PAPER

Image Categorization Using Scene-Context Scale Based on Random Forests

Yousun KANG[†a)], **Hiroshi NAGAHASHI**[††], ***and*** **Akihiro SUGIMOTO**[†††], ***Members***

SUMMARY Scene-context plays an important role in scene analysis and object recognition. Among various sources of scene-context, we focus on scene-context scale, which means the effective scale of local context to classify an image pixel in a scene. This paper presents random forests based image categorization using the scene-context scale. The proposed method uses random forests, which are ensembles of randomized decision trees. Since the random forests are extremely fast in both training and testing, it is possible to perform classification, clustering and regression in real time. We train multi-scale texton forests which efficiently provide both a hierarchical clustering into semantic textons and local classification in various scale levels. The scene-context scale can be estimated by the entropy of the leaf node in the multi-scale texton forests. For image categorization, we combine the classified category distributions in each scale and the estimated scene-context scale. We evaluate on the MSRC21 segmentation dataset and find that the use of the scene-context scale improves image categorization performance. Our results have outperformed the state-of-the-art in image categorization accuracy.

key words: *scene-context scale, image categorization, randomized decision trees, random forests, multi-scale texton forests*

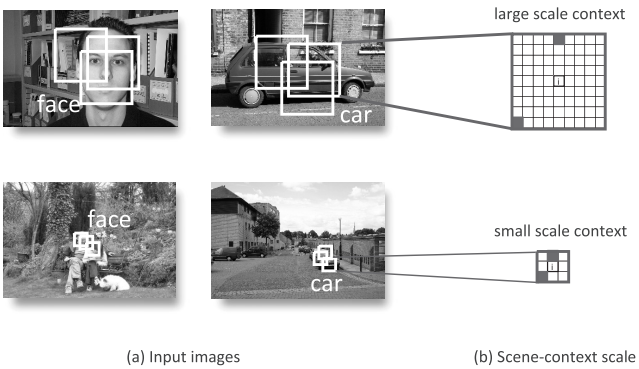

Fig. 1 The example images with different scene-context scale. The scene-context scales of an object (faces or car) strongly differs in each image.

1. Introduction

When people wish to find an image in large database such as web albums, it is important to categorize an image not only simply but as quickly and accurately as possible. As more and more digital images are shared online by individual users, automatically categorizing an image has become one of the most important tasks. Popular search engines have begun to provide tags based on simple characteristics of images, but they still have difficulty in automatically assigning a set of text labels to an image based on its visual content.

Image categorization is to determine, for a given image, the category to which the image belongs (e.g. dog images, beach images, indoor images). Image categorization is one way in which we can perform image retrieval, and it can be used to inform other tasks, such as semantic segmentation or object detection. For example, an image retrieval system will become easy to use if semantic categories and keywords for an image are provided. Furthermore, image categorization can enhance the understanding of visual content for easy browsing in the website.

Prior categorization frameworks have tackled the problems of extracting features as image cues [1], or combining features [2] to improve the performance. Recently not a few results have shown that the dense sampling of visual words and their combinations with image cues can improve categorization performance significantly [3]. They were developed in such a way that visual words are integrated into the bag-of-words model for learning of each category. In this paper, we propose a new framework for image categorization that exploits new context information, namely scene-context scale, and incorporate it into the classified category distributions.

Computer vision approaches have demonstrated that the use of context improves recognition performance [4]–[8]. While the term *context* is frequently used in the literature as one of important keywords, it is difficult to give its clear definition. There are many sources of context, and thus numerous psychophysical studies that have presented new theories on context for human object recognition [9]–[11]. As a result, a specific agreement has been achieved in the community that scene-context plays an important role in scene categorization and object recognition.

When the scene-context is used on a per-pixel level, we can capture the local context where image pixels within a region of interest carry useful information. Some image pixels/patches have ambiguous features at a very local scale, because the color and texture of the local level are insufficient to identify the pixel-class. The more the region of

Manuscript received December 8, 2010.
Manuscript revised April 28, 2011.

[†]The authors is with the Department of Applied Computer Science, Tokyo Polytechnic University, Atsugi-shi, 243–0297 Japan.
[††]The author is with the Imaging Science and Engineering Laboratory, Tokyo Institute of Technology, Yokohama-shi, 226–8503 Japan.
[†††]The author is with National Institute of Informatics, Tokyo, 101–8430 Japan.
a) E-mail: yskang@cs.t-kougei.ac.jp

DOI: 10.1587/transinf.E94.D.1809

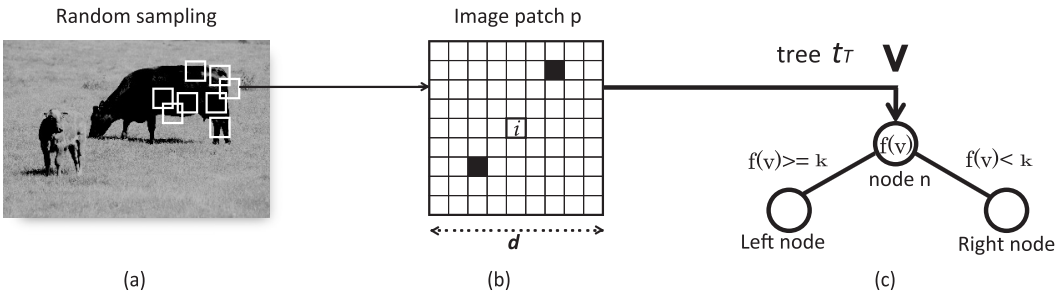

Fig. 2 (a) Image patches. The random subset of the training data is composed of image patches sampled from input image. (b) A region of interest. An image patch p includes raw image pixels within a $(d \times d)$ window. (c) Node branching. The split nodes n of decision trees t_T is branched into left or right node by computing simple functions f and threshold κ with feature vector $\mathbf{v}$.

interest increases, the more it includes the neighborhoods of pixels. Therefore, increasing the size of a region of interest is one of the common methods to include valid local context [12].

The size of a region of interest depends on an object in a scene. Given object presence and location, its scale or the relative size in a scene can be a significant cue for recognizing the object in the scene. We refer to this scale as the scene-context scale. We focus in this work on the scene-context scale that is present in a scene, but rarely used as context to improve the recognition performance. Various scene-context scales of images are illustrated in Fig. 1. The size of an object (face or car) strongly differs in each image. When the object is recognized in a scene, the scene-context scale should be considered to improve the recognition performance.

There are several possible sources to estimate the scene-context scale in an image. If the actual scale of objects within an image is provided, or the absolute distance between the observer and a scene can be measured, we may easily estimate the scene-context scale in each image. Torralba and Oliva inferred the scene scale and estimate the absolute depth in the image [13]. Saxena et al. presented an algorithm for predicting depth from a single still image [14]. They dealt with the scale problem in a scene, however, they did not use scale information as a cue to recognize the object in a scene. We motivate the scale of an object in a scene to be performed as an important cue for categorization and segmentation.

On the other hand, *textons* have been proven effective in categorizing materials [15] as well as generic object classes [16]. The term texton means a compact representation for the range of different appearances of an object. The collection of textons are clustered to produce a codebook of visual words in bag-of-textons model. Recently, textonization process is performed using random forests to generate semantic textons. Random forests are powerful tools with high computational efficiency in vision applications [17].

In this paper, we estimate the scene-context scale using multiple random forests. We propose multi-scale texton forests, which can generate different textons according to scale space. For categorization and segmentation, Shotton et al. [18] proposed semantic texton forests as efficient texton codebooks without using of scene-context scale. We investigate how scene-context scale combines with multi-scale texton forests to show the accuracy improvement of categorization.

To assess the utility of multi-scale texton forests and the scene-context scale, we compare the clustering and classification accuracy and the categorization accuracy with that of the state-of-the-art [18]. The results show that our method achieves better classification and categorization accuracy than those of the state-of-the-art that is without using of scene-context scale.

This paper is organized as follows: Sect. 2 explains the multi-scale texton forests in detail. Section 3 describes how to combine the scene-context scale into the image categorization module. Section 4 shows experimental results on performance and our conclusions are presented in the final section.

2. Scale Extension Based on Random Forests

The textons facilitates dense representation for visual words in bag-of-word model. When the textonization process is performed on randomized decision forests, there are many advantages in the framework. Firstly, the texton codebooks are available without computing expensive filter-banks or descriptors. There is no need to be time-consuming clustering method and nearest-neighbor assignment using k-means. Secondly, using the learned randomized decision forests, we can simultaneously exploit both hierarchical clustering and classification with category distribution. In this section, we explain how to generate semantic textons according to various scale-level using randomized decision forests.

The randomized decision forest, which is simply called random forest, is a machine learning technique which can be used to categorize individual pixels of an image [19]. We perform textonization process using random forests to formulate multi-scale texton forests. We employ the semantic texton forest proposed by Shotton et al. [18] and extend its concept to obtain multi-scale texton forests.

Each data for each node of random forest indicates an image patch **p** as shown in Fig. 2 (a). The image patch **p** has

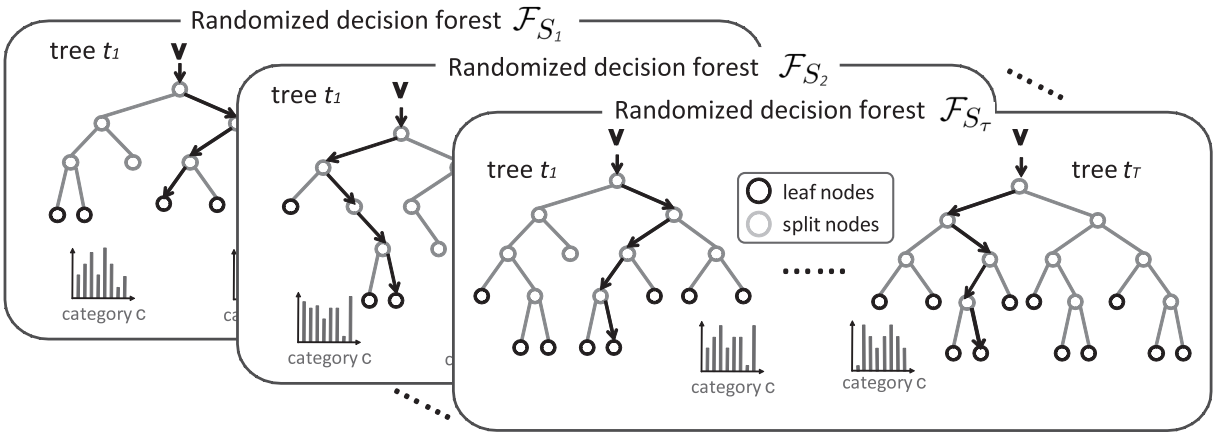

Fig. 3 Multi-scale texton forest. The multi-scale texton forest consists of several random forests with various scale space and each random forest consists of many decision trees with same scale level.

the size of (d×d) image pixels as shown in Fig. 2 (b). At each node, a simple test is performed as shown in Fig. 2 (c), and the result of that test is used to determine which child node to choose in a decision tree. In a decision tree, the recursive node branching continues to the maximum depth or until no further information gain is possible [21]. We employed a depth-first manner, which recursively splits nodes until a maximum depth is reached.

To formulate multi-scale texton forests, we employ the method to expand scale level of random forests that is to be increased in size of image patches **p** for split functions. Each random forest has its own scale level and its scale level can be expanded by increasing the region of interest in multi-scale texton forests. The effective region size for local context can be chosen among the multi-scale texton forests with different scale.

Multi-scale texton forests are random forests created in different scale space for textonization of an image. The multi-scale texton forests consist of several random forests $\mathcal{F}_S$ with various scale space $\mathcal{S} = (S_1, \ldots, S_\tau)$. As shown in Fig. 3, a random forest $\mathcal{F}_S$ is a combination of T decision trees at each scale space S_k, where k is the level of scale ($k = 1, \ldots, \tau$). The nodes in the trees efficiently provide a hierarchical clustering into semantic textons with scale-contextual features.

The split nodes in multi-scale texton forests use split functions of image pixels within a region of interest. Each random forest $\mathcal{F}_{S_k}$ has a different set of pixel combinations within a region of interest as shown in Fig. 2 (b). We can increase the scale level k of a random forest by dilatation of a region of interest.

At the first scale space S_1, the region of interest R_{S_1} covers whole pixels within a $(d \times d)$ image patch, on which the split functions in $\mathcal{F}_{S_1}$ act. In next scale space S_2, the region of interest R_{S_2} deals with the pixels within the difference of $(dk \times dk)$ image patch from the region R_{S_1} of a previous scale space S_1. Therefore, the region of interest R_{S_k} increases within a $(dk \times dk) - (d(k-1) \times d(k-1))$ image patch as illustrated in Fig. 4. The number of possible combinations of selecting two pixels inside a region of interest also increases quadratically with respect to the scale level k.

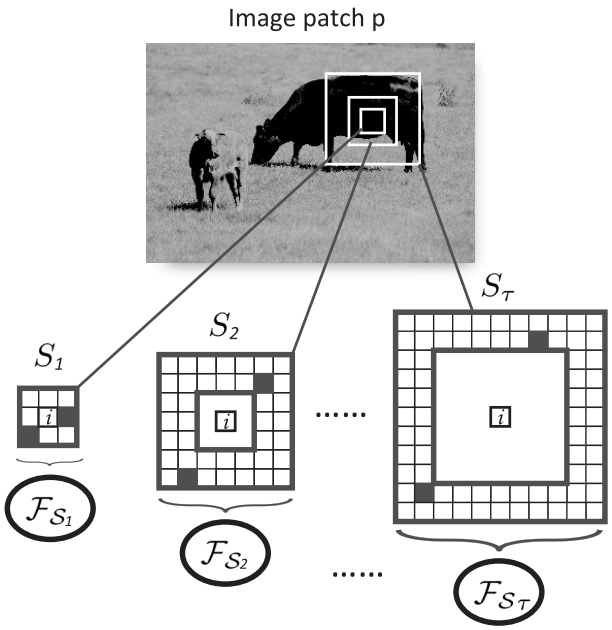

Fig. 4 Dilatation of a region of interest according to scale space S_k. Various sizes of a region of interest are used for node split function in the multi-scale texton forests.

To textonize an image according to scale space, image patches centered at each pixel with various sizes are passed down the multi-scale texton forests. Each random forest consists of many leaf nodes $L = (l_1, \ldots, l_T)$ and we can compute the averaged class distribution in the leaf node. The class distribution is $\mathcal{F}_S\{p(c|L)\}$, where c is a category. The textons generated by each random forest can be extracted in different scales from other forests. By pooling the statistics of semantic textons and class distributions over an image region, the bag of textons presents a powerful feature for image categorization.

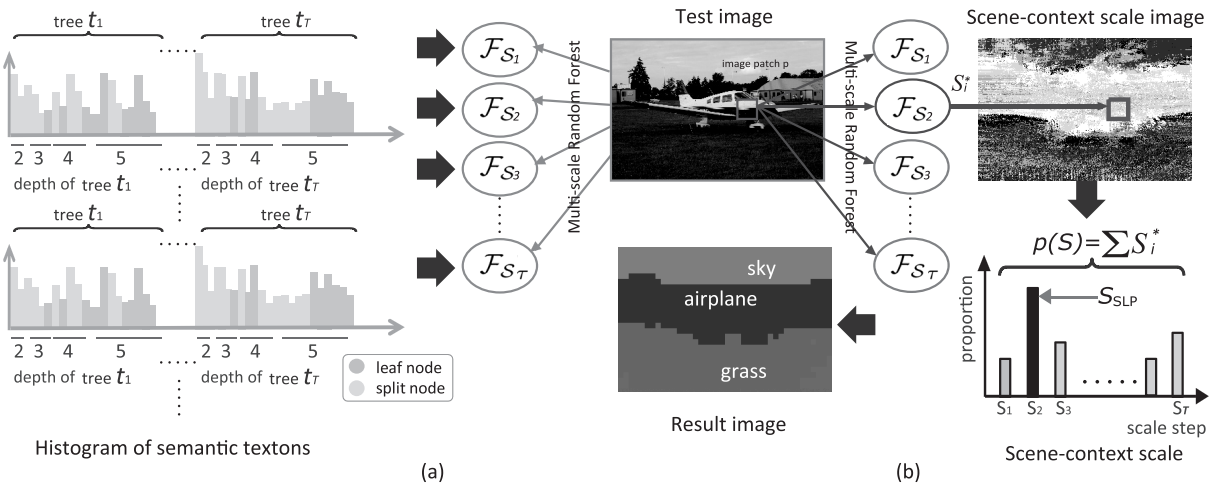

Fig. 5 (a) The histogram of the bag of textons. The histogram contains both leaf nodes and split nodes. The depth of the nodes shows in histogram of each random forest. (b) The scene-context scale of an image pixel can be estimated by computing the minimum entropy of each image patch p. Darker pixels correspond to smaller scale, so white pixels represent the largest scale S_τ.

3. Image Categorization Using Scene-Context Scale

In this section, we explain how to estimate the scene-context scale of each image pixel using multi-scale texton forests. To obtain classified category distributions, we adopt the pyramid match kernel and non-linear support vector machine (SVM) in each scale. Finally, we combine the estimated scene-context scale and classified category distributions in the end of this section.

3.1 Scene-Context Scale and Its Estimation

Scene-context scale is the effective scale of local context to classify an image pixel in a scene. Since the scale of objects strongly differs in each input image, the use of the scene-context scale improves image categorization performance.

The scene-context scale of each image pixel is obtained by computing the entropies of an image patch in the leaf nodes of each random forest. Since the objects in various size and background/foreground appear together in the image, we should compute scene-context scale per pixel. The confidence of each random forest is computed as the entropies of the class label distribution in leaf nodes. We regard the confidence as the criterion of an optimal scale level to be chosen. At each image pixel, therefore, one scale level with minimum entropy is chosen as the scene-context scale among the multi-scale texton forests.

At first, we compute the entropy $E(\mathbf{p}|L)$ of each image patch $\mathbf{p}$ at leaf nodes L of a random forest as

$$E(\mathbf{p}|L) = -P(c|L) \times \log P(c|L). \tag{1}$$

The entropy $E(\mathbf{p}|L)$ can be computed in each random forest $\mathcal{F}_{S_k}$ with scale level $k = (1, \ldots, \tau)$ and we note the entropy of a random forest as $\mathcal{F}_{S_k}\{E(\mathbf{p}|L)\}$. Among the whole scale space $\mathcal{S} = (S_1, \ldots, S_\tau)$, only one scale space S_i is chosen that contains the leaf nodes of a random forest $\mathcal{F}_{S_i}$ with minimum entropy as

$$S_i^* = \arg\min_{S_i}(\mathcal{F}_{S_i}\{E(\mathbf{p}|L)\}). \tag{2}$$

The scene-context scale of an image pixel is the instance S_i^* of the best scale space of image patches as shown in Fig. 5 (b). Now, we can estimate the scene-context scale in an image as the proportion of the instances of scale space S_k^* of image pixels. This gives the distribution of scale space $P(\mathcal{S})$ in input image as

$$P(\mathcal{S}) = \sum_{\mathbf{p}} S_i^*. \tag{3}$$

We can also determine the Scale-Level Prior [18] that is the most likely scale space S_{SLP}^* in whole image like as

$$S_{SLP}^* = \arg\max_i S_i^*. \tag{4}$$

3.2 Integration of Scene-Context Scale into Classifier for Categorization

We use a bag of textons model [22] computed across the whole image for image categorization. The bag of textons model can construct the histogram of semantic textons and calculate the node prior distributions over the whole image, even discarding spatial layout. A histogram consists of the nodes of each random forest, containing both leaf nodes and split nodes as shown in the left histogram of Fig. 5 (a). Because the hierarchical clusters are better than the leaf node clusters alone, we use both of leaf and split nodes for constructing histograms. The histogram of each random forest is used as an input of a classifier for recognizing object categories.

For a classifier for categorization, we use a non-linear

support vector machine (SVM). The non-linear SVM depends on a kernel function, which measures similarity between input images. Grauman et al. [23] proposed a pyramid match kernel over unordered feature sets that allows them to be used effectively and efficiently in kernel-based learning methods. We employ the pyramid match kernel to efficiently compute the approximation of global correspondence between sets of features in two images. Therefore, the first categorization process for each random forest we use is very similar to those in [18], but our classifier is different from that in [18] because ours involves scene-context scale to include scale information of object in a scene.

We build a 1-vs-others SVM classifier, which gives probability of every class per each image. The probability can be computed per each random forest as $P_{\mathcal{F}_S}(c|I)$, where I indicates the whole image. At scene-context scale S_i^*, the category distributions $\mathcal{F}_{S_i}\{p(c|I)\}$ are also available for local classification and consist in the histogram of a bag-of-textons model.

For each test image, therefore, we estimate the scene-context scale and we combine the output of SVM categorization algorithm with it. The categorization performance increases by multiplying the distributions of each category $P(c|\mathcal{F}_S)$ and of scene-context scale $P(\mathcal{S})$ as

$$P'(c|\mathcal{F}_S) = \sum_{k=1}^{\tau} P(c|\mathcal{F}_{S_k}) \times P(S_k). \tag{5}$$

And the SLP is used to emphasize likely categories and discourage unlikely categories, by multiplying the average distribution of the multi-scale texton forests and the distributions at SLP as

$$P'(c|\mathcal{F}_S) = \left(\frac{l}{\tau} \sum_{k=1}^{\tau} P(c|\mathcal{F}_{S_k}) \right) \times P(S_{SLP})^{\alpha} \tag{6}$$

using parameter α to soften the prior.

4. Experimental Results

This section presents our experimental results for image categorization using multi-scale texton forests. To assess the utility of the scene-context scale and multi-scale texton forests in image categorization, we compare the classification accuracy with that of conventional semantic texton forests method [18] without using the scene-context scale.

We evaluate our algorithm using challenging MSRC (Microsoft Research Cambridge) segmentation dataset [25]. The MSRC database is composed of 591 photographs of the 21 objects. The dataset includes 21 object classes such as building, grass, tree, cow, sheep, sky, airplane, water, face, car, bicycle, flower, sign, bird, book, chair, road, cat, dog, body, boat. Note that the ground-truth labeling of the 21-class database contains pixels labeled as 'void'. Void pixels are ignored for both training and testing. We used the 45% training and 10% validation data for training, and the hand-labeled ground truth to train the classifiers. The remaining 45% images were used for test data.

Before presenting categorization accuracy, let us show the clustering and classification results using the multi-scale texton forests. The multi-scale texton forests provide both a hierarchical clustering into semantic textons and local classification in various scale space. We separately train the forests in different scale space.

To train the multi-scale texton forest, we prepared six scale steps $\mathcal{S} = (S_1, \ldots, S_6)$ and an initial image patch size is (15×15). Therefore, the size of image patches for split function is $(15k \times 15k)$ at each scale step S_k. A random forest $\mathcal{F}_S$ has the following parameters: $T = 5$ trees, maximum depth $D = 10$, $500k$ feature tests and 10 threshold tests per split, and 0.25 of the data per tree, resulting in approximately 500 leaves per tree. Training a semantic texton forest took approximately 30×2^k minutes on MSRC dataset at each scale step k, however, testing an image took 0.1 second per a semantic texton forest.

At test time, the most likely class in the averaged category distribution gives the clustering and classification results for each pixel as shown in Fig. 6. Clustering and local 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). Figure 7 shows the results of the clustering and local classification based on scene-context scale. We estimate the scene-context scale per image pixel using multi-scale texton forests as shown in the third row of Fig. 7. Since each image pixel has the category distribution at the scene-context scale, we can infer the most likely category $c_i^* = \arg\max_{c_i} P(c_i|L)$ of leaf nodes $L = (l_1, \ldots, l_T)$ for each pixel i as shown in Fig. 7 (a). On the other hand, Fig. 7 (b) shows the results of the state-of-the-art [18] without using scene-context scale based on single-scale semantic texton forests. The single-scale semantic texton forests used the same parameter of the multi-scale texton forests with the first scale level $\mathcal{F}_{S_1}$.

As shown in Fig. 8, a pixel level classification based on the local distributions $P(c|L)$ gives poor, but still good performance. The global classification accuracy without scene-context scale gives 50.2% and the result with using scene-context scale based on multi-scale texton forests gives 53.0%. In particular, significant improvement can be observed in most of the classes except some classes: tree, water, car, bicycle, sign and road. It should seem that they have not influence on scene-context scale. Across the whole MSRC dataset, using the scene-context scale achieved a class average performance of 48.3%, which is better than the 38.4% of (b) as shown in the table of Fig. 8. Therefore, we can see that the proposed scene-context scale can be powerful and effective context information for category classification and clustering.

In Fig. 9, we plot the class average accuracy against the number of scene-context scale. The number of scene-context scale corresponds to the number of semantic texton forests according to scale step k. A noticeable improvement is obtained until the scale step k is 5. From the Fig. 9, we can see that class average increases with more scene-context

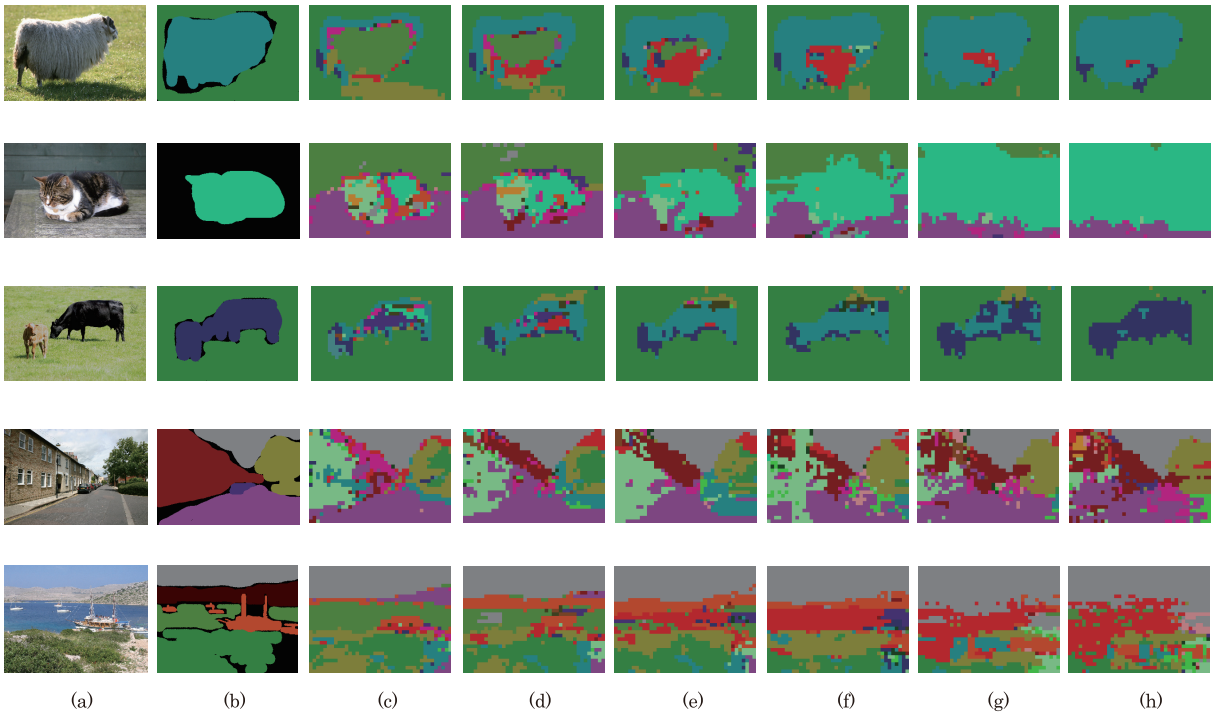

Fig. 6 Clustering and classification results using the multi-scale texton forest. The multi-scale texton forest can generate the different textons according to scale steps. (a) Input images. (b) Ground-truth images. (c) - (h) Clustering results according to scale space $\mathcal{S} = (S_1, \ldots, S_6)$. The results correspond to each scale space such as S_1 = (c), S_2 = (d), S_3 = (e), S_4 = (f), S_5 = (g), and S_6 = (h).

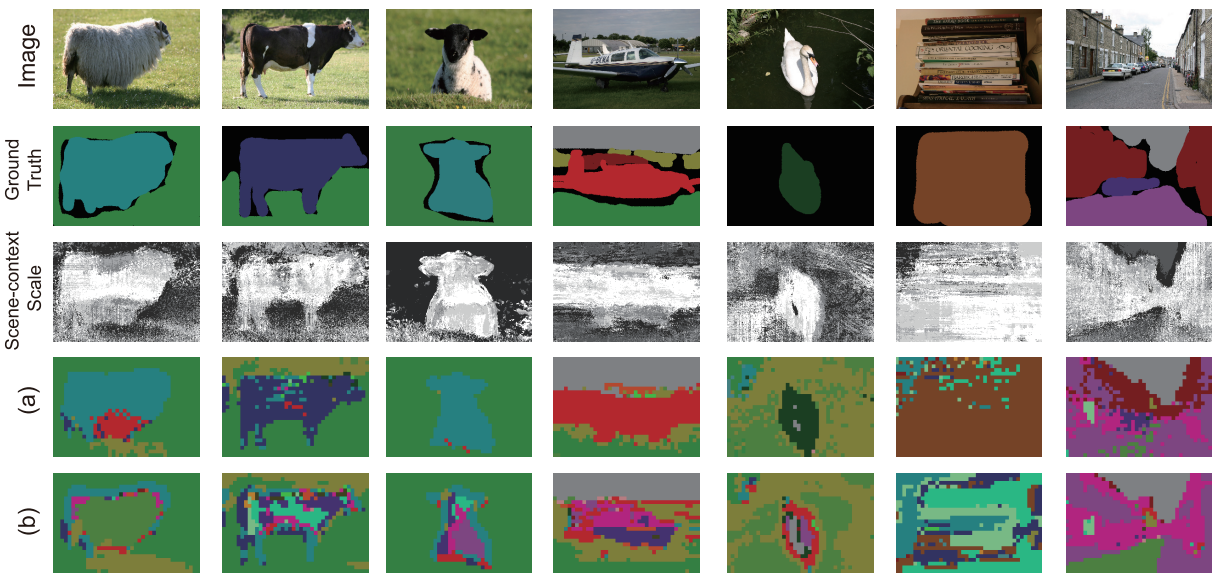

Fig. 7 Clustering and classification results using scene-context scale. (a) Classification result with using scene-context scale based on multi-scale texton forests. (b) Classification result without using scene-context scale based on single-scale semantic texton forests [18]

scale.

The pixel level classification based on the local distributions gives different results according to each scale step. Figure 6 shows the clustering and classification results of each random forest $\mathcal{F}_{\mathcal{S}_i}$ with different scales. As can be seen, each image has the best accuracy according to category in some scene-context scale. Therefore, there is the scene-context scale in not each image but each image patch. Using the multi-scale texton forest, we find the scene-context scale per image patch in a test image. By multiplying the distributions of each category and the proportion of the scene-context scale in the test image, we can finally improve the

	building	grass	tree	cow	sheep	sky	aeroplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat	global	class average
(a)	12	90	43	60	79	90	89	36	89	28	34	65	19	6	64	14	46	42	22	36	49	53.0	48.3
(b)	2	89	49	43	39	84	36	45	74	30	58	66	28	3	17	6	57	19	14	17	28	50.2	38.4

Fig. 8 Clustering and classification results on MSRC datasets. Classification accuracies (percent) over the whole dataset, without the scene-context scale (b), and with the scene-context scale (a). Our new highly efficient scene-context scale achieve a significant improvement on previous work (b).

	building	grass	tree	cow	sheep	sky	aeroplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat	class average
(a) None	64	86	75	**86**	**92**	90	74	66	64	**88**	72	84	70	**53**	90	**67**	67	57	**36**	64	**77**	72.8
(b) SLP	63	93	**81**	82	73	**97**	64	**75**	**84**	61	**73**	**85**	**72**	49	93	56	**77**	57	35	**84**	64	72.2
(c) Mean	**71**	88	80	83	77	95	**87**	70	73	86	71	83	67	**53**	**94**	62	71	59	33	74	70	73.7
(d) Distribution	64	**94**	79	84	82	**97**	83	**75**	**84**	79	**73**	**85**	59	49	91	58	75	**68**	32	**84**	**77**	**74.9**

Fig. 10 Image categorization results on MSRC datasets. Categorization accuracies (percent) over the whole dataset. Scene-context scale achieves a improvement on previous work.

Fig. 9 Global accuracy vs number of scene-context scale.

learned per-category distribution.

We obtained the categorization accuracy as shown in Fig. 10 (a) without using the scene-context scale, and (b), (c) and (d) with using scene-context scale. (a) None in the first row of the table used only one scale space, as the previous work [18]. (b) SLP in the second row of the table used the Eq. (11) in Sect. 4. (c) Mean in the third row of the table used the average of categorization accuracies over the whole randomized decision forests in the multi-scale texton forests. (d) Distribution in the forth row of the table used the proportion of the scene-context scale in a test image as like Eq. (10).

The proposed method (d) using the distribution of scene-context scale gives better results than any other methods without using scene-context scale. Across the whole challenging dataset, using the distribution of scene-context scale achieved a class average performance of 74.9%, which is better than all the 72.8% of (a), the 72.2 % of (b), and the 73.7 % of (c). The proposed method improves performance for all but three classes. This is probably because the proportion of the scene-context scale is inappropriate for the objects in an image such as dog and bird, therefore poor at categorizing the objects. In addition, the three classes also have low performance in clustering and classification process with low class average such as dog (22%), bird (6%), and chair (14%). We should devote to estimate more accurate scene-context scale and to generate more discriminate texton for various objects in future works. In particular, significant improvement can be observed difficult classes: grass and cat.

5. Conclusion

This paper presented a new framework for image categorization using multi-scale texton forest and scene-context scale. We have (i) introduced the concept of scene-context scale in object recognition, (ii) expanded the random forests to multi-scale texton forests, and (iii) achieved efficient categorizing by using a combination of scene-context scale and multi-scale texton forest. In experiments, we confirmed that the proposed method using the scene-context scale gives better results than any other methods without using scene-context scale. The multi-scale texton forest can be utilized in semantic segmentation and object recognition by integrating scene-context scale with bag of textons method.

In future work, we improve accuracy per-category distribution by using geometric transformations and affine photometric transformations on training/test dataset. In addition, the results of image categorization are utilized as region priors for object recognition and semantic segmentation.

Acknowledgment

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

References

[1] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2003.
[2] G. Csurka, C.R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," ECCV International Workshop on Statistical Learning in Computer Vision, 2004.
[3] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," Proc. European Conf. Computer Vision, 2006.
[4] C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-ocurrence, location and appearance," Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2008.
[5] G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," Proc. European Conf. Computer Vision, 2008.
[6] A. Torralba, "Contextual priming for object detection," Int. J. Comput. Vis., vol.53, no.2, pp.169–191, 2003.
[7] D. Hoiem, A.A. Efros, and M. Hebert, "Putting objects in perspective," Int. J. Comput. Vis., vol.80, no.1, pp.3–15, 2008.
[8] K. Murphy, A. Torralba, and W.T. Freeman, "Using the forest to see the trees: A graphical model relating features, objects and scenes," Proc. NIPS, MIT Press, 2003.
[9] A. Oliva and A. Torralba, "The role of context in object recognition," Trends Cogn Sci, November, 2007.
[10] P. Carbonetto, N. de Freitas, and K. Barnard, "A statistical model for general contextual object recognition," Proc. European Conf. Computer Vision, 2004.
[11] L. Wolf and S. Bileschi, "A critical view of context," Int. J. Comput. Vis., vol.69, no.2, pp.251–261, 2006.
[12] S. Divvala, D. Hoiem, J. Hays, A. Efros, and M. Hebert, "An empirical study of context in object detection," Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2009.
[13] A. Torralba and A. Oliva, "Depth estimation from image structure," IEEE Trans. Pattern Anal. Mach. Intell., vol.24, no.9, pp.1226–1238, 2003.
[14] A. Saxena and S. Chung and A, "Ng. 3-D depth reconstruction from a single still image," Int. J. Comput. Vis., vol.76, no.1, pp.53–69, 2008.
[15] M. Varma and A. Zisserman, "A statistical approach to texture classification from single images," Int. J. Comput. Vis., vol.62, no.1, pp.61–81, 2005.
[16] J. Winn, A. Criminisi, and T. Minka, "Categorization by learned universal visual dictionary," Proc. Int. Conf. Computer Vision, pp.2:1800–1807, 2005.
[17] L. Breiman, "Random forests," Mach. Learn., vol.45, no.1, pp.5–32, 2001.
[18] J. Shotton, M. Johnson, and R. Cipolla, "Semantic Texton Forests for image categorization and segmentation," Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2008.
[19] F. Moosmann, B. Triggs, and F. Jurie, "Fast discriminative visual codebooks using randomized clustering forests," Proc. NIPS, 2006.
[20] V. Lepetit, P. Lagger, and P. Fua, "Randomized trees for real-time keypoint recognition," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.2:775–781, 2005.
[21] G. Rogez, J. Rihan, S. Ramalingam, C. Orrite, and P. Torr, "Randomized trees for human pose detection," Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2008.
[22] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," Int. J. Comput. Vis., vol.81, no.1, pp.2–23, 2009.
[23] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," Proc. Int. Conf. Computer Vision, 2005.
[24] M. Johnson, Semantic Segmentation and Image Search, Phd Thesis, University of Cambridge, 2008.
[25] The Microsoft Research Cambridge 21 Class Database. http://research.microsoft.com/vision/cambridge/recognition/

Yousun Kang received her B.S. and Dr. Eng. degrees from Chosun University in 1993 and 1999 respectively, and Ph.D. degree from Tokyo Institute of Technology in 2010. She worked with TOYOTA CENTRAL R&D LABS., INC. for three years from 2007. From 2010 to 2011, she had been a researcher in the National Institute of Informatics, Japan. She is currently an associate professor in 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.

Hiroshi Nagahashi received his B.S. and Dr. Eng. degrees from Tokyo Institute of Technology in 1975 and 1980, respectively. Since 1990, he has been with Imaging Science and Engineering Laboratory, Tokyo Institute of Technology, where he is currently a professor. His research interests include pattern recognition, computer graphics, image processing, computer vision. Dr. Nagahashi is a member of IEEE Pattern Analysis and Machine Intelligence Society, the Institute of Electrical Engineers of Japan, the Information Processing Society 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. From 2006 to 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. In particular, his current main research interests include discrete mathematics, approximation algorithm, vision geometry, and modeling of human vision.