Random Forests Based Image Categorization Using Scene-Context Scale

Yousun KANG[†] and Akihiro SUGIMOTO[†]

[†] National Institute of Informatics
Hitotsubashi 2–1–2, Chiyoda-ku, Tokyo, 101–8430 Japan
E-mail: †{yskang,sugimoto}@nii.ac.jp

Abstract Scene-context plays an important role in scene analysis and object recognition. This paper presents random forests based image categorization using scene-context scale. The proposed method uses the random forests, which are ensembles of randomized decision trees. Since the randomized decision trees are extremely fast to both train and test, it is possible to perform classification, clustering and regression in real time. We train the multi-scale texton forests which efficiently provide both a hierarchical clustering into semantic textons and local classification in various scale space. The use of the scene-context scale improves image categorization performance. We evaluate on MSRC21 segmentation dataset. Our results advance the state-of-the-art in image categorization accuracy, and the use of efficient decision forests facilitates execution speed.

Key words Random forests, Scene-context scale, Image categorization

1. Introduction

Scene-context plays an important role in scene understanding. In fact, computer vision approaches have demonstrated that the use of context improves recognition performance [1, 2, 3]. While the term context is frequently used in the literature as an important keyword, it is difficult to give its clear definition. There are many sources of scene-context and numerous psychophysics studies have presented new theories of context for human object recognition [4, 5].

When the context is used on a per-pixel level, we can capture the local context that 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 local level can not be enough to identify the pixel class. The more the region of 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 [6].

The size of a region of interest is related to size of objects in a scene. Given object presence and location, its scale or relative size in a scene can be a significant cue for recognizing the objects in the scene. We refer this scale 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.

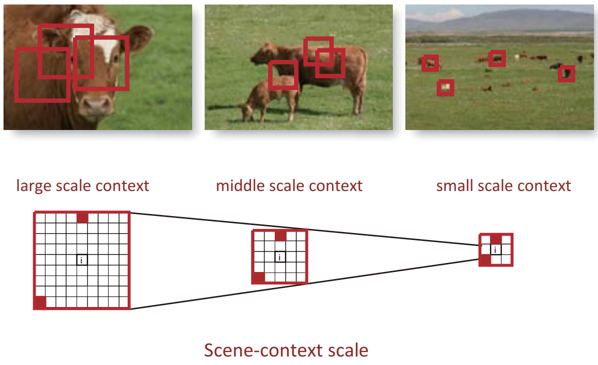


Fig. 1 **Three examples of images with different scene-context scale.** The objects strongly differ in their scale in an image.

The various scene-context scales of images are illustrated in Fig. 1. 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. However, the estimation of the scene-context scale still remains difficult and unreliable in current computational approaches.

In this paper, we estimate the scene-context scale using multi-scale texton forests, which consist of several randomized decision forests with different scales. Random forests [7] have proven powerful tools with high computational efficiency in vision applications [8, 9, 10]. For categorization and segmentation, Shotton *et* *al.* [11] proposed semantic texton forests as efficient texton codebooks without using of scene-context scale. We propose multi-scale texton forests, which can generate different textons according to scale space. We investigate how scene-context scale combines with multi-scale texton forests to improve the accuracy of categorization.

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 the state-of-the-art [11]. The results show that the proposed method achieves better classification accuracy than the methods without using of scene-context scale.

2. Randomized Decision Forests

In this section, we begin with a brief review of randomized decision forests [7]. A decision forest is an ensemble of T decision trees. 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. The whole forest achieves an accurate and robust classification by averaging the class distributions over the leaf nodes $L = (l_1, ..., l_T)$:

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

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 $\mathbf{v}$:

$$I_l = \{i \in I_n | f(\mathbf{v}_i) < \kappa\}, \tag{2}$$

$$I_r = I_n \backslash I_l. \tag{3}$$

At each split node, several candidates for function f and threshold κ are generated randomly, and the one that maximizes the expected gain in information about the node categories is chosen [9]:

$$\Delta E = -\frac{|I_l|}{|I_n|}E(I_l) - \frac{|I_r|}{|I_n|}E(I_r), \tag{4}$$

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.

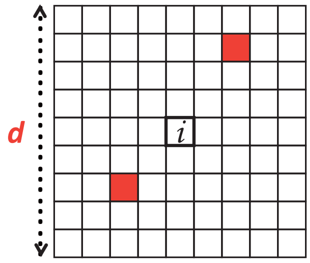


Fig. 2 **A region of interest for node split function of randomized decision trees.** The split nodes of decision trees use simple functions of raw image pixels within a $(d \times d)$ image patch.

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 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.

The amount of training data may be significantly biased towards certain classes in some datasets. A classifier learned on this data will have a corresponding prior preference for those classes. To normalize this bias, we weight each training example by the inverse class frequency: $w_i = \xi_{c_i}$, where $\xi_c = (\sum_{i \in I}[c = c_i])^{-1}$. The classifiers trained using this weighting tend to give a better class average performance. After training, an improved estimate of the class distributions is obtained using all training data I, but not just the random subset.

3. Multi-scale Texton Forest

Textons [12] have proven powerful discrete image representations for categorization and segmentation. 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 [13].

By using random forests, we can build powerful texton codebooks without computing expensive filter-banks or descriptors, and without performing costly k-means clustering and nearest-neighbor assignment. Therefore, when the bag of textons method is employed for categorization and segmentation, the random forests

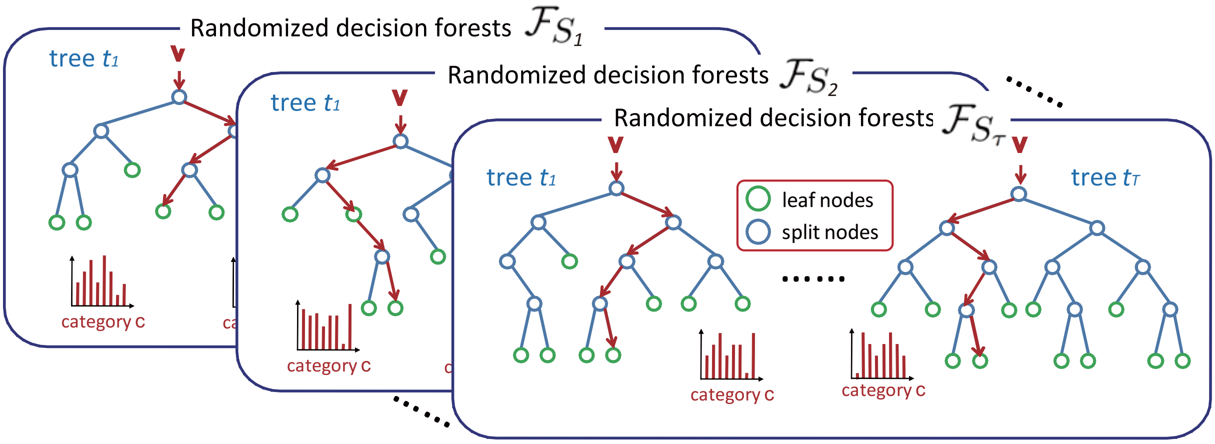

Randomized decision forests $\mathcal{F}_{S_1}$
Randomized decision forests $\mathcal{F}_{S_2}$
Randomized decision forests $\mathcal{F}_{S_\tau}$
tree t_1
tree t_T
leaf nodes
split nodes
category c

Fig. 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 step.

have the advantage of being extremely fast and high performance.

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 step 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 of Section 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 step 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 step 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 step 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.

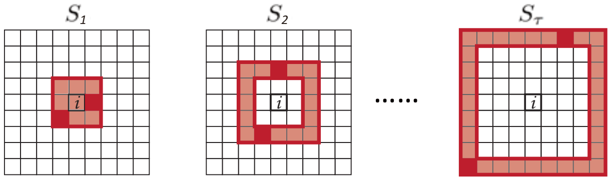


Fig. 4 **Dilatation of a region of interest according to scale space S_k.** Various sizes of a region of interest are used for node split function in the multi-scale texton forests.

To textonize an image according to scale steps, 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 semantic textons presents a much more powerful feature for image categorization.

4. Image Categorization

One of the most important tasks in computer vision is image categorization. Categorizing an image consists of determining those categories (e.g. forest images, office images, moon images) to which the image belongs. Image categorization is one way in which we can perform image retrieval and segmentation or object detection.

4. 1 Non-linear SVM

We use a bag of textons model [11] computed across the whole image for image categorization. The bag of textons model uses the histogram of semantic textons and the node prior distributions over the whole image, even discarding spatial layout. The histogram is used as an input to a classifier to recognize object categories.

For a classifier we use a non-linear support vector machine (SVM). The non-linear SVM depends on a kernel function K that defines the similarity measure between images. To take advantage of the hierarchy in the multi-scale texton forest, we adapt the pyramid match kernel method (based on [11]) to act on a pair of bag of textons histograms computed across the whole image.

The kernel function (based on [14]) is then

$$K(P,Q) = \frac{l}{\sqrt{Z}}\tilde{K}(P,Q), \tag{5}$$

where Z is a normalization term for images of different sizes

$$Z = \tilde{K}(P,P)\tilde{K}(Q,Q), \tag{6}$$

and $\tilde{K}$ is the actual matching function, computed over levels of the tree as

$$\tilde{K}(P,Q) = \sum_{d=1}^{D} \frac{l}{2^{D-d+1}}(\Gamma_d - \Gamma_{d+1}), \tag{7}$$

using the histogram intersection Γ

$$\Gamma_d = \sum_j \min(P_d[j], Q_d[j]), \tag{8}$$

where D is the depth of the tree, P and Q are bag of textons, and P_d and Q_d are the portions of the histograms at depth d, with j indexing over all nodes at depth d. There are no nodes at depth $D+1$, hence $\Gamma_{D+1} = 0$. If the tree is not full depth, missing nodes j are simply assigned $P_d[j] = Q_d[j] = 0$.

The kernel over all trees in a random forest $\mathcal{F_S}$ is calculated as $K = \sum_t \gamma_t K_t$ with mixture weights γ_t. We build a 1-vs-others SVM kernel K_c per category, in which the count for node n in the bag of semantic textons histogram is weighted by the value $P_{\mathcal{F_S}}(c|n)$. This helps balance the categories, by selectively down-weighting those that cover large image areas (e.g. grass, water) and thus have inappropriately strong influence on the pyramid match, masking the signal of smaller classes (e.g. cat, bird).

4. 2 Estimation of scene-context scale

At each image patch, a random forest with minimum entropy of leaf nodes can be chosen among the multi-scale texton forests. We use the entropy of a leaf node as the criterion of an optimal scale space to be chosen. The optimal scale space can be obtained by computing entropy of the class label distribution and the scene-context scale of an image is obtained by integrating the optimal scale space at each image patch.

In order to estimate the scene-context scale of an image, we compute the entropy $E(I|L_{\mathcal{S}})$ of each image patch I at leaf nodes $L_{\mathcal{S}}$ of every random forest $\mathcal{F_S}$. Among the scale space $\mathcal{S} = (S_1, ..., S_\tau)$, 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

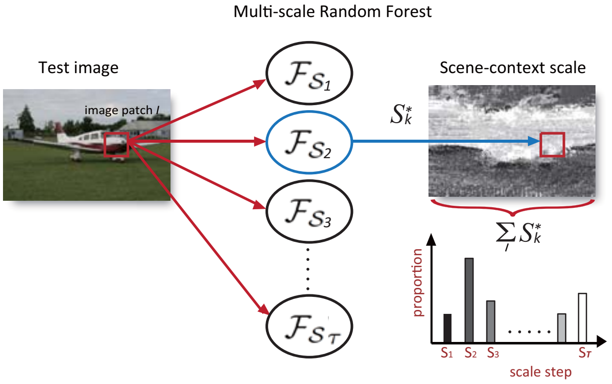


Fig. 5 **The scene-context scale of an images can be estimated by computing the minimum entropy of each image patch.** Darker pixels correspond to smaller scale, so white pixels represent the largest scale S_τ.

$$S_k^* = \arg\min_{S_k} E(I|L_{S_k}). \tag{9}$$

We can estimate the scene-context scale in an image as the proportion of the instances of scale space S_k^* of image patches. This gives the distribution of scale space $P(\mathcal{S})$ as shown in Fig. 5. We can determine the Scale-Level Prior (SLP) that is the most likely scale space S_k in whole image.

For each test image, 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

$$P'(c|\mathcal{F_S}) = \sum_{k=1}^{\tau} P(c|\mathcal{F}_{S_k}) \times P(S_k). \tag{10}$$

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

$$P'(c|\mathcal{F_S}) = (\frac{l}{\tau}\sum_{k=1}^{\tau} P(c|\mathcal{F}_{S_k}))\times P(S_{SLP})^{\alpha} \tag{11}$$

using parameter α to soften the prior.

5. Experimental Results

We evaluate our algorithm using challenging MSRC21 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

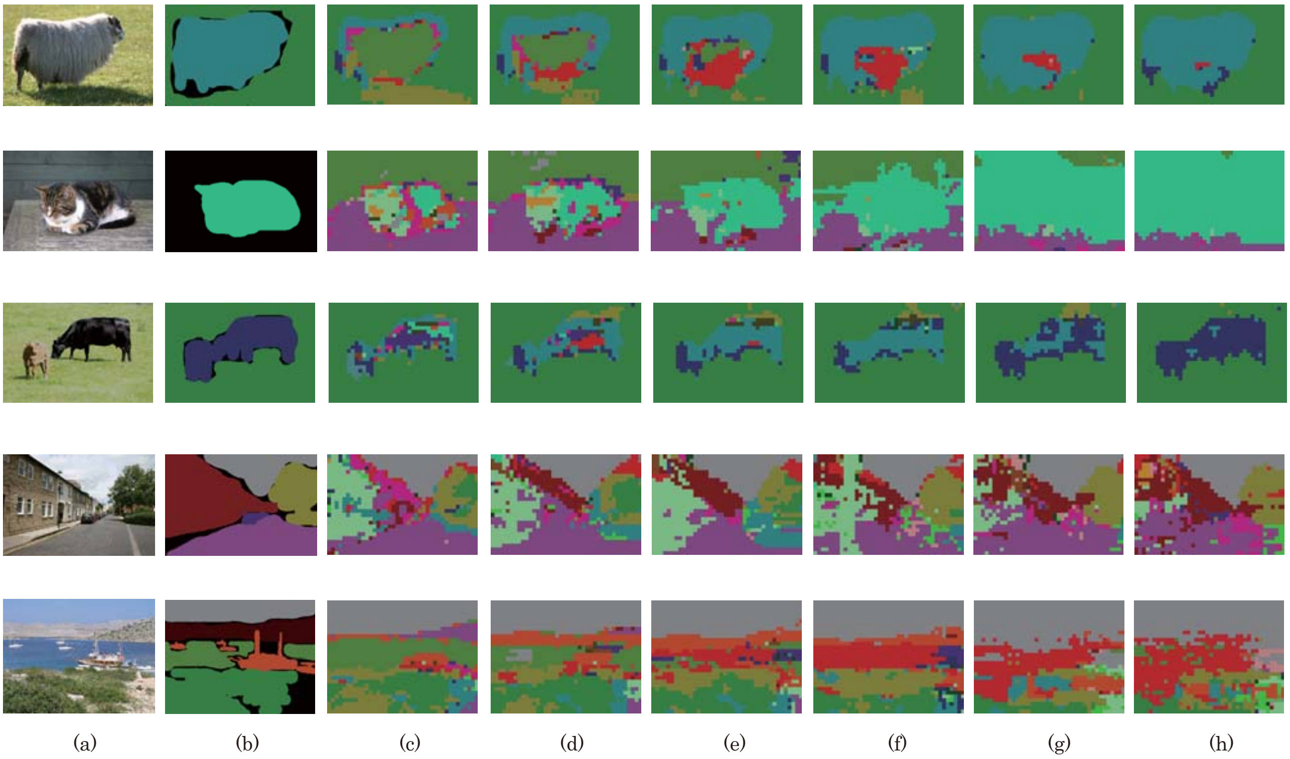


Fig. 6 **Clustering and classification results using the multi-scale texton forest.** The multi-scale texton forest can generate the different textons according to scale steps. (a) Input images. (b) Ground-truth images. (c) - (h) Clustering results according to scale space $\mathcal{S} = (S_1, ..., S_6)$. 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).

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.

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.

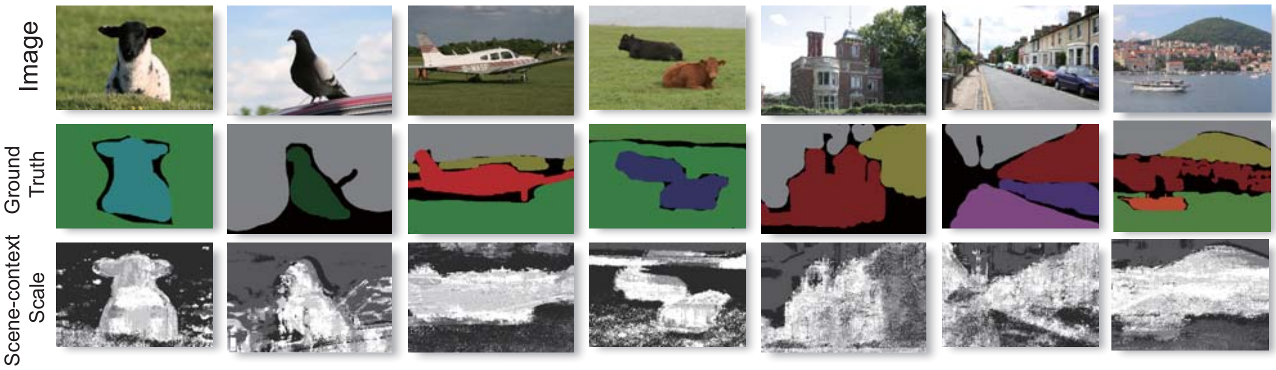
To train the multi-scale texton forest, we prepared six scale steps $\mathcal{S} = (S_1, ..., S_6)$ and a initial image patch size is (15×15). Therefore, the size of image patches for split function f is $(15k \times 15k)$ at each scale step S_k. A randomized decision forest $\mathcal{F}_{\mathcal{S}}$ has following parameters : $T = 5$ trees, maximum depth $D = 10$, 500 feature tests and 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 step.

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. As can be seen, the pixel level classification based on the local distributions gives poor, but still good performance and gives different results according to each scale step.

Using the multi-scale texton forest, we estimate the scene-context scale in test images. Fig. 7 shows the test images and ground-truth images and its scene-context scale. The categorization accuracies (percent) over the whole dataset, are also shown in table of Fig. 7.

We obtained the results (a) and (c) without using the scene-context scale, and (b) 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 [11]. (b) SLP in the second row of the table used the equation (11) in Section 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 equation (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



	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. 7 **Image categorization results on MSRC21 datasets**. Above: Scene-context scale of test images using multi-scale texton forests. Below: Categorization accuracies (percent) over the whole dataset. Scene-context scale achieves a improvement on previous work.

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. In particular, significant improvement can be observed difficult classes: grass and cat.

6. Conclusion

This paper presented a new framework for image categorization using the multi-scale texton forest and scene-context scale. 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) achieved efficient categorizing by using a combination of scene-context scale and multi-scale texton forest. The multi-scale texton forest can be utilized in semantic segmentation and object recognition by integrating scene-context scale with bag of textons method.

Acknowlgement

This work was in part supported by JST, CREST.

References

[1] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. Int. Journal of Computer Vision, 80(1), 2008.

[2] K. Murphy, A. Torralba, and W. T. Freeman. Using the forest to see the trees: a graphical model relating features, objects and scenes. In Proc. NIPS. MIT Press, 2003.

[3] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-ocurrence, location and appearance. In CVPR, 2008.

[4] A. Oliva and A. Torralba. The role of context in object recognition. Trends Cogn Sci, November, 2007.

[5] P. Carbonetto, N. de Freitas, and K. Barnard. A statistical model for general contextual object recognition. In Proc. ECCV, 2004.

[6] L. Wolf and S. Bileschi. A critical view of context. Int. Journal of Computer Vision, 69(2):251–261, 2006.

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

[8] F. Moosmann, B. Triggs, and F. Jurie. Fast discriminative visual codebooks using randomized clustering forests. In Proc. NIPS, 2006.

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

[10] G. Rogez, J. Rihan, S. Ramalingam, C. Orrite and P. Torr. Randomized trees for human pose detection. In Proc. CVPR, 2008.

[11] J. Shotton, M. Johnson, and R. Cipolla. Semantic Texton Forests for Image Categorization and Segmentation. In Proc. CVPR, 2008.

[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.

[13] J. Winn, A. Criminisi, and T. Minka. Categorization by learned universal visual dictionary. In Proc. ICCV, pages 2:1800–1807, 2005.

[14] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In Proc. ICCV, 2005.