Texton Clustering for Local Classification using Scene-Context Scale

Yousun Kang
Tokyo Polytechnic University
Atsugi, Kanakawa, Japan 243-0297
Email: yskang@cs.t-kougei.ac.jp

Sugimoto Akihiro
National Institute of Informatics
Tokyo, Japan 101-8430
Email: sugimoto@nii.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 texton clustering for local classification 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. In our experiments, we use MSRC21 segmentation dataset to assess our clustering algorithm and show that the usage of the scene-context scale improves recognition performance.**

I. Introduction

There are many sources of scene-context, which play an important role in scene analysis and object recognition [1], [2]. 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 or increasing the size of a region of interest 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. 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 [3]. Saxena *et al.* presented an algorithm for predicting depth from a single still image [4]. 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.

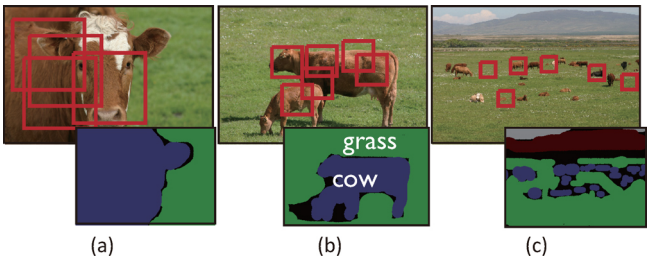


Fig. 1. **Example images of 'cow' category on MSRC dataset.** The objects have different scale such as large scale (a), middle scale (b), and small scale (c) in large dataset.

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 the clustering for semantic segmentation and object recognition. We propose the multi-scale texton forests, which can generate different textons according to scale levels. In addition, scene-context scale can combine the bag-of-textons model with a histogram of the category distribution for semantic segmentation.

To assess the utility of the scene-context scale based on multi-scale texton forests for local classification, we compare the classification accuracy with that of the state-of-the-art [6]. The results show that our clustering method achieves better local classification accuracy than the methods without using of scene-context scale.

II. Multi-scale Texton Forests

In this section, we explain multi-scale texton forests using random forests. The scene-context scale can be obtained by using multi-scale texton forests, which consist of several random forests [5] with different scales. Recently, textonization process is performed on random forests to generate semantic texton by Shotton *et al.* [6] for image categorization and segmentation. We employ the semantic texton forests proposed by Shotton *et al.* [6] and generate different scale levels to obtain multi-scale texton forests.

We increase the size of image patches to expand scale level of random forests. Each random forests have their own scale level and its scale level can expand for multi-scale texton

978-1-4673-5621-3/13/$31.00 ©2013 IEEE

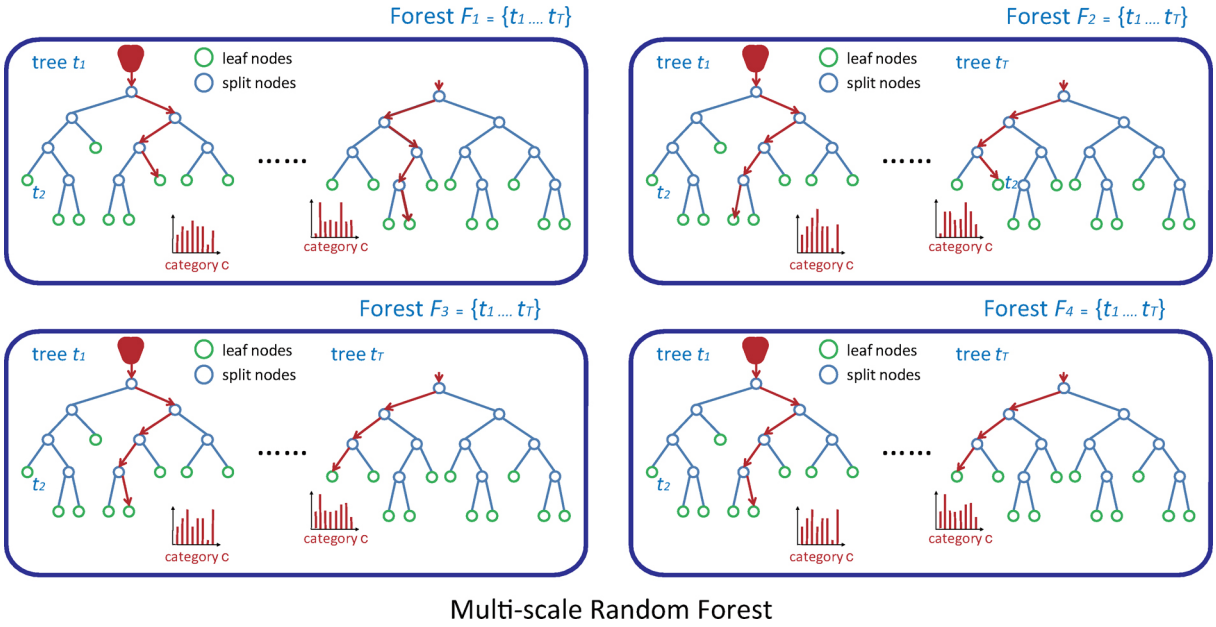


Fig. 2. **Multi-scale texton forest** The multi-scale texton forest consists of several semantic texton forests [6] with various scale levels and each semantic texton forest consists of randomized decision trees with same scale level

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

In the first scale level $k = 1$, an image patch p_1 covers whole pixels within a $(d \times d)$ size on which the split node functions for the first semantic texton forest $\mathcal{F}_1$ act. In the next scale level $k = 2$, the increased image patch p_2 covers the pixels within a $(2d \times 2d)$ size excluding the former image patch p_1. Therefore, the size of image patch p_k is increased to $(kd \times kd)$ pixels excluding the image patch $p_{(k-1)}$ that is for the former scale level $(k-1)$ as shown in Fig. 3.

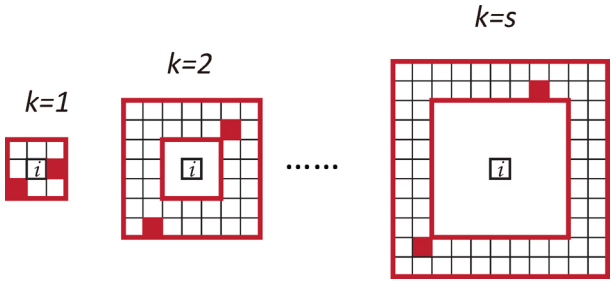


Fig. 3. **Dilatation of a region of interest according to scale space k.** Various sizes of a region of interest are used for node split function in the multi-scale texton forests.

Multi-scale texton forests have been utilized in classifiers [7] or clustering with the fast and powerful performance. To textonize an image, an image patch p_k are passed down the multi-scale texton forest according to their scale level. We can obtain the class distributions $P_k(c|L_k)$ by averaging the local distributions over the leaf nodes $L_k = (l_1, l_2, ..., l_T)$ at scale k as

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

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

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

III. Scene-Context Scale

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

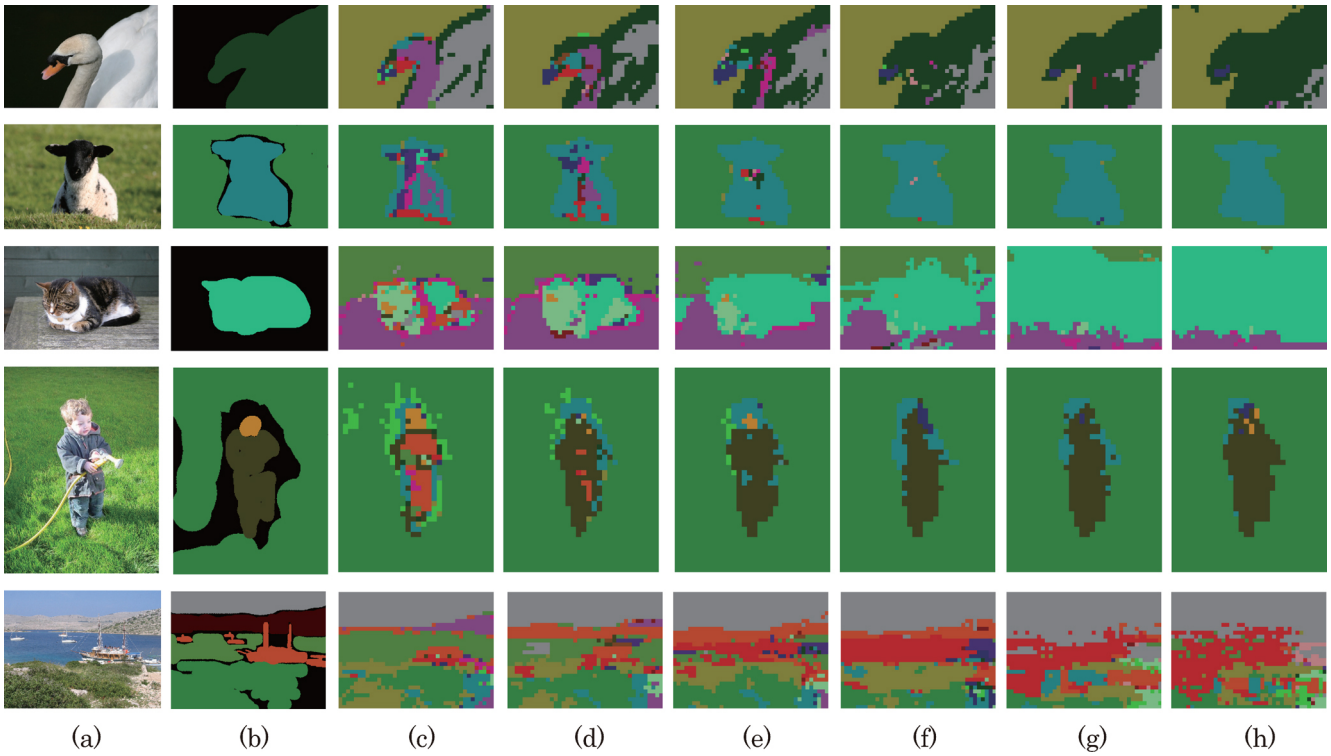


Fig. 4. **Clustering and classification results on MSRC segmentation dataset using multi-scale texton forests.** The multi-scale texton forest can generate the different textons according to scale levels. (a) Input images. (b) Ground-truth images. (c) – (h) Clustering and classification results according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

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

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

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

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

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

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

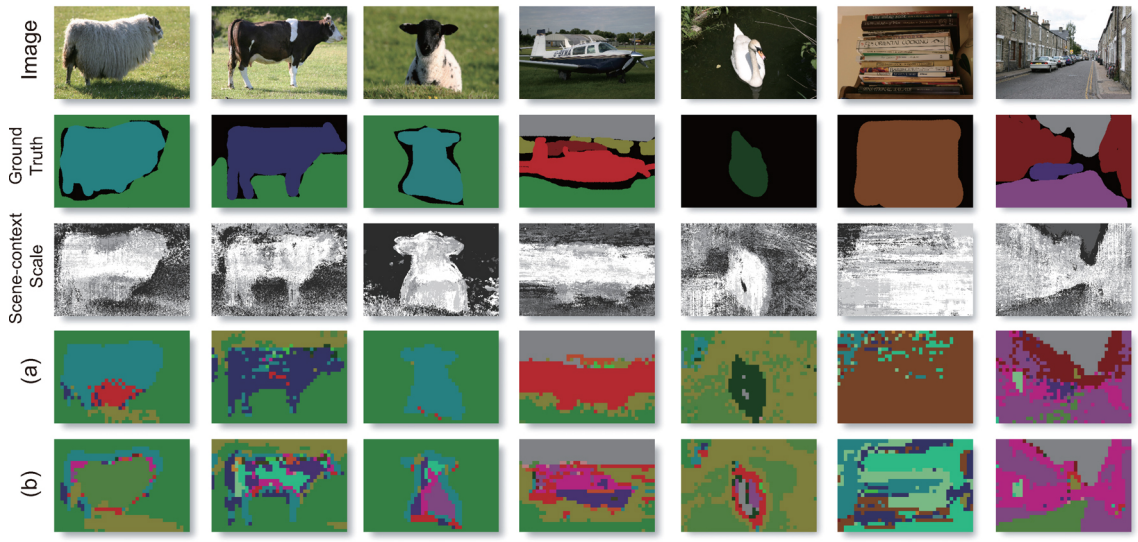
Among multi-scale texton forests, a semantic texton forest $\mathcal{F}_{k^*}$ is selected in the textonization process. The semantic texton forest $\mathcal{F}_{k^*}$ has the instance k^* of the most likely scene-context scale. We can define the texton generated by the semantic texton forest $\mathcal{F}_{k^*}$ as our scale-optimized texton. Our textonization process exploits the class distributions $P_{k^*}(c|L_{k^*})$ in the semantic texton forest $\mathcal{F}_{k^*}$ with the scene-context scale k^*.

IV. Experimental Results

This section presents our experimental results for texton clustering for semantic segmentation and object recognition using scene-context scale. To assess the utility of the scene-context scale, we compare the classification accuracy with that of the state-of-art [6] 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 [8]. 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. 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. Clustering and local classification performance is measured as both the category average accuracy and the global accuracy.

Fig. 4 shows the results of the clustering and class clas-



	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. 5. **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 [6] 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).

sification 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. From the results of the first and second rows of Fig. 4, we notice that the more the scale level increases, the more their performance also increases. However, the image of the third row shows the roughest result in the largest scale level. Most intriguingly of all, the fourth row's image has different performance between category according to scale level, i.e., the 'face' is classified at the smallest scale, but 'body' is classified at larger scale. At the image of the last row, it shows the good performance in the smallest scale as we expected. As a result, we can see that the scene-context scale is to be estimated per image pixel.

Fig. 5 shows the results of the clustering and local classification using scene-cotext scale. We estimated the scene-context scale per image pixel as shown in the third row of Fig. 5. 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. 4(a). On the other hand, Fig. 4(b) shows the results of the state-of-art [6] 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 $\mathcal{F}_{\mathcal{S}_1}$.

As shown in Fig. 5, 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. 4. Therefore, we can see that the proposed scene-context scale can be powerful and effective context information for category classification and clustering.

V. Conclusion

We have introduced the concept of scene-context scale in object recognition and described the multi-scale texton forest to estimated the scene-context scale. In experiments, we confirmed that the proposed texton clustering method using scene-context scale gives better results than any other methods without using scene-context scale. In future work, we can image categorization and semantic segmentation using the proposed texton clustering method.

Acknowlgement

This work was supported by JSPS and CREST.

REFERENCES

[1] 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.

[2] 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.

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

[4] 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.

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

[6] 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.

[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.

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

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