Semantic Segmentation and Object Recognition using Scene-Context Scale

Yousun Kang[1] Hiroshi Nagahashi[2] Akihiro Sugimoto[1]

[1]National Institute of Informatics
{yskang, sugimoto}@nii.ac.jp

[2]Tokyo Institute of Technology
longb@isl.titech.ac.jp

Abstract

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 region size of local context to classify an image pixel in a scene. This paper presents semantic segmentation and object recognition using scene-context scale. The scene-context scale can be estimated by the entropy of the leaf node in multi-scale texton forests. The multi-scale texton forests efficiently provide both hierarchical clustering into semantic textons and local classification depending on different scale levels. For semantic segmentation, we combine the classified category distributions of scene-context scale with the bag-of-textons model. In our experiments, we use MSRC21 segmentation dataset to assess our segmentation algorithm and show that the usage of the scene-context scale improves recognition performance.

1. Introduction

There are many sources of scene-context, which play an important role in scene analysis and object recognition [3]. When the 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, using the multi-scale features [4] or increasing the size of a region of interest [2] is one of the common methods to include valid local context in computer vision approaches.

In object recognition process, the size of a region of interest means available range to search local context for an image pixel. Given object presence and location in a scene, its scale or relative size in the scene is related to this range and it can be a strong cue for recognizing the objects in the scene. We refer the effective region size for local context as scene-context scale. We focus in this work on the scene-context scale that is present in a scene, but rarely used as a context to improve the recognition performance.

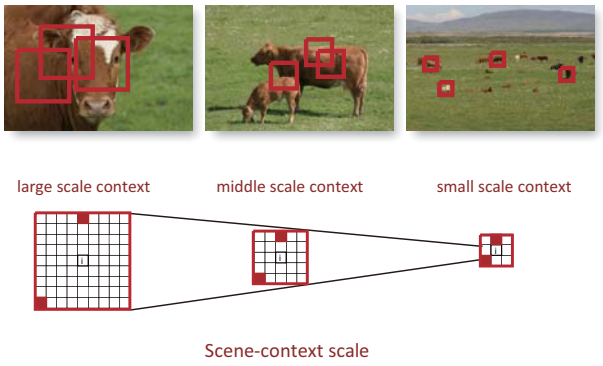


Figure 1. **Three examples of images with different scene-context scale.** The objects strongly differ in their scale in an image. When the object (cow) is recognize in a scene, the scene-context scale should be considered to improve the recognition performance.

The various scene-context scales of images are illustrated in Fig. 1. There are several helpful 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 is measured, we may straightforwardly estimate the scene-context scale in each image. Torralba and Oliva inferred the scene scale and estimate the absolute depth in the image [14]. Saxena *et. al* presented an algorithm for predicting depth from a single still image [10]. They dealt with the scale problem in a scene, however, they did not use the scale information as a cue to recognize the object in a scene.

In this work, we estimate the scene-context scale of objects in a scene using multi-scale texton forests and use the scene-context scale to improve the accuracy of segmentation and recognition. We propose the multi-scale texton forests, which can generate different textons according to scale levels. In addition, we combine the bag-of-textons model [12] with a histogram of the category distribution at scene-context scale for semantic segmentation. The histogram is used as an input to a classifier to recognize object

978-0-7695-4285-0/10 $26.00 © 2010 IEEE
DOI 10.1109/PSIVT.2010.14

categories. At last, we illustrate performance by using the scene-context scale for semantic segmentation and object recognition. To assess the utility of the scene-context scale based on multi-scale texton forests for semantic segmentation, we compare the classification accuracy with that of the state-of-the-art [11]. The results show that our segmentation method achieves better classification accuracy than the methods without using of scene-context scale.

This paper is organized as follows: Section 2 explains the multi-scale texton forests in detail. Section 3 describes the scene-context scale and how to combine the scene-context scale with the semantic segmentation module using bag-of-textons model. Section 4 shows experimental results on performance and our conclusions are presented in the final section.

2. Multi-scale Texton Forests

Textons have been proven effective in categorizing materials [15] as well as generic object classes [16]. Recently, textonization process is performed on random forests to generate semantic texton by Shotton *et.al* [11] for image categorization and segmentation. By using random forests, texton codebooks are available without computing filter-banks or descriptors, and without performing expensive k-means clustering and nearest-neighbor assignment. Therefore, the random forests have the advantage of being extremely fast and high performance.

In this section, we explain multi-scale texton forests using random forests [1]. The scene-context scale can be obtained by using multi-scale texton forests, which consist of several random forests with different scales.

2.1. Textonization using random forests

In general, random forests is an ensemble of randomize T decision trees [8]. A learned class distribution $P(c|n)$ is associated with each node n in the tree, where c is a category label of a pixel. A decision tree works by recursively branching left or right down the tree according to a learned binary function of the feature vector, until a leaf node l is reached [11].

Each tree is trained separately using a small random subset of the training data I. Learning proceeds recursively, splitting the training data I_n at node n into left and right subsets I_l and I_r according to a threshold κ of some split function f of the feature vector $\mathbf{v}$:

$$I_l = \{i \in I_n | f(\mathbf{v}_i) < \kappa\}, \tag{1}$$

$$I_r = I_n \backslash I_l. \tag{2}$$

The split functions f act on small image patches $\mathbf{p}$ of size $(d \times d)$ pixels as shown in Fig. 2. These functions can be computed with simple operations of raw image pixels

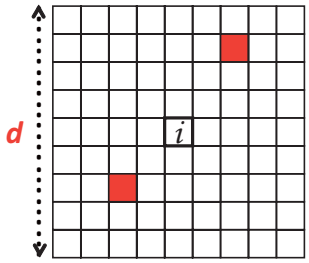


Figure 2. **A region of interest for node split function of randomized decision trees.** The split nodes of decision trees use simple functions of raw image pixels within a $(d \times d)$ image patch.

within a $(d \times d)$ patch: one of the raw value of a single pixel, the sum, difference, and absolute difference of a pair of pixels, namely,

$$\begin{aligned} f(\mathbf{p}) &= p_{x_1,y_1,b_1} \\ f(\mathbf{p}) &= p_{x_1,y_1,b_1} + p_{x_2,y_2,b_2} \\ f(\mathbf{p}) &= p_{x_1,y_1,b_1} - p_{x_2,y_2,b_2} \\ f(\mathbf{p}) &= |p_{x_1,y_1,b_1} - p_{x_2,y_2,b_2}|, \end{aligned}$$

where p is the value of a pixel at (x, y), and b_1 and b_2 are possibly different color channels.

At each split node, several candidates for function f and threshold κ are generated randomly. We compute the expected gain information and choose the one that maximizes in the information about the node categories as follows [7]:

$$\Delta E = -\frac{|I_l|}{|I_n|} E(I_l) - \frac{|I_r|}{|I_n|} E(I_r), \tag{3}$$

where $E(I)$ is the Shannon entropy of the classes in the set of examples I. The recursive training continues to the maximum depth D or until no further information gain is possible. The class distributions $P(c|n)$ are estimated empirically using a histogram of the class labels c_i of the training examples i that reached node n.

A random forest achieves an accurate and robust classification by averaging the class distributions over the leaf nodes of whole trees $L = (l_1, ..., l_T)$:

$$P(c|L) = \frac{1}{T} \sum_{t=1}^{T} P(c|l_t). \tag{4}$$

At this time, the number of training data may be biased towards certain classes in some datasets. To normalize this bias, we adjustment the number of each training data using by weight factor such as the inverse class frequency, $w_i = \xi_{c_i}$. The ξ_c means $(\sum_{i \in I}[c = c_i])^{-1}$.

2.2. Random forests with multi-scale

We increase the size of image patches of split functions to expand scale level of random forests. Each random forests have their own scale level and its scale level

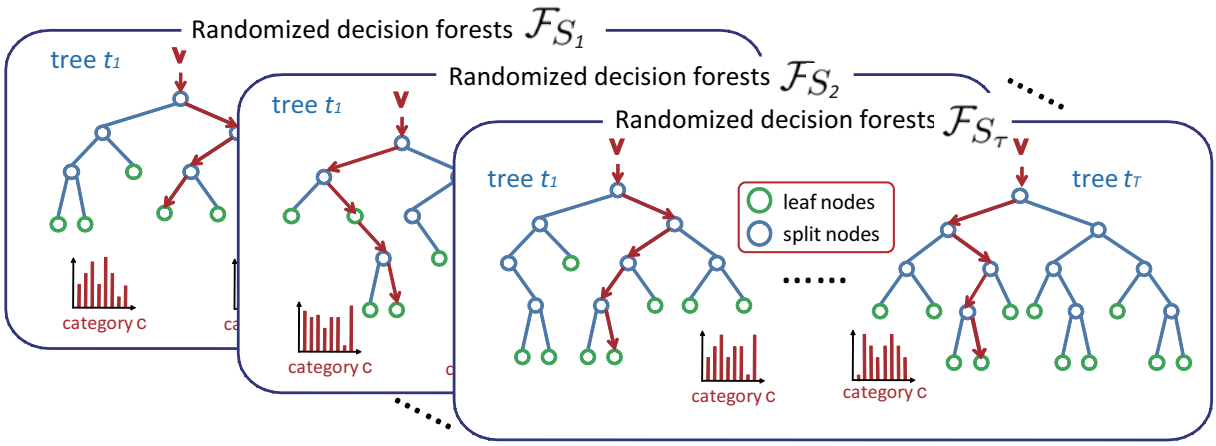


Figure 3. **Multi-scale texton forest.** The multi-scale texton forest consists of several randomized decision forests with various scale space and the randomized decision forest consists of many decision trees at each scale level.

can expand by using 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 randomized decision forests created in different scale space for textonization of an image. The multi-scale texton forests consist of several randomized decision forests $\mathcal{F}_\mathcal{S}$ with various scale space $\mathcal{S} = (S_1, ..., S_\tau)$. As shown in Fig. 3, a random forest $\mathcal{F}_\mathcal{S}$ is a combination of T decision trees at each scale space S_k, where the level of scale space is $k = (1, ..., \tau)$. 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 $\mathcal{F}_\mathcal{S}$ has different set of pixel combinations within a region of interest as shown in Fig. 2. We can increase the scale space $\mathcal{S}$ of a random forest by dilatation of scale of a region of interest.

At the first scale level S_1, the region of interest R_{S_1} covers whole pixels within a $(d \times d)$ image patch, where the split functions f in $\mathcal{F}_{S_1}$ act on. In next scale level 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 level 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 factor k.

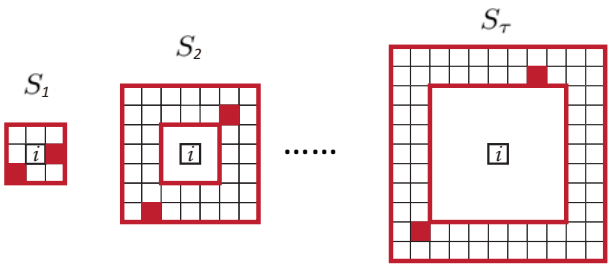


Figure 4. **Dilatation of a region of interest according to scale space S_k.** Various sizes of a region of interest are used for node split function in the multi-scale texton forests.

To textonize an image according to scale levels, image patches centered at each pixel with various size are passed down the multi-scale texton forests resulting in semantic texton leaf nodes $L = (l_1, ..., l_T)$ and the averaged class distribution of each random forest $P_{\mathcal{F}_\mathcal{S}}(c|L)$. The textons generated by each randomized decision forest can be extracted in different scales from the other forests. By pooling the statistics of semantic textons L and distributions $P_{\mathcal{F}_\mathcal{S}}(c|L)$ over an image region, the bag-of-textons presents a powerful feature for semantic segmentation.

3. Segmentation and Recognition

The study of textons facilitates a compact representation for image decomposition and the collection of textons produce a codebook of visual words in bag-of-textons model. The bag-of-textons model treats an object class as an unordered collection of textons and it has the advantage of simplicity and good performance. In this section, we explain how to estimate the scene-context scale in an image pixel and combine it with the bag-of-textons model for segmentation and recognition.

3.1. Scene-context scale

To estimate the scene-context scale of each image pixel, we use the entropies in the leaf nodes of each random forest. Since the confidence of each random forest can be com-

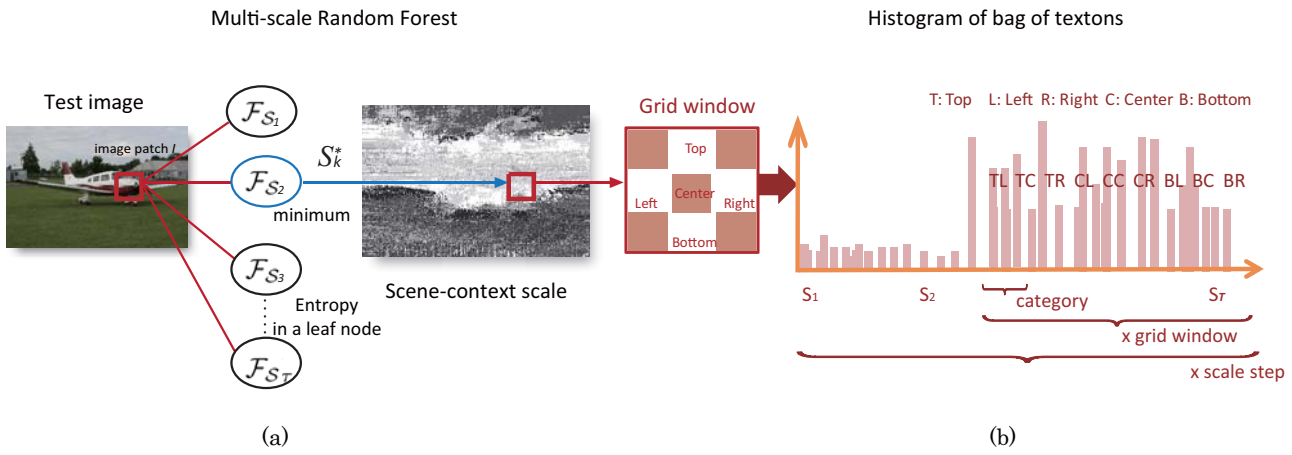


Figure 5. **(a) Scene-context scale of an images.** Darker pixels correspond to smaller scale, so black pixels represent the first scale level S_1 and white pixels represent the largest scale level S_τ. **(b) Histogram of bag of textons model.** The dimension of a histogram is number of grid window times number of category times total scale level

puted as the sum of 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, the scale level of a random forest with minimum entropy of leaf nodes is chosen as the scene-context scale among the multi-scale texton forests.

At first, we compute the entropy $E(I|L_\mathcal{S})$ of each image patch I at leaf nodes $L_\mathcal{S}$ of every random forest $\mathcal{F}_\mathcal{S}$ as

$$E(I|L_\mathcal{S}) = -\sum p(c|L_\mathcal{S}) \times \log_2 p(c|L_\mathcal{S}). \quad (5)$$

The computed entropy $E(I|L_\mathcal{S})$ is summed according to the each forest $\mathcal{F}_\mathcal{S}$. Among the scale level $\mathcal{S} = (S_1, ..., S_\tau)$ of the forest $\mathcal{F}_\mathcal{S}$, the one S_k that contains the leaf node L_{S_k} of a random forest $\mathcal{F}_{S_k}$ with minimum entropy is chosen as

$$S_k^* = \arg\min_{S_k} E(I|L_{S_k}). \quad (6)$$

The scene-context scale of an image pixel is the instances of a scale level S_k^* of image patches as shown in Fig. 5(a). At scene-context scale S_k^*, the category distributions $P(c|L)$ are also available for local classification and consist in the histogram of a bag-of-textons model. The histogram is, therefore, used as input to a classifier to recognize object categories.

3.2. Bag-of-textons model

To demonstrate the efficiency of the scene-context scale for semantic segmentation, we adapt the bag-of-textons algorithm. The goal is to segment an image into coherent regions and simultaneously infer the class label of each region (see Figure 7). We make the histogram of the localized bag-of-textons model using category distributions at the scene-context scale, illustrated in Fig. 5.

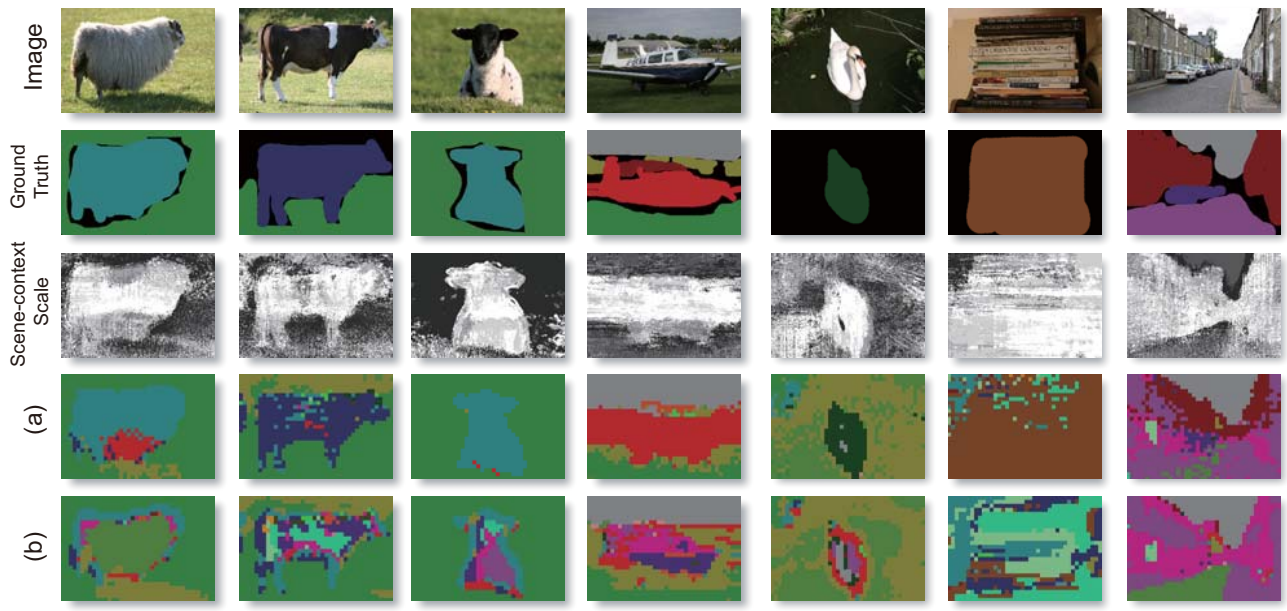
Since the bag-of-textons models discard spatial layout, we use a local grid window as shown in Fig. 5(b). The local grid window consists of nine sub-grid such as Top-Left (TL), Top-Center (TC), Top-Right (TR), Center-Left (CL), Center-Center (CC), Center-Right (CR), Bottom-Left (BL), Bottom-Center (BC), and Bottom-Right (BR). To learn layout and context information automatically, we use class distributions with different scales in a local grid window.

The scene-context scale S_k^* is chosen by using the entropy of class distribution and the class distributions consist in a histogram computed from nine grid windows from top-left (TL) to bottom-right (BR). S_1 are first chosen covering about $(d \times d)$ the pixel area. We concatenated histograms consisting of the class distributions of a scene-context scale among from S_1 to S_τ. Finally, the normalized histogram with multi-scale grid windows is used as a feature vector for object recognition.

Outside the image boundary, there is no contribution to the response. We employ the joint boosting algorithm [13] to select discriminative features of the bag-of-textons model. Random feature selection and subsampling reduce training time to generate several thousand weak learners. The learned strong classifier is an additive model of the form

$$H(c, i) = \sum_{m=1}^{M} h_m(c, i), \quad (7)$$

summing the classification confidence of M weak classifiers. This confidence value can be reinterpreted as a probability distribution using the soft-max transformation [6] to give the energy for optimal labeling.



	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

Figure 6. **Clustering and classification results using scene-context scale.** Above : (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 [11] Below: Classification accuracies (percent) over the whole dataset, without-(b), and with-(a), the scene-context scale. Our new highly efficient scene-context scale achieve a significant improvement on previous work (b).

4. Experimental Results

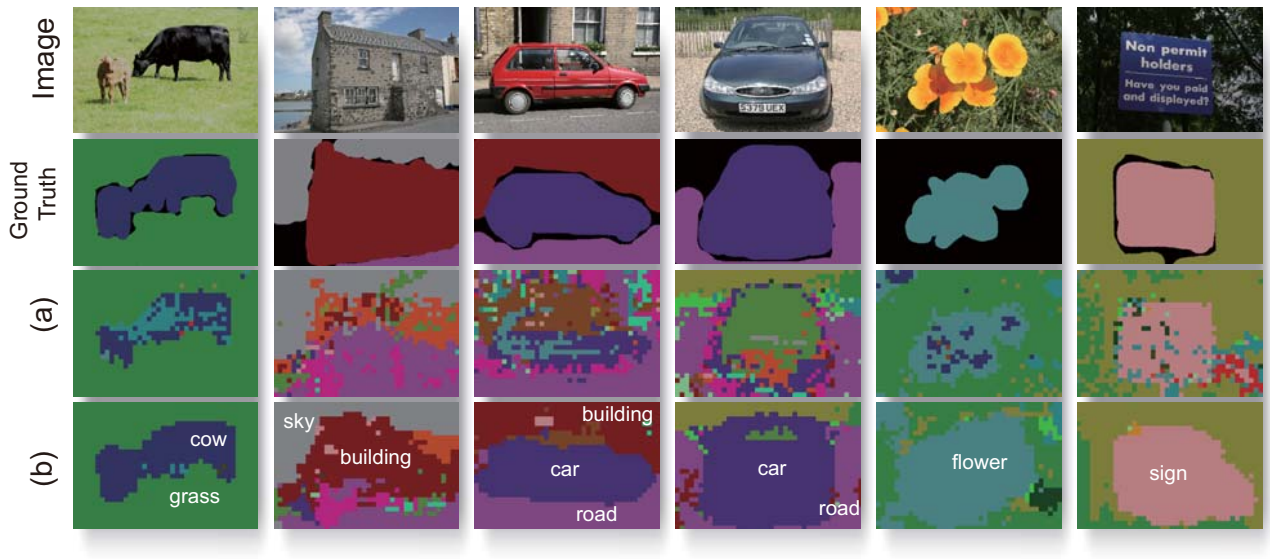
This section presents our experimental results for semantic segmentation and object recognition using scene-context scale. We show two experimental results using scene-context scale: one is the result of clustering and local classification in the multi-scale texton forests and the other is the result of semantic segmentation with bag-of-textons model. To assess the utility of the scene-context scale, we compare the classification accuracy with that of the state-of-art [11] based on single-scale semantic texton forests without using of the scene-context scale. The state-of-art is simulated on C# open source code obtained by "Semantic Texton Forests" site [5]. We use the same train/test split for ours and the state-of-art in all experiments.

We evaluate our algorithm using challenging MSRC21 [9] segmentation dataset that includes a variety of objects such as building, grass, tree, cow, sheep, sky, aeroplane, water, face, car, bike, 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'. These were included both to cope with pixels that do not belong to any database class, and to allow for a rough and quick hand-segmentation which does not align exactly with the object boundaries. Void pixels are ignored for both training and testing.

4.1. Clustering and local classification

To train the multi-scale texton forest, we prepared six scale levels $\mathcal{S} = (S_1, ..., S_6)$ and separately trained the multi-scale texton forests in the six scale levels. The size of image patches has initial size (30×30) and expands their size for split function f to $(30k \times 30k)$ at each scale level S_k. A randomized decision forest $\mathcal{F}_\mathcal{S}$ has following parameters : $T = 5$ trees, maximum depth $D = 10$, $400k$ feature tests at each scale level S_k, 10 threshold tests per split, and 0.25 of the data per tree, resulting in approximately 500 leaves per tree. Training the randomized decision forest on the MSRC dataset took only 10 minutes at each scale level.

We use the standard train/test splits, and the hand-labeled ground truth to train the classifiers. Clustering and local classification performance is measured as both the category



	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)	**45**	89	60	62	65	86	80	**50**	**89**	**70**	**58**	**73**	**48**	20	**80**	**44**	68	37	**31**	**57**	43	**65.2**	**59.8**
(c)	37	86	**62**	**65**	74	83	74	42	87	69	**58**	**73**	47	**24**	77	42	**70**	**45**	28	47	40	64.7	58.6

Figure 7. **Semantic segmentation results on MSRC21 datasets**. Above: (a) The result images of clustering and local classification using scene-context scale with noisy local classification. (b) The result image of semantic segmentation using bag-of-textons model. Below: Segmentation accuracies (percent) over the whole dataset. (a) Our clustering and classification results. (b) Our semantic segmentation results using bag-of-textons model. (c) The state-of-art results [11].

average accuracy (the average proportion of pixels correct in each category) and the global accuracy (total proportion of pixels correct). Fig. 6 shows the results of the clustering and local classification based on randomized decision forest. We estimate the scene-context scale per image pixel using multi-scale texton forests as shown in the third row of Fig. 6. 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 $L = (l_1, ..., l_T)$ for each pixel i as shown in Fig. 6(a). On the other hand, Fig. 6(b) shows the results of the state-of-art [11] 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. 6, 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 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 MSRC21 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. 6. Therefore, we can see that the proposed scene-context scale can be powerful and effective context information for category classification and clustering.

4.2. Semantic segmentation

To train the proposed bag-of-textons model, we select 5000 training samples of each category equally on a random subset. The training error decreases nonlinearly as the number of iterations increases, thus, the experiment was performed with 6000 iterations for week learner of the bag-of-textons model. Since random feature selection improves training time, we have a 10% random feature selection proportion.

Fig. 7(a) shows the results of the clustering and local classification in Section 4.1. After that, the classifier of the bag-of-textons model allows us to semantic segmentation on an image pixel with semantic-context. The semantic

segmentation results on test images show in Fig. 7(b). As can be seen, the proposed segmentation algorithm greatly improves the accuracy in the local classification process, specially the classes with the result of noisy clustering in Section 4.1 such as water, car, bicycle, sign and road show good performance in this process.

Note that we do not use a Markov or conditional random field which could clean up the segmentations to precisely follow image edges. We obtained the segmentation results global 65.2%, class average 59.8% using the bag-of-textons model with scene-context scale. We compared the proposed method with state-of-art using random forest and TextonBoost [12] for semantic segmentation in the table of Fig. 7. In fact, the results of state-of-art is better than 58.6% in their paper [11], 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. In addition, they separately run the categorization and segmentation algorithms and multiply the distributions with image-level prior (ILP) to emphasize the likely categories and discourage unlikely categories using the results of image categorization. However, we exclude the ILP of image categorization results for all experiments. Therefore, across the whole dataset with same experimental condition, the proposed method by using scene-context scale achieved a class average performance of 59.8%, which is better than the 58.6% of the state-of-art.

5. Conclusion

This paper presented a new framework for semantic segmentation and object recognition using the scene-context scale based on multi-scale texton forest. We have (i) introduced the concept of scene-context scale in object recognition, (ii) described the randomize decision forests and expanded it to multi-scale texton forest, and (iii) estimated the scene-context scale and combined with bag-of-textons model for semantic segmentation. 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.

In future work, we can improve accuracy per-category distribution by using geometric transformations and affine photometric transformations on training/test dataset. In addition, the ILP of image categorization result is also utilize as region priors for object recognition.

Acknowlgement

This work was in part supported by JST, CREST.

References

[1] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. 2

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2005. 1

[3] S. Divvala, D. Hoiem, J. Hays, A. Efros, and M. Hebert. An empirical study of context in object detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2009. 1

[4] X. He, R. Zemel, and M. Carreira-Perpinan. Multiscale conditional random fields for image labeling. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2004. 1

[5] M. Johnson. Semantic texton forests reference implementation. In *http://www.matthewajohnson.org/research/stf.html*. 5

[6] S. Kumar and M. Hebert. A hierarchical field framework for unified context-based classification. In *Int. Conf. on Computer Vision*, 2005. 4

[7] V. Lepetit, P. Lagger, and P. Fua. Randomized trees for real-time keypoint recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2005. 2

[8] F. Moosmann, B. Triggs, and F. Jurie. Fast discriminative visual codebooks using randomized clustering forests. In *Neural Information Processing Systems*, 2006. 2

[9] MSRC21. The microsoft research cambridge 21 class database. In *http://research.microsoft.com/vision/cambridge/recognition/*. 5

[10] A. Saxena, S. Chung, and A. Ng. 3-d depth reconstruction from a single still image. *Int. Journal of Computer Vision*, 76(1):53–69, 2008. 1

[11] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2008. 2, 5, 6, 7

[12] 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. Journal of Computer Vision*, 81(1):2–23, 2009. 1, 7

[13] 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. 4

[14] A. Torralba and A. Oliva. Depth estimation from image structure. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(9):1226–1238, 2003. 1

[15] M. Varma and A. Zisserman. A statistical approach to texture classification from single images. *Int. Journal of Computer Vision*, 62(1):61–81, 2005. 2

[16] J. Winn, A. Criminisi, and T. Minka. Categorization by learned universal visual dictionary. In *Int. Conf. on Computer Vision*, 2005. 2