Learning Visual Categories based on Probabilistic Latent Component Models with Semi-supervised Labeling

Masayasu Atsumi

***Abstract*—This paper proposes a learning method of object and scene categories based on probabilistic latent component models in conjunction with semi-supervised object class labeling. In this method, a set of object segments extracted from scene images of each scene category is firstly clustered by the probabilistic latent component analysis with the variable number of classes, next the probabilistic latent component tree is generated as a classification tree of all the object classes of all the scene categories, and then object classes are incrementally labeled by propagating prior scene category labels and posterior object category labels given to representative object instances of some object classes as teaching signals. Through experiments by using images of plural categories in an image database, it is shown that the method works effectively in learning a labeled object category tree and object category composition of scene categories and achieves high performance for object and scene recognition.**

***Index Terms*—categorization, computer vision, labeling, learning.**

I. Introduction

THE human ability of categorization makes it possible to identify object categories and also scene categories as composites of them. The problem to be addressed in this paper is learning a classification tree of object appearance, category labels of object classes and object category composition of various scenes from a set of scene images each of which is labeled with one of plural objects in a scene. Here a labeled object in a scene is an object which is considered as a foreground object and other objects in background are unlabeled. A set of scene images whose foreground objects have the same label forms a scene category and a scene image can be contained in plural scene categories dependent on which object is considered to be in foreground. In this paper, we propose a learning method for this problem which consists of 1) the probabilistic latent component analysis [1] with the variable number of classes (V-PLCA) for clustering a set of object segments extracted from scene images in each scene category, 2) generation of the probabilistic latent component tree (PLCT) as a classification tree of all the object classes of all the scene categories and 3) semi-supervised labeling of object classes by propagating prior scene category labels and posterior object category labels given to representative object instances of some object classes as teaching signals.

As for related work, probabilistic latent variable models have been applied to learning object and scene categories [2], [3], [4], [5]. Since hierarchical representation enables systematic classification of object appearance and efficient identification of object and scene categories, there have been proposed hierarchical models for object and scene categorization [6], [7], [8], [9]. In [10], the hierarchical Latent Dirichlet Allocation has been applied to automatically discover a visual object hierarchy from a collection of unlabeled images though the depth of hierarchy is prefixed. It is known that context improves category recognition of ambiguous objects in a scene [11] and there have been proposed several methods [12], [13], [14] which incorporate context into object categorization. Our problem is closely related to recent research of multi-instance multi-label learning [15], [16] which learns multiple labels of multiple object instances in a scene image.

The one of main difference of our method from these existing ones is that it simultaneously learns a classification tree of categorical object appearance and probabilistic composition of object categories in scene categories. Another main difference is that our method incrementally learns object category labels in a semi-supervised manner.

This paper is organized as follows. Section II formulates the problem of learning object and scene categories and section III describes the proposed method in detail. Experimental results are shown in section IV and we conclude our work in section V.

II. Problem Formulation

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 foreground and other categorical objects in 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) histogram [17] of its local feature. In order to calculate a BoF histogram, first of all, grey or color SIFT descriptors [18], [19] are extracted from object segments at interest points or on a dense grid [20]. Then, all the SIFT features of all the segments are clustered by the K-tree method [21] to obtain a set of key features as a code book. Finally, a BoF histogram of each segment is calculated by using this code book. Let F

Manuscript received February 29, 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 192-8577, Japan (e-mail: matsumi@t.soka.ac.jp).

DOI: 10.5176_2010-2283_2.1.133

be a set of key features, f_n be a n-th key feature of F and N_F be the number of key features. Then an object segment s_{c,i_j} is represented by a BoF histogram of key features $H(s_{c,i_j}) = [h_{c,i_j}(f_1), ..., h_{c,i_j}(f_{N_F})]$.

Let $H_c = \{H(s_{c,i_j})|s_{c,i_j} \in S_c\}$ be a set of BoF histograms 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 problem to be solved is to compute a set of classes Q_c, which represents object categories, from H_c of each scene category $c \in C$, then to generate a classification tree of all classes $\cup_{c \in C} Q_c$ each class of which is located at a leaf of the tree and to assign object category labels to those classes.

The probabilistic latent component analysis with the variable number of classes (V-PLCA) is proposed for the first problem of computing a set of classes $Q_c = \{q_{c,r}|r = 1, ..., N_{Q_c}\}$ which represents object categories of each scene category $c \in C$, where N_{Q_c} is the number of classes in Q_c. A probability distribution of classes $\{p(q_{c,r})|q_{c,r} \in Q_c\}$, conditional probability distributions of instances, that is, object segments $\{p(s_{c,i_j}|q_{c,r})|s_{c,i_j} \in S_c\}$ for any $q_{c,r} \in Q_c$ and conditional probability distributions of key features $\{p(f_n|q_{c,r})|f_n \in F\}$ for any $q_{c,r} \in Q_c$ are calculated by this method where 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 probability distribution of key features represents feature of object categories.

The probabilistic latent component tree (PLCT) method is proposed for the second problem of generating a classification tree of all classes $\cup_{c \in C} Q_c$. The PLCT is a binary tree in which similar classes are located at close leaves where the similarity is calculated by using the conditional probability distribution of key features. Branch nodes also have the probability distribution of key features which characterizes subtrees whose roots are those branch nodes. The PLCT can be seen as a kind of thesaurus which defines a classification system of appearance of object categories.

The semi-supervised labeling method is proposed for the third problem of assigning object category labels to classes located at leaves of the PLCT. The category hypothesis rule is introduced to infer category labels of leaf nodes through branch nodes by propagating prior scene category labels and posterior object category labels incrementally given to representative object instances of some object classes as teaching signals. An instance whose conditional probability for a class is maximum is used as a representative instance for the class.

III. Proposed Method

A. Probabilistic Latent Component Analysis of Scene Categories

The problem of learning object category composition of each scene category $c \in C$ is estimating probabilities $p(s_{c,i_j}, f_n) = \sum_r p(q_{c,r})p(s_{c,i_j}|q_{c,r})p(f_n|q_{c,r})$, namely $\{p(q_{c,r})|q_{c,r} \in Q_c\}$, $\{p(s_{c,i_j}|q_{c,r})|s_{c,i_j} \in S_c, q_{c,r} \in Q_c\}$, $\{p(f_n|q_{c,r})|f_n \in F, q_{c,r} \in Q_c\}$, and the number of latent classes N_{Q_c} that maximize the following log-likelihood

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

for a set of BoF histograms $H_c = \{H(s_{c,i_j})|s_{c,i_j} \in S_c\}$. When the number of latent classes is given, these probabilities are estimated by the EM algorithm in which the following E-step and M-Step are iterated until convergence

[E-step]

$$p(q_{c,r}|s_{c,i_j}, f_n) = \frac{[p(q_{c,r})p(s_{c,i_j}|q_{c,r})p(f_n|q_{c,r})]^\beta}{\sum_{q_{c,r'}} [p(q_{c,r'})p(s_{c,i_j}|q_{c,r'})p(f_n|q_{c,r'})]^\beta} \tag{2}$$

[M-step]

$$p(f_n|q_{c,r}) = \frac{\sum_{s_{c,i_j}} h_{c,i_j}(f_n)p(q_{c,r}|s_{c,i_j}, f_n)}{\sum_{f_{n'}} \sum_{s_{c,i_j}} h_{c,i_j}(f_{n'})p(q_{c,r}|s_{c,i_j}, f_{n'})} \tag{3}$$

$$p(s_{c,i_j}|q_{c,r}) = \frac{\sum_{f_n} h_{c,i_j}(f_n)p(q_{c,r}|s_{c,i_j}, f_n)}{\sum_{s_{c,i_j'}} \sum_{f_n} h_{c,i_j'}(f_n)p(q_{c,r}|s_{c,i_j'}, f_n)} \tag{4}$$

$$p(q_{c,r}) = \frac{\sum_{s_{c,i_j}} \sum_{f_n} h_{c,i_j}(f_n)p(q_{c,r}|s_{c,i_j}, f_n)}{\sum_{s_{c,i_j}} \sum_{f_n} h_{c,i_j}(f_n)} \tag{5}$$

where β is a temperature coefficient.

The number of latent classes is determined through an 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 index, which is called the degree of scatter,

$$\delta_{c,r} = \sum_{s_{c,i_j}} \Big(\sum_{f_n} |p(f_n|q_{c,r}) - D(s_{c,i_j}, f_n)| \Big) \times p(s_{c,i_j}|q_{c,r}) \tag{6}$$

where

$$D(s_{c,i_j}, f_n) = \frac{h_{c,i_j}(f_n)}{\sum_{f_{n'}} h_{c,i_j}(f_{n'})} \tag{7}$$

for $\forall q_{c,r} \in Q_c$. Then a class whose degree of scatter takes a maximum value among all classes whose degrees of scatter are above a threshold and class probabilities are above a threshold is divided into two classes. This iterative process is continued until degrees of scatter or class probabilities of all the classes become less than those thresholds.

The latent class is divided into two classes as follows. Let q_{c,r_0} be a source class to be divided and let q_{c,r_1} and q_{c,r_2} be target classes after division. Then, for a segment $s_{c,i_j^*} = \arg\max_{i_j} \{p(s_{c.i_j}|q_{c,r_0})\}$ which has the maximum conditional instance probability and its BoF histogram $H(s_{c,i_j^*}) = [h_{c,i_j^*}(f_1), ..., h_{c,i_j^*}(f_{N_F})]$, one class q_{c,r_1} is set by specifying its conditional probability distribution of key features, conditional probabilities of instances and a class probability as

$$p(f_n|q_{c,r_1}) = \frac{h_{c,i_j^*}(f_n) + \alpha}{\sum_{f_{n'}} (h_{c,i_j^*}(f_{n'}) + \alpha)} \quad for\ \forall f_n \in F \tag{8}$$

$$p(s_{c,i_j}|q_{c,r_1}) = \begin{cases} p(s_{c,i_j^*}|q_{c,r_0}) & for\ i_j = i_j^* \\ \frac{1 - p(s_{c,i_j^*}|q_{c,r_0})}{N_{S_c} - 1} & for\ \forall i_j (i_j \neq i_j^*) \in S_c \end{cases} \tag{9}$$

$$p(q_{c,r_1}) = \frac{p(q_{c,r_0})}{2} \tag{10}$$

respectively where α is a positive correction coefficient. Another class q_{c,r_2} is set by specifying its conditional probability distribution of key features $\{p(f_n|q_{c,r_2})|f_n \in F\}$ at random, conditional probabilities of instances $\{p(s_{c,i_j}|q_{c,r_2})|i_j \in S_c\}$ as 0 for s_{c,i_j^*} and equal probability $\frac{1}{N_{S_c}-1}$ for other instances $s_{c,i_j}(i_j \neq i_j^*)$, and a class probability as $p(q_{c,r_2}) = \frac{p(q_{c,r_0})}{2}$.

The temperature coefficient β is set 1.0 until the number of classes is fixed and after that it is gradually decreased according to a given schedule of the tempered EM until convergence.

B. Probabilistic Latent Component Tree for Object Categorization

The problem of learning a classification tree PLCT of categorical object appearance is generating a binary tree of all classes $Q^* = \cup_{c \in C} Q_c$ of all the scene categories by using their conditional probability distributions of key features and class probabilities.

Let $B(Q^0)$ be a branch node where $Q^0 (\subseteq\ Q^*)$ is a set of classes which are located at leaf nodes of a subtree whose root is the branch node. Note that $Q^0 = Q^*$ for a root node of a PLCT. Then two child nodes of the parent node $B(Q^0)$ are generated as follows. First of all, for each key feature $f_n \in F$, Q^0 is divided into two subsets of classes $Q^1_{f_n} = \{q_{c,r}|p(f_n|q_{c,r}) \leq \epsilon, q_{c,r} \in Q^0\}$ and $Q^2_{f_n} = \{q_{c,r}|p(f_n|q_{c,r}) > \epsilon, q_{c,r} \in Q^0\}$ according to whether a probability value of the key feature f_n of each class in Q^0 is below ϵ or not where ϵ is 0 or a small positive value and 0 by default. Next, mean probability distributions of key features of classes in $Q^1_{f_n}$ and $Q^2_{f_n}$ are calculated as $\{\mu_{Q^1_{f_n}}(f_{n'})|f_{n'} \in F\}$ and $\{\mu_{Q^2_{f_n}}(f_{n'})|f_{n'} \in F\}$ respectively and the following distance

$$D_{f_n} = \sum_{q_{c,r} \in Q^1_{f_n}} p(q_{c,r}) \Big(\sum_{f_{n'} \in F} p(f_{n'}|q_{c,r}) \log \frac{p(f_{n'}|q_{c,r})}{\mu_{Q^1_{f_n}}(f_{n'})} \Big) \tag{11}$$
$$+ \sum_{q_{c,r} \in Q^2_{f_n}} p(q_{c,r}) \Big(\sum_{f_{n'} \in F} p(f_{n'}|q_{c,r}) \log \frac{p(f_{n'}|q_{c,r})}{\mu_{Q_2}(f_{n'})} \Big)$$

is computed based on the KL information between each and mean probability distributions of key features. Finally, Q^0 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. However, when the number of classes of a branch node $B(Q^0)$ is 2, two leaf nodes each of which has one of these two classes are generated as child nodes. 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. By the way, it rarely happens that the number of classes in either Q^1 or Q^2 becomes 0. In that case, a mean probability value $\sum_{q_{c,r} \in Q^0} p(f_n|q_{c,r})/|Q^0|$ is used as ϵ for dividing Q^0 into two subsets of classes Q^1 and Q^2 where $|Q^0|$ represents the number of classes in Q^0.

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 n_p be a branch node and n_{c1} and n_{c2} be its child nodes. For class probabilities $p(n_{c1})$ and $p(n_{c2})$ and conditional probability distributions of key features $\{p(f_n|n_{c1})|f_n \in F\}$ and $\{p(f_n|n_{c2})|f_n \in F\}$ of child nodes, the branch node has a class probability $p(n_p) = p(n_{c1}) + p(n_{c2})$ and a conditional probability distribution of key features $\{p(f_n|n_p)|f_n \in F\}$ a probability value of which is obtained by

$$p(f_n|n_p) = \frac{p(n_{c1})}{p(n_p)} \times p(f_n|n_{c1}) + \frac{p(n_{c2})}{p(n_p)} \times p(f_n|n_{c2}). \tag{12}$$

C. Learning Object Category Labels

The problem of learning object category labels is inferring labels of classes located at leaf nodes through branch nodes by propagating prior scene category labels and posterior object category labels incrementally given to representative object instances of some object classes as teaching signals. Class labels are initialized by using scene category labels when a PLCT is generated and they are incrementally modified by using object category labels given for representative instances of some object classes. An instance whose conditional probability for a class is maximum is used as a representative instance for the class. The category hypothesis rule is introduced to infer class labels of leaf nodes through branch nodes in the following steps:

1) Assign teaching signals to leaf nodes where a teaching signal of each leaf node is a label of a scene category for the leaf class at the PLCT generation time and an object category label given for a representative instance of the leaf class while incremental modification time,
2) Infer branch category hypotheses of branch nodes by propagating class probabilities of leaf nodes and teaching category labels assigned to leaf nodes,
3) Infer object category labels of leaf classes by using branch category hypotheses and teaching category labels.

In the step 2, a branch category hypothesis is inferred as follows based on class probabilities of leaf nodes and teaching category labels assigned to leaf nodes. Let $B(Q)$ be a branch node where $Q = \{q_{c,r}\}$ is a set of classes which are located at leaf nodes of a subtree whose root is the branch node and $\Gamma_{B(Q)} = \{(p(q_{c,r}), l_{c,r})|q_{c,r} \in Q, l_{c,r} \in L_c\}$ be a set of pairs of class probabilities and teaching category labels of those leaf nodes where L_c is a set of category labels. Then for each category label $l \in L_c$, $p_{B(Q),l} = \sum_{(p(q_{c,r}),l) \in \Gamma_{B(Q)}} p(q_{c,r})$ gives a certainty value that $B(Q)$ represents an object category l. Accordingly, a branch category hypothesis is obtained as $l^* = \arg\max_{l \in L_c} \{p_{B(Q),l}\}$ that gives the maximum certainty value among all the categories in L_c.

In order to infer object category labels of leaf classes in the step 3, a node attribute that is called the categorial root is introduced by using branch category hypotheses and teaching category labels. Let n_p be a branch node whose branch category hypothesis is $l_{n_p} \in L_c$, n_s be a sibling node of n_p, n_{c1} and n_{c2} be child nodes of n_p. Let $l_{n_{ck}} (k = 1, 2)$ or $l_{n_s} \in L_c$ be a branch category hypothesis if n_{ck} or n_s is a branch node or a teaching category label if n_{ck} or n_s is a leaf node. Then a branch node n_p becomes a categorial root if $l_{n_p} \neq l_{n_s}$ and $l_{n_{c1}} = l_{n_{c2}}$ hold. In case $l_{n_p} \neq l_{n_s}$ holds but $l_{n_{c1}} \neq l_{n_{c2}}$, n_{c1} or n_{c2} becomes a categorial root if it is a leaf node. An object category label of a leaf class is inferred by using the categorial root attribute of branch nodes and leaf nodes as follows: (1) if a leaf node is a categorial root, a teaching category label given to the leaf node is set as an object category label of the leaf class, (2) otherwise, a branch category hypothesis of the nearest ancestor node that is a categorial root is set as an object category label of the leaf class.

When a PLCT is generated, object category labels of leaf classes are initialized by firstly assigning their scene category labels to leaf nodes as teaching signals, then inferring branch category hypotheses and finally inferring object category labels of leaf classes based on the category hypothesis rule. An object category label of a leaf class is modified when the leaf node receives a teaching category label which is different from a current object category label. Here, a label of an instance whose conditional instance probability for the leaf class is maximum is selected as a teaching category label for the leaf node. If a given teaching category label is different from a previous teaching category label, which is a scene category label assigned to a leaf node as an initial teaching signal in case of the initial PLCT, firstly the new teaching category label is assigned to the leaf node, then branch category hypotheses and the categorial root attributes are modified for nodes on a path from the leaf node to a PLCT root node and their sibling nodes, and finally according to the category hypothesis rule, object category labels are modified for all the leaf classes which are located at leaf nodes of subtrees whose roots are these modified nodes. On the other hand, if a given teaching category label is the same as a previous teaching category label, which is a scene category label assigned to a leaf node as an initial teaching signal in case of the initial PLCT, an object category label is modified to the teaching category label. In this modification, a teaching category label given to a leaf node is propagated to neighbor leaf nodes so that their object category labels are also modified through the category hypothesis rule.

IV. Experiments

A. Experimental Framework

Experiments of category learning were conducted by using the MSRC labeled image database v2 [1]. Scene image sets of 16 categories each category of which contained about 27 images and was labeled with its foreground object were prepared and used for experiments. The total numbers of images was 429.

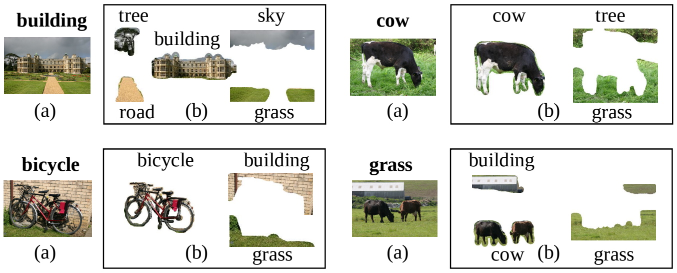


Fig. 1. Examples of (a) scene images and (b) object segments with labels. Scene images and object segments of 16 categories ("airplane", "bicycle", "bird", "building", "car", "cat", "chair", "cow", "dog", "grass", "road", "sheep", "sign", "sky", "tree", "water") were used in experiments.

An image contains a few object segments each of which has one of 16 category labels. 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, that is, each of five parts was used as a recognition test set and the others as a learning set. Main learning parameters were set as follows. In determining the number of classes of V-PLCA, thresholds of the degree of scatter and class probability were 1.0 and 0.2 respectively and a correction coefficient α in the expression (8) was 1.0. In the tempered EM, a temperature coefficient β was decreased by multiplying it by 0.95 at every 20 iterations until it became 0.8.

Two types of local feature descriptors, the 384-dimensional opponent color SIFT descriptor on a dense grid in addition to the 128-dimensional grey SIFT descriptor at interest points were used for experiments as it was known that dense representation performed better than sparsely detected interest point representation and opponent color SIFT descriptor was in general recommended among various color SIFT descriptor [20], [19]. The code book size of grey SIFT features and opponent color SIFT features were 719 and 720 respectively. These two features are abbreviated as DOCS (dense opponent color SIFT) and IPGS (interest point grey SIFT) respectively.

B. Experimental Results

The mean of the total numbers of classes which were generated by the V-PLCA from 16 scene categories were 106.2 and 89.2 for IPGS and DOCS respectively. By using scene category labels as teaching signals at the generation time, initial IPGS PLCTs had 58.6 correct object category labels for leaf classes on average and initial DOCS PLCTs had 36.6 correct object category labels for leaf classes on average, which were 55.2% and 40.8% of 106.2 and 89.8 total classes respectively. Here a correct label for a leaf class was given by a label of an instance whose conditional instance probability for the leaf class was maximum. Fig. 2 shows a part of a PLCT at the generation time. It is observed that an object category label of a leaf node "L3" is correctly inferred as a "water" by the category hypothesis rule though its scene category label is a "bird".

When teaching category labels, which were labels of object segments whose conditional instance probabilities for the leaf classes were maximum, were given to leaf nodes in the order

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

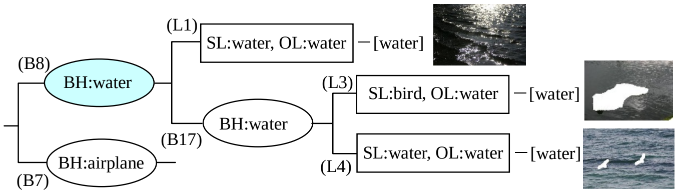
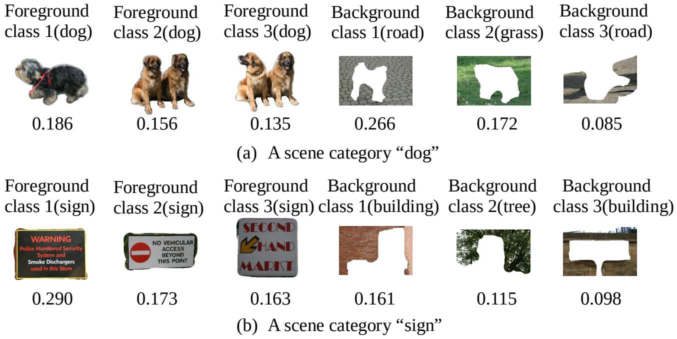


Fig. 2. A part of an initial PLCT. An ellipse is a branch node and a rectangle is a leaf node. In an ellipse, "BH" indicates a branch category hypothesis. In a rectangle, "SL" and "OL" indicate a scene category label and an object category label respectively. A colored ellipse represents it is a categorial root. A label in a square bracket is a correct label and a segment image shows a representative instance of a leaf class.



(a) A scene category "dog"

(b) A scene category "sign"

Fig. 4. Examples of object category composition of scene categories

of higher to lower class probability, IPGS PLCTs had 99.6 correct object category labels and DOCS PLCTs had 82.4 correct object category labels for leaf classes on average according to the category hypothesis rule, which were 93.8% and 91.8% of 106.2 and 89.8 total classes respectively. Fig. 3 shows a part of a modified PLCT. It is observed that object category labels of leaf nodes "L106" and "L110" are correctly inferred as "cow"s according to the category hypothesis rule by giving a teaching category label "cow" to the leaf node "L106". By another repetition of the modification of object category labels by giving same teaching category labels to the leaf nodes, the correctness became 100% since the same teaching category labels were given twice in succession.

Through learning object category labels, it turns out whether each class of a scene category represents a foreground object category or a background object category in the scene category and composition ratio of object categories in the scene category is obtained by their class probabilities. Fig. 4 shows foreground and background object classes and their composition ratio of some scene categories where a foreground class represents a foreground object category and a background class represents a background object category.

The feature of a scene category is represented by composing 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 $Q_c^f(\theta_f) = \{q_{c,r} | q_{c,r} \in Q_c^f, p(q_{c,r}) \geq \theta_f\}$ and $Q_c^b(\theta_b) = \{q_{c,r} | q_{c,r} \in Q_c^b, p(q_{c,r}) \geq \theta_b\}$ 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

$$p(f_n|Q_c^f(\theta_f), Q_c^b(\theta_b)) = \sum_{q_{c,r} \in Q_c^f(\theta_f) \cup Q_c^b(\theta_b)} \lambda(q_{c,r}) \times p(f_n|q_{c,r}) \tag{13}$$

$$\lambda(q_{c,r}) = \frac{p(q_{c,r})}{\sum_{q_{c,r'} \in Q_c^f(\theta_f) \cup Q_c^b(\theta_b)} p(q_{c,r'})} \tag{14}$$

for $\forall f_n \in F$ where θ_f and θ_b were set as 0.1 in experiments. For a given scene image, objects in the scene are recognized based on similarity between conditional probability distributions of key features for object classes and bags of features of those objects. Also a scene is recognized based on similarity between composite probability distributions of key features for scene categories and a composite bag of features of objects in the scene which is sum of bags of features of those objects. The most similar object categories and scene category are selected for the given scene image.

Two methods of recognition - the object-to-scene recognition method and the scene-to-object recognition method - are devised and their recognition performance was evaluated through 5-fold cross validation. In the object-to-scene recognition method, firstly object categories are selected for objects in a scene by computing similarity between bags of features of those objects and conditional probability distributions of key features of object classes. Then selected object categories are used for shortlisting candidate scene categories which are scene categories whose foreground object categories are same with selected object categories. Finally a scene category is selected by computing similarity between composition of bags of features of objects and probability distributions of key features of candidate scene categories. In the scene-to-object recognition method, firstly a scene category is selected by computing similarity between composition of bags of features of objects in a scene and probability distributions of key features of scene categories. Then object categories are selected by computing similarity between bags of features of objects and conditional probability distributions of key features of object categories in the scene category. Table. I shows mean recognition rates of objects and scenes by these two

TABLE I
RECOGNITION RATES OF SCENES AND OBJECTS

Recognition method	Object-to-scene		Scene-to-object	
Feature	DOCS	IPGS	DOCS	IPGS
Scene recognition accuracy	0.807	0.676	0.568	0.626
Object recognition accuracy	0.724	0.649	0.685	0.600
Foreground object recall	0.979	0.996	0.966	0.987

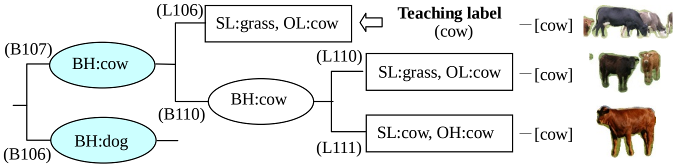


Fig. 3. A part of a modified PLCT

recognition methods. The object-to-scene recognition method achieved higher recognition performance than the scene-to-object recognition method. The DOCS representation performed better than the IPGS representation especially in case of the object-to-scene recognition method. The recognition performance was very high for foreground objects that were objects whose categories were same with their scene categories.

C. Discussion

The categorization and labeling of objects and scenes are essential to recognize and understand the world. In our method, categorization is achieved by unsupervised V-PLCA and PLCT and labeling is achieved by semi-supervised manner through the category hypothesis rule. 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, the depth of an object class tree is not necessary to be fixed in advance and is determined dependent of 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. Since an object class labeling process is incremental, it can be performed in an interactive mode in which a system equipped with this method communicates with its users to acquire object labels in a real world situation.

Our method can learn and recognize both object and scene categories at the same time. The DOCS-based object-to-scene recognition method achieved recognition rates of 0.807 for scenes and 0.724 for objects, especially 0.979 for foreground objects. The recognition performance depends on not only learning and recognition methods but also feature coding and pooling methods and learning data sets [22], [23]. Our results are high enough in comparison with existing methods which uses simple SIFT-based features [19], [22].

V. Conclusions

We have proposed a learning method of both object and scene categories based on probabilistic latent component models V-PLCA and PLCT in conjunction with semi-supervised object class labeling. Through experiments by using images of plural categories in the MSRC labeled image database, it was shown that the method worked effectively in learning a labeled object category tree and object category composition of scene categories. It was also confirmed that the proposed object-to-scene recognition method achieved high recognition performance for object and scene categories. We are currently extending our object and scene representation to include mid-level features.

References

[1] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, vol. 42, pp. 177–196, 2001.

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

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

[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*, 2004, pp. 331–335.

[5] M. Atsumi, "Learning visual object categories and their composition based on a probabilistic latent variable model," in *Neural Information Processing: Theory and Algorithms (Proceedings of ICONIP 2010, Part I)*, ser. Lecture Notes in Computer Science, vol. 6443. Springer-Verlag, 2010, pp. 247–254.

[6] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky, "Describing visual scenes using transformed objects and parts," *International Journal of Computer Vision*, vol. 77, pp. 291–330, 2008.

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

[8] L. Zhu and Y. Chen, "Unsupervised learning of a probabilistic grammar for object detection and parsing," in *Advances in Neural Information Processing Systems 19*, 2006.

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

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

[11] M. Bar, "Visual objects in context," *Nature Reviews Neuroscience*, vol. 5, pp. 617–629, 2004.

[12] A. Torralba, "Contextual priming for object detection," *International Journal of Computer Vision*, vol. 53, pp. 169–191, 2003.

[13] A. Rabinovich, C. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in context," in *Proc. of IEEE Int. Conf. on Computer Vision*, 2007.

[14] C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-occurrence, location and appearance," in *Proc. of IEEE CS Conf. on Computer Vision and Pattern Recognition*, 2008.

[15] 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*, 2006, pp. 1609–1616.

[16] Z.-J. Zha, X.-S. Hua, T. Mei, J. Wang, G.-J. Qi, and Z. Wang, "Joint multi-label multi-instance learning for image classification," in *Proc. of IEEE CS Conf. on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

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

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

[19] 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.

[20] F. Jurie and B. Triggs, "Creating efficient codebook for visual recognition," in *Proc. of 10th IEEE Int. Conf. on Computer Vision*, 2005, pp. 604–610.

[21] 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.

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

[23] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *Computer Vision - ECCV 2006, Part IV*, ser. Lecture Notes in Computer Science, vol. 3954. Springer-Verlag, 2006, pp. 490–503.

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.