Object and Scene Recognition Based on Learning Probabilistic Latent Component Tree with Boosted Features

Masayasu Atsumi

***Abstract*—This paper proposes an object and scene categorization method based on the probabilistic latent component tree with boosted features. In this method, object classes are firstly obtained by clustering a set of object segments extracted from scene images in each scene category through the probabilistic latent component analysis with the variable number of classes. Then the probabilistic latent component tree with boosted features at its branch nodes is generated as a classification tree of all the object classes of all the scene categories followed by labeling object classes. Lastly, each scene category is characterized according to the composition of its labeled object classes. Object and scene recognition is simultaneously performed based on the probabilistic latent component tree search by using composite boosted features for the tree traversal. Through experiments by using images of plural categories in an image database, it is shown that performance of object and scene recognition is high and improved by using composite boosted features in the probabilistic latent component tree search.**

***Index Terms*—Boosting, categorization, computer vision, probabilistic learning.**

I. Introduction

Object and scene categorization is a basic ability for a human to understand the visual world and also one of the fundamental problems in computer vision. In object and scene categorization, it is effective to organize object categories into a classification tree by their discriminative appearance and to associate scene categories with frequent object categories appeared in them. The problem to be addressed in this paper is learning a classification tree of object appearances which are related to scenes through the object composition for object and scene recognition. By the way, for a scene which contains plural objects, a human perceives one object in the foreground and other objects in the background. Thus, in this problem, a set of scene images each of which is labeled with one of plural objects in a scene is provided for learning. Here a labeled object in a scene is an object which is considered to be in the foreground and other objects are in the background. A set of scene images each of which contains the same foreground object forms a scene category and a scene image can be contained in plural scene categories dependent on which object is considered to be in the foreground.

In this paper, we propose an object and scene categorization method based on the probabilistic latent component tree [1] with boosted features. In this method, for a set of object segments extracted from scene images in each scene category, object classes are firstly obtained by clustering the object segments through the probabilistic latent component analysis with the variable number of classes (V-PLCA). Then the probabilistic latent component tree with boosted features (PLCT-BF) is generated based on similarity among object classes as a classification tree of all the object classes of all the scene categories followed by labeling object classes by using their representative instances whose category names are given as teaching signals. In this probabilistic latent component tree (PLCT) generation, discriminative features are selected through a boosting procedure on branch nodes to improve the PLCT search for recognition. Lastly, each scene category is characterized according to the composition of its labeled object classes. In object and scene recognition, for a scene image which contains several object segments, object categories are obtained through the PLCT-BF search and a scene category is determined among scene categories which contains either of those object categories in the foreground.

As for related work, probabilistic latent variable models have been applied to learning object and scene categories [2], [3], [4] and there have been proposed hierarchical models for object and scene categorization [5], [6], [7]. There have also been proposed methods of image classification through boosting [8], [9], [10] since AdaBoost has been applied to object detection [11]. The main difference of our method from these existing ones is that it is not necessary to fix the number of object classes and the depth of a classification tree in advance, object and scene categories are recognized at the same time, and discriminative features are selected with their confidences on a classification tree through boosting.

This paper is organized as follows. Section II describes the proposed method, experimental results are shown in Section III and we conclude our work in Section IV.

II. Proposed Method

Let C be a set of categories and N_C be the number of categories. A scene category $c \in C$ is a set of scene images each of which contains an object of the category in the foreground and other categorical objects in the background. Let s_{c,i_j} be a j-th object segment extracted from a scene image i of a scene category c, S_c be a set of object segments extracted from any scene images of a scene category c and N_{S_c} be the number of object segments in S_c. An object segment is represented by a bag of features (BoF) [12] of its local feature. Let F be a set of key features as a code book, f_n be a n-th key feature of F and N_F be the number of key

Manuscript received September 19, 2012; revised November 12, 2012. This work was supported in part by Grant-in-Aid for Scientific Research (C) No.23500188 from Japan Society for Promotion of Science.

M. Atsumi is with the Department of Information Systems Science, Faculty of Engineering, Soka University, 1-236 Tangi, Hachioji, Tokyo, Japan (e-mail: matsumi@t.soka.ac.jp).

10.7763/IJMLC.2012.V2.232

features. Then an object segment s_{c,i_j} is represented by a BoF of key features $H(s_{c,i_j}) = [h_{c,i_j}(f_1),...,h_{c,i_j}(f_{N_F})]$.

Let $H_c = \{H(s_{c,i_j}) \mid s_{c,i_j} \in S_c\}$ be a set of BoFs obtained from a set of scene images of a scene category $c \in C$ and $\{H_c\}_{c\in C}$ be given for a set of scene categories. The learning problem is to compute a set of object classes Q_c from H_c, a classification tree PLCT of all classes $\cup_{c\in C}Q_c$ each class of which is located at a leaf of the PLCT and labeled with its object category, boosted key features with their confidences at all the branch nodes of the PLCT, and the characterization of each scene category $c \in C$ by using a set of key features F. The recognition problem is to simultaneously obtain object categories and a scene category through the PLCT-BF search for a given scene image which contains several object segments whose categories are contained in a set of categories C.

A. Probabilistic Latent Component Analysis of Scene Cate- gories

The probabilistic latent component analysis with the variable number of classes (V-PLCA) computes a set of classes $Q_c=\{q_{c,r}|r=1,\ldots,N_{Q_c}\}$ which represents object categories of each scene category $c \in C$, where N_{Q_c} is the number of classes in Q_c. The problem to be solved is estimating probabilities, $p(s_{c,i_j}, f_n) = \sum_r p(q_{c,r})p(s_{c,i_j} \mid q_{c,r})p(f_n \mid q_{c,r})$, namely class probabilities $\{p(q_{c,r}) \mid q_{c,r} \in Q_c\}$, conditional probabilities of instances $\{p(s_{c,i_j} \mid q_{c,r}) \mid s_{c,i_j} \in S_c, q_{c,r} \in Q_c\}$, conditional probabilities of key features $\{p(f_n \mid q_{c,r}) \mid f_n \in F, q_{c,r} \in Q_c\}$, and the number of classes N_{Q_c} that maximize the following log-likelihood

$$L_c = \sum_{i_j}\sum_n h_{c,i_j}(f_n)\log p(s_{c,i_j}, f_n) \quad (1)$$

for a set of BoFs $H_c = \{H(s_{c,i_j}) \mid s_{c,i_j} \in S_c\}$. The class probability represents the composition ratio of object categories in a scene category, the conditional probability of instances represents the degree that object segments are instances of an object category and the conditional probability distribution of key features represents feature of an object category.

These probabilities are estimated by the tempered EM algorithm and the number of classes is determined through the EM iterative process with subsequent class division. The process starts with one or a few classes, pauses at every certain number of EM iterations less than an upper limit and calculates the following dispersion index

$$\delta_{c,r} = \sum_{s_{c,i_j}}((\sum_{f_n}|p(f_n \mid q_{c,r}) - \kappa(s_{c,i_j}, f_n)|)\times p(s_{c,i_j} \mid q_{c,r})) \quad (2)$$

where

$$\kappa(s_{c,i_j}, f_n) = \frac{h_{c,i_j}(f_n)}{\sum_{f_{n'}} h_{c,i_j}(f_{n'})} \quad (3)$$

for $\forall q_{c,r} \in Q_c$. Then a class whose dispersion index takes a maximum value among all classes is divided into two classes. This iterative process is continued until dispersion indexes or class probabilities of all the classes become less than given thresholds [1].

B. Probabilistic Latent Component Tree for Object Catego- rization

The probabilistic latent component tree (PLCT) is generated as a classification tree of all classes $Q^*=\cup_{c\in C}Q_c$. The PLCT is a binary tree in which similar classes are located at close leaf nodes where the similarity is calculated by using the conditional probability distribution of key features and class probabilities.

Let $B(Q)$ be a branch node where $Q(\subseteq Q^*)$ is a set of classes which are located at leaf nodes of a subtree whose root is the branch node. Note that $Q = Q^*$ for a root node of a PLCT. Then two child nodes of the parent node $B(Q)$ are generated as follows. First of all, for each key feature $f_n \in F$,

Q is divided into two subsets of classes $Q^1_{f_n} = \{q_{c,r} \mid p(f_n \mid q_{c,r}) \le \varepsilon_f, q_{c,r} \in Q\}$ and $Q^2_{f_n} = \{q_{c,r} \mid p(f_n \mid q_{c,r}) > \varepsilon_f, q_{c,r} \in Q\}$ where ε_f is 0 or a small positive value and 0 by default. Next, mean probability distributions of key features of classes $Q^1_{f_n}$ and $Q^2_{f_n}$ are calculated as $\{\mu_{Q^1_{f_n}}(f_{n'}) \mid f_{n'} \in F\}$ and $\{\mu_{Q^2_{f_n}}(f_{n'}) \mid f_{n'} \in F\}$ respectively and the following distance

$$D_{f_n} = \sum_{q_{c,r}\in Q^1_{f_n}} p(q_{c,r})(\sum_{f_{n'}\in F} p(f_{n'} \mid q_{c,r})\log\frac{p(f_{n'} \mid q_{c,r})}{\mu_{Q^1_{f_n}}(f_{n'})}) + \sum_{q_{c,r}\in Q^2_{f_n}} p(q_{c,r})(\sum_{f_{n'}\in F} p(f_{n'} \mid q_{c,r})\log\frac{p(f_{n'} \mid q_{c,r})}{\mu_{Q^2_{f_n}}(f_{n'})}) \quad (4)$$

is computed based on the KL information between each and mean probability distributions of key features. Finally, Q is divided into two subsets of classes Q^1 and Q^2 which give the minimal value of D_{f_n} for any key feature $f_n \in F$. Then for each of $Q^k (k = 1, 2)$, a branch node $B(Q^k)$ is generated as a child node if the number of classes in Q^k is greater than 1 and a leaf node $L(Q^k)$ is generated as a child node if the number of classes in Q^k is 1. The generation of child nodes by dividing a set of classes is started from a root node $B(Q^*)$ and is recursively repeated on branch nodes until leaf nodes are generated. A leaf node $L(\{q_{c,r}\})$ has one class $q_{c,r}$ so that its class probability, conditional probability distribution of key features and conditional probabilities of instances are maintained in the leaf node where the class probability is normalized as $p(q_{c,r})/N_C$ by dividing $p(q_{c,r})$ by the number of scene categories N_C. A branch node also has a class probability and a conditional probability distribution of key features. Let ν be a branch node and ν^1 and ν^2 be its child nodes. For class probabilities $p(\nu^1)$ and $p(\nu^2)$ and conditional probability distributions of key features $\{p(f_n \mid \nu^1) \mid f_n \in F\}$ and $\{p(f_n \mid \nu^2) \mid f_n \in F\}$ of child nodes, the branch node has a class probability $p(\nu) = p(\nu^1) + p(\nu^2)$ and a conditional probability distribution of key features $\{p(f_n \mid \nu) \mid f_n \in F\}$ a probability value of which is obtained by

$$p(f_n \mid \nu) = \frac{p(\nu^1)}{p(\nu)}\times p(f_n \mid \nu^1) + \frac{p(\nu^2)}{p(\nu)}\times p(f_n \mid \nu^2). \quad (5)$$

Leaf classes are labeled by using object category labels given for representative instances of those leaf classes in a semi-supervised manner [1]. An instance whose conditional probability for a class is the maximum is used as a representative instance for the class. Through labeling object classes, it turns out whether each class of a scene category represents a foreground object category or a background object category in the scene category. The feature of a scene category is represented by the composition of conditional probability distributions of key features for foreground and background object categories in the scene category. Let Q_c^f and Q_c^b be sets of classes which represent foreground and background object categories in a scene category $c \in C$ and $Q_c^f(\theta_f) = \{q_{c,r} \mid q_{c,r} \in Q_c^f, p(q_{c,r}) \geq \theta_f\}$ and $Q_c^b(\theta_b) = \{q_{c,r} \mid q_{c,r} \in Q_c^b, p(q_{c,r}) \geq \theta_b\}$ be subsets of Q_c^f and Q_c^b respectively. Then a probability distribution of key features for the scene category c is expressed by

$$p(f_n \mid Q_c^f(\theta_f), Q_c^b(\theta_b)) = \sum_{q_{c,r} \in Q_c^f(\theta_f) \cup Q_c^b(\theta_b)} \psi(q_{c,r}) \times p(f_n \mid q_{c,r}) \quad (6)$$

$$\psi(q_{c,r}) = \frac{p(q_{c,r})}{\sum_{q_{c,r'} \in Q_c^f(\theta_f) \cup Q_c^b(\theta_b)} p(q_{c,r'})} \quad (7)$$

for $\forall f_n \in F$.

C. Boosting Features on Probabilistic Latent Component Tree

Branch and leaf nodes of the PLCT have conditional probability distributions of key features which encode features of object categories. In recognition of an object category based on the PLCT, for a BoF of an object segment, a leaf node of the object category is obtained through the PLCT search in which distances between those conditional probability distributions of key features and the BoF are calculated and used to traverse branches. In order to improve the performance of the search, a subset of key features is selected with their confidences through a boosting procedure at each branch node. Training samples for the boosting are generated as follows from a conditional probability distribution of key features $\{p(f_n \mid q) \mid f_n \in F\}$ for a class $q \in Q^*$ of each leaf node. First, for each sample u, the number of local feature points N_u is selected uniformly from a range $[N_{u1}\ N_{u2}]$. Next, a BoF of key features $H(u)(= [h_u(f_1), \ldots, h_u(f_{N_f})])$ is generated by selecting N_u key features according to a conditional probability distribution of key features. Then, a distribution of the BoF $D(u)(= [h_u(f_1)/\sum_{f_i} h_u(f_i), \ldots, h_u(f_{N_f})/\sum_{f_i} h_u(f_i)])$ is obtained from the $H(u)$. A given number of samples, that is, distributions of BoFs are generated for each leaf node class.

A subset of key features is selected with their confidences at each branch node $B(Q)$ as follows, where Q is a set of classes which are located at leaf nodes of a subtree whose root is the branch node. Let $\Phi = \{(u_i, w_i, v_i) \mid i = 1, \ldots, N_\Phi\}$ be a set of samples with their weights for boosting where u_i is a sample which is generated from a class in Q and is represented by a distribution of a BoF, w_i is a weight of the sample, $v_i \in \{1, -1\}$ is a label of the sample and N_Φ is the number of samples. A label v_i takes a value 1 or −1 according to whether u_i is a sample which is generated from one of leaf node classes of a left child subtree or a right child subtree of a branch node $B(Q)$. Let T be the number of boosted key features.

Step1: Initialize weights of all samples with $w_i = 1/N_\Phi$.

Step2: For $t = 1, \ldots, T$, select an index of a key feature λ_t and its confidence α_t under a distribution of sample weights. Let $p_1(\lambda)$ and $p_2(\lambda)$ be the λ-th elements of conditional probability distributions of key features of left and right child nodes respectively and $u(\lambda)$ be the λ-th element of a distribution of a BoF of a sample u. Then let us define $\eta(\lambda, u)$ as follows.

$$\eta(\lambda, u) = \begin{cases} 1 & |p_1(\lambda) - u(\lambda)| < |p_2(\lambda) - u(\lambda)| \\ 0 & |p_1(\lambda) - u(\lambda)| = |p_2(\lambda) - u(\lambda)| \\ -1 & |p_1(\lambda) - u(\lambda)| > |p_2(\lambda) - u(\lambda)| \end{cases} \quad (8)$$

Step2-1: For each unselected key feature f_λ, calculate the following weighted sum of errors for all the samples $(u_i, w_i, v_i) \in \Phi$.

$$\varepsilon = \sum_{i: v_i \neq \eta(\lambda, u_i)} w_i \quad (9)$$

Then select an index of a key feature λ_t that minimize the weighted sum of errors. When there are plural key features which minimize the weighted sum of errors, select the key feature which maximizes $|p_1(\lambda_t) - p_2(\lambda_t)|$.

Step2-2: Calculate the following confidence α_t from the weighted sum of errors.

$$\alpha_t = \frac{1}{2} \log\left(\frac{1-\varepsilon}{\varepsilon}\right) \quad (10)$$

However, $\alpha_t = 0$ when $\varepsilon \geq 0.5$.

Step2-3: Update sample weights by the following expression.

$$w_i = w_i \times \exp(-\alpha_t v_i \eta(\lambda_t, u_i)) \quad (11)$$

Step2-4: Normalize sample weights so that the sum of them is 1.

$$w_i = \frac{w_i}{\sum_{j=1}^{N_S} w_j} \quad (12)$$

Step3: Record pairs of indexes of key features and their confidences $\{(\lambda_t, \alpha_t) \mid t = 1, \ldots, T\}$.

D. Object and Scene Recognition

For a given scene image which contains several object segments, object categories are recognized based on the PLCT search and a scene category is recognized based on a result of object recognition.

Object category recognition is performed by selecting a leaf node through traversing a PLCT from its root node to leaf nodes. During the PLCT traversal, for a given object segment s, distances between a distribution of a BoF $H(s)(= [h_s(f_1), \ldots, h_s(f_{N_f})])$ and conditional probability distributions of key features of nodes are calculated and used to traverse the nodes in ascending order of their distance. This PLCT traversal is characterized by the definition of distance and the number of leaf nodes to be traversed. A leaf node with the minimum distance is selected among plural traversed leaf nodes.

$$D(s) = [d_s(f_1),\dots,d_s(f_{N_f})], d_s(f_i) = \frac{h_s(f_i)}{\sum_{j=1}^{N_f} h_s(f_j)} \quad (13)$$

Two distances are introduced for the PLCT traversal. One is a distance which simply uses a conditional probability distribution of key features itself and is called the KFPD (Key Feature Probability Distribution) distance. The KFPD distance $E(s, v)$ for a node v is defined by

$$E(s,v) = \sum_{i=1}^{N_f} | d_s(f_i) - p(f_i \mid v) | \quad (14)$$

where $p(f_i \mid v)$ is a conditional probability distribution of key features g of the node v. The other is a distance which uses a composite of boosted key features with confidences for all the branch nodes and is called the CBKF(Composite Boosted Key Features) distance. Let $\{(\lambda_t^v, \alpha_t^v) \mid t = 1,\dots,T\}$ be a set of pairs of indexes of key features and their confidences for a branch node v and let us compute the following composite confidence $\{\alpha^*\}$ for all branch nodes of a PLCT.

$$\alpha_i^* = \frac{\sum_v \hat{\alpha}_i^v}{\sum_{j=1}^{N_f} \sum_v \hat{\alpha}_j^v}, \quad i = 1,\dots,N_f \quad (15)$$

$$\hat{\alpha}_i^v = \begin{cases} \alpha_t^v & \exists t \in T, i = \lambda_t^v \\ 0 & otherwise \end{cases} \quad (16)$$

Then the CBKF distance $E(s, v^k)$ for a left child node v^1 and a right child node v^2 of the node v is defined by

$$E(s,v^k) = \sum_{i=1}^{N_f} \alpha_i^* \times | d_s(f_i) - p(f_i \mid v^k) |, \quad k = 1,2 \quad (17)$$

where $p(f_i \mid v^k)$ is a conditional probability distribution of key features of the node v^k.

In scene category recognition, selected object categories are firstly used for shortlisting candidate scene categories which are scene categories whose foreground object categories are same with selected object categories. Then a scene category is selected based on distance between a sum of BoFs of given object segments and probability distributions of key features of candidate scene categories.

III. Experiments

A. Experimental Framework

Experiments were conducted by using 429 images of 16 scene categories which were arranged from the MSRC labeled image database v2[1]. Each scene category contains about 27 images and each image of the category contains an object segment of the category in the foreground and several object segments of other categories in the background. Fig.1 shows some categorical scene images and object segments with labels. These images were split into five parts with equal size for 5-fold cross validation. Main parameters were set as follows. In determining the number of classes of V-PLCA, thresholds of the dispersion index and class probabilities were 1.0 and 0.2 respectively. In the expressions (6) and (7), thresholds θ_f and θ_b were set to 0.1. In boosting, a range of local feature points N_{u1} and N_{u2} were set to 50 and 500 respectively, the number of samples N_Φ was set to 500 and the upper limit number of boosted key features T was set to 500.

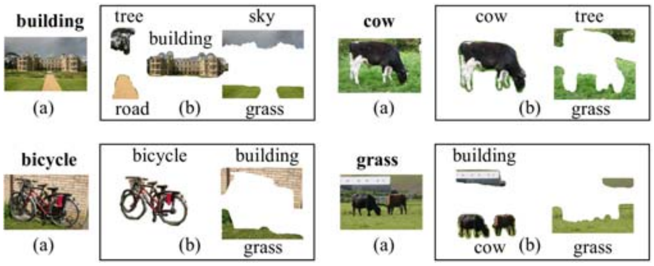

Fig. 1. Examples of (a) scene images and (b) object segments with labels. Scene images and object segments of 16 categories (“airplane”, “bicycle”, “bird”, “building”, “car”, “cat”, “chair”, “cow”, “dog”, “grass”, “road”, “sheep”, “sign”, “sky”, “tree”, “water”) were used in experiments.

Two types of local feature descriptors, the 128-dimensional gray SIFT descriptor [13] at interest points and the 384-dimensional opponent color SIFT descriptor [14] on a dense grid were used for experiments and code books were obtained by the K-tree method [15]. We abbreviate these two features as IGPS (interest point gray SIFT) and DOCS (dense opponent color SIFT) respectively. The code book sizes of IPGS and DOCS were 719 and 720 respectively.

B. Experimental Results

The mean of the total numbers of object classes which were generated by the V-PLCA from 16 scene categories was 97.6 for IPGS and DOCS and the mean depth of IPGS and DOCS PLCTs which had these classes at their leaf nodes was 11.93. Fig. 2 shows examples of foreground and background object classes and their composition ratio for scene categories. A scene category consisted of 6.1 object classes on average.

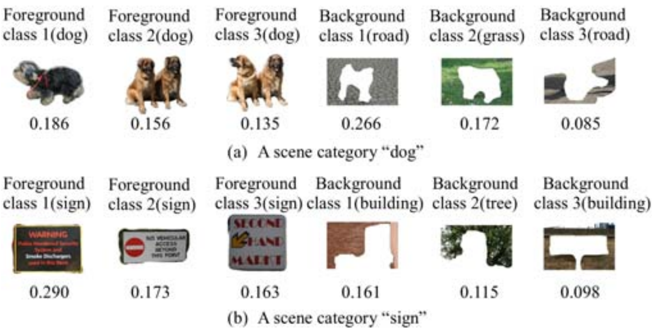

Fig. 2. Examples of object category composition of scene categories. Images are representative segments whose instance probabilities are maximal. Class probabilities which represent composition ratio are shown under those segment images.

Table I shows the mean classification accuracy of object and scene recognition for three recognition methods of the PLCT- BF search method, the PLCT search method and the direct matching method and two feature descriptors of IPGS and DOCS. The PLCT-BF search method is a recognition method based on searching the PLCT using the CBKF distance. The PLCT search method is a recognition method based on searching the PLCT using the KFPD distance. The direct matching method is a recognition method which

[1] http://research.microsoft.com/vision/cambridge/recognition/

selects an object class with the minimum distance by directly computing distances for conditional probability distribution of key features of all the object classes. The mean classification accuracy is obviously same for the PLCT search method and the direct matching method since they use the same KFPD distance. The PLCT-BF search method achieved higher classification accuracy than other methods for both of IPGS and DOCS features.

TABLE I: CLASSIFICATION ACCURACY OF OBJECT AND SCENE CATEGORIES

Recognition Method (Distance)	Feature descriptor			
	IPGS		DOCS	
	Object	Scene	Object	Scene
PLCT-BF search (CBKF distance)	0.655	0.687	0.742	0.821
PLCT search (KFPD distance)	0.649	0.676	0.728	0.807
Direct matching (KFPD distance)	0.649	0.676	0.728	0.807

Fig.3 shows the mean classification accuracy to the number of traversed leaf nodes. The PLCT-BF search method achieved higher classification accuracy for objects and scenes by the less number of leaf node traversal in comparison with the PLCT search method in both cases of IPGS and DOCS features.

C. Discussion

In our method, categorization is performed through unsupervised V-PLCA and PLCT-BF followed by semi-supervised object labeling. In the V-PLCA, the number of object classes in scene categories is not necessary to be fixed in advance and is determined dependent on learning samples. Also in the PLCT-BF, the depth of a tree is not necessary to be fixed in advance and is determined dependent on object classes generated through the V-PLCA. These characteristics of our method make it easy to adapt to various features and data sets for learning without tuning size parameters of the method. In addition, selection of key features with their confidences through boosting makes it possible to obtain discriminative key features independent of a given set of key features. Our method can learn and recognize both object and scene categories at the same time. The recognition performance depends on not only learning and recognition methods but also feature coding and pooling methods and learning data sets [16]. The PLCT-BF search method using boosted key features achieved higher classification accuracy for both objects and scenes than the direct matching method using the KFPD distance [1] and also showed competitive performance in comparison with existing methods which uses SIFT-based features [14], [16].

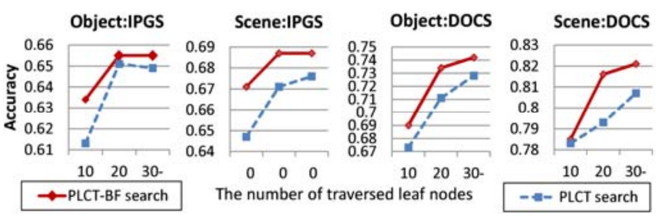

Fig. 3. Classification accuracy to the number of traversed leaf nodes.

IV. CONCLUSIONS

We have presented an object and scene categorization method based on the probabilistic latent component tree with boosted features. Through experiments by using images of plural categories in the MSRC labeled image database, it was shown that performance of object and scene recognition was high and improved by using composite boosted features in the probabilistic latent component tree search.

REFERENCES

[1] M. Atsumi, "Learning visual categories based on probabilistic latent component models with semi-supervised labeling," *GSTF International Journal on Computing*, vol. 2, pp. 88–93, 2012.

[2] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects and their location in images," in *Pro. IEEE Int. Conf. on Computer Vision*, pp. 370–377, 2005.

[3] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification via pLSA," in *Proc. the European Conf. on Computer Vision*, pp. 517–530, 2006.

[4] S. Huang and L. Jin, "A PLSA-based semantic bag generator with application to natural scene classification under multi-instance multi-label learning framework," in *Proc. of the 5th Int. Conf. on Image and Graphics*, 2009, pp. 331–335.

[5] B. Epshtein and S. Ullman, "Feature hierarchies for object classification," in *Proc. of IEEE Int. Conf. on Computer Vision*, 2005.

[6] E. Bart, I. Porteous, P. Perona, and M. Welling, "Unsupervised learning of visual taxonomies," in *Proc. IEEE CS Conf. on Computer Vision and Pattern Recognition*, 2008.

[7] [7] J. Sivic, B. Russell, A. Zisserman, W. Freeman, and A. Efros, "Unsu- pervised discovery of visual object class hierarchies," in *Proc. of IEEE CS Conf. on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[8] Z. Tu, "Probabilistic boosteing-tree: Learning discriminative models for classification, recognition and clustering," in *Proc. 10th IEEE Int. Conf. on Computer Vision*, vol. 2, pp. 1589–1596, 2005.

[9] M. Antenreiter, C. Savu-Krohn, and P. Auer, "Visual classification of images by learning geometric appearances through boosting," in *Proc. of the Second IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, pp. 233–243, 2006.

[10] Z. H. Zhou and M. L. Zhang, "Multi-instance multi-label learning with application to scene classification," in *Advances in Neural Information Processing Systems 19*, pp. 1609–1616, 2006.

[11] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in Proc. *of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 511–518, 2001.

[12] G. Csurka, C. Bray, C. Dance, and L. Fan, "Visual categorization with bags of keypoints," in *Proc. ECCV Workshop on Statistical Learning in Computer Vision*, pp. 1–22, 2004.

[13] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[14] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 1582–1596, 2010.

[15] G. Shlomo, "K-tree; a height balanced tree structured vector quantizer," in *Proc. of the 2000 IEEE Signal Processing Society Workshop*, vol. 1, 2000, pp. 271–280.

[16] Y. L. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *Proc. 2010 IEEE Conf. on Computer Vision and Pattern Recognition*, 2010, pp. 2559–2566.

Masayasu Atsumi received Doctor's degree of Engineering from Tokyo Institute of Technology in 1996. He is currently an associate professor in the Department of Information Systems Science, Faculty of Engineering at Soka University. He is a member of the Japanese Society for Artificial Intelligence, Information Processing Society of Japan, Japanese Neural Network Society, the Association for Natural Language Processing, the Robotics Society of Japan and the Association for the Advancement of Artificial Intelligence. He received a Best Paper Award in Joint 5th International Conference on Soft Computing and Intelligent Systems and 11th International Symposium on Advanced Intelligent Systems in 2010. His research interests include artificial intelligence, computer vision and human-robot interaction.