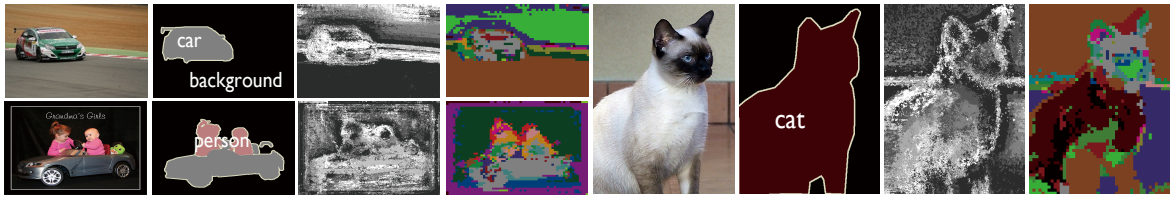
Fig. 5 shows a several scene-context scale image of the MSRC test dataset. Using the scene-context scale, we can obtain scale-optimized textons, and can infer the most likely category for each pixel as shown in Fig. 5(a). Then Fig. 5(b) shows results of the state-of-the-art [22] based on single-scale semantic texton forests. The single-scale semantic texton forest used the same parameter of the multi-scale texton forests with the first scale level $\mathscr{F}_1$.

Clustering and class classification performance are measured as both the class average accuracy (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. 5. The global classification accuracy without scale-optimized textons gives 50.2%, although that with using scale-optimized textons scale gives 53.0%. Particularly significant improvement is visible in most classes. For some classes such as trees or water, however, no improvement is apparent. This lack of benefit might derive from the fact that they have no influence on scale-optimized textons because of their strong textural property. Across the whole MSRC dataset, using the scale-optimized textons achieved a class average performance of 48.3%, which is greater than the 38.4% of (b), as shown in Fig. 5.

VOC 2007 Segmentation Dataset [8] Fig. 7 shows the results of our scale-optimized textonization. As presented in Fig. 7, a pixel-level classification based on the class distributions gives a good performance (13.7%), even if it does not cooperate with any spatial-layout information. Therefore, we can confirm that the proposed scale-optimized textons can be powerful and discriminative visual words for the bag-of-features model.

5.2 Categorization and Segmentation

As a result of image categorization, we obtained the accuracy of VOC 2007 and MSRC categories, as shown respectively in the last row of the table of Fig. 7 and the upper side of table in Fig. 6. For a non-linear SVM classifier, we compared the class average using a radial basis function (RBF) kernel and pyramid match kernel (PMK) [10] to the state-of-the-art [22]. We confirmed that the RBF kernel gives



Semantic Segmentation	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	lable	dog	horse	motorbike	person	plant	sheep	sofa	train	monitor	class average
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 7: **Result images of clustering and the class classification (upper) on VOC 2007.** The VOC 2007 contains 21 challenging categories including the background. The bottom table shows the accuracy of the clustering and the class classification and also image categorization (last row).

improved results than the PMK. As might be apparent, the proposed method using the scale-optimized textons provides considerably better results than the selected state-of-the-art method. It shows improved performance for all categories.

To demonstrate the power of the scale-optimized textons as features for segmentation, we employed the joint boosting algorithm [26] to select discriminative features of the bag-of-features model. The semantic segmentation results for MSRC test data are shown at the bottom of Fig. 6. As might be readily apparent, the proposed segmentation algorithm improves the accuracy in the local classification process. In particular, classes with the result of noisy clustering such as water, car, bicycle, sign and road, show good performance in this process. We obtained segmentation results with global 65.2% and class average 59.8% using the bag-of-features model with scale-optimized textons.

We compared results obtained using the proposed method with those obtained using the state-of-the-art method in the table of Fig. 6. In fact, the results obtained from the state-of-the-art method are better than 58.6% in their paper [22], because they augmented the training data with image copies that are artificially transformed geometrically and photometrically. However, our experiments use no geometric transformations, or affine photometric transformations such as rotation, scaling, and left-right flipping.

Additionally, they separately run the categorization and segmentation algorithms and multiply the distributions with image-level prior (ILP) to emphasize the likely categories and to discourage unlikely categories using the results of image categorization. However, we exclude the ILP of image categorization results for all experiments. Nevertheless, across the whole dataset under the same experimental conditions, the proposed method achieved a class average performance of 59.8%, which is better than the 58.6% that was obtained using the state-of-the-art method.

6 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 information into the textonization process. By extending random forests into multi-scale texton forests, our method generates different textons in scale. Then, using the scene-context scale, it 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 will be combined with texture-layout filters [24] to improve segmentation accuracy further.

Acknowledgment

This work was supported by JST, CREST, and JSPS KAKENHI Grant (No. 23500237).

References

[1] S. Battiato, G. Farinella, G. Gallo, and D. Ravi. Spatial hierarchy of textons distributions for scene classification. In *Proc. of the 15th International Multimedia Modeling Conference, Sophia-Antipolis, France, LNCS*, volume 5371, pages 333–343. Springer-Verlag, January 2009.

[2] A. Bosch, A. Zisermann, and X. Munoz. Image classification using random forests and ferns. In *Proc. of the International Conference on Computer Vision*, pages 1–8. IEEE, June 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. `http://www.csie.ntu.edu.tw/libsvm`, 2001.

[5] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. *Workshop on statistical learning in computer vision, ECCV*, 1(1-22):1–2, 2004.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. of the Computer Vision and Pattern Recognition*, pages 886–893. IEEE, June 2005.

[7] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert. An empirical study of context in object detection. In *Proc. of the Computer Vision and Pattern Recognition*, pages 1271–1278. IEEE, June 2009.

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

[9] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proc. of the Computer Vision and Pattern Recognition*, pages 524–531. IEEE, June 2005.

[10] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *Proc. of the Computer Vision and Pattern Recognition*, pages 1458–1465. IEEE, June 2005.

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

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

[13] Y. Kang, H. Nagahashi, and A. Sugimoto. Image categorization using scene-context scale based on random forests. *IEICE Transactions on Information and Systems*, E94-D(9):1809–1816, 2011.

[14] J. Kivinen, E. B. Sudderth, and M. I. Jordan. Learning multiscale representations of natural scenes using dirichlet processes. In *Proc. of the International Conference on Computer Vision*, pages 1–8. IEEE, October 2007.

[15] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. of the Computer Vision and Pattern Recognition*, pages 2169–2178. IEEE, June 2006.

[16] V. Lepetit, P. Lagger, and P. Fua. Randomized trees for real-time keypoint recognition. In *Proc. of the Computer Vision and Pattern Recognition*, pages 775–781. IEEE, June 2005.

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

[18] J. Malik, S. Belongie, T. Leung, , and J. Shi. Contour and texture analysis for image segmentation. *Int. Journal of Computer Vision*, 43:7–27, 2001.

[19] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *Proc. of the Computer Vision and Pattern Recognition*, pages 257–263. IEEE, June 2003.

[20] F. Moosmann, B. Triggs, and F. Jurie. Fast discriminative visual codebooks using randomized clustering forests. In B. Scholkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 985–992. MIT Press, 2006.

[21] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *Proc. of the European Conf. on Computer Vision, Graz, Austria, LNCS*, volume 3954, pages 490–503. Springer-Verlag, May 2006.

[22] J. Shotton, M. Johnson, , and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *Proc. of the Computer Vision and Pattern Recognition*, pages 1–8. IEEE, June 2008.

[23] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling formulti-class object recognition and segmentation. In *Proc. of the European Conf. on Computer Vision, Graz, Austria, LNCS*, volume 3951, pages 1–15. Springer-Verlag, May 2006.

[24] 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.

[25] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc. of the International Conference on Computer Vision*, pages 1470–1447. IEEE, June 2003.

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

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

[28] C. Wallraven, B. Caputo, and A. Graf. Recognition with local features: the kernel recipe. In *Proc. of the Computer Vision and Pattern Recognition*, pages 257–264. IEEE, June 2003.

[29] G. Wang, Y. Zhang, and L. Fei-Fei. Using dependent regions for object categorization in a generative framework. In *Proc. of the Computer Vision and Pattern Recognition*, pages 1597–1604. IEEE, June 2006.

[30] S. Wang and Y. Wang. A multi-scale learning framework for visual categorization. In *Proc. of the 12th Asian Conference on Computer Vision, Queenstown, New Zealand, LNCS*, volume 6492, pages 310–322. Springer-Verlag, November 2011.

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

[32] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *Proc. of the International Conference on Computer Vision*, pages 1800–1807. IEEE, October 2005.

[33] C. Wojek and B. Schiele. A dynamic conditional random field model for joint labeling of object and scene classes. In *Proc. of the 10th European Conference on Computer Vision, Marseille, France, LNCS*, volume 5305, pages 733–747. Springer-Verlag, October 2008.

[34] 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*, 73:213–238, 2007.

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

Author Biography

Yousun Kang received a Ph.D. degree from Tokyo Institute of Technology in 2010. She worked with Toyota Central R&D LABS., Inc. for three years from 2007. During 2010–2011, she was a researcher in the National Institute of Informatics, Japan. She is currently an Associate Professor at Tokyo Polytechnic University. Her research interests include texture analysis, scene understanding, pattern recognition, image processing, and computer vision. She is a member of the RSJ and IEICE of Japan.

Akihiro Sugimoto received his B.S, M.S, and Dr. Eng. degrees in Mathematical Engineering from The University of Tokyo in 1987, 1989, and 1996, respectively. After working at Hitachi Advanced Research Laboratory, ATR, and Kyoto University, he joined the National Institute of Informatics, Japan, where he is currently a professor. During 2006–2007, he was a visiting professor at ESIEE, France. He received a Paper Award from the Information Processing Society in 2001. He is a member of IEEE. He is interested in mathematical methods in engineering. Particularly his current main research interests include discrete mathematics, approximation algorithm, vision geometry, and modeling of human vision.