Probabilistic Object Learning through
Attention-based Organized Perception

Masayasu Atsumi
Dept. of Information Systems Science, Faculty of Engineering, Soka University
1-236 Tangi, Hachioji, Tokyo 192-8577, Japan
Email: matsumi@t.soka.ac.jp

Abstract—This paper proposes a model of probabilistic object learning in conjunction with attention-based organized perception. This model consists of the following two sub-models: the former is a model of attention-based perceptual organization of segments and the latter is a model of learning object composition of categories based on a bag of features representation of segments. In attention-based perceptual organization, concurrent figure-ground segmentation is performed on dynamically-formed Markov random fields around salient preattentive points and co-occurring segments of objects and their context are grouped in the neighborhood of selective attended segments. In probabilistic learning of categorical object composition, multi-class classifiers are learned based on intra-categorical probabilistic latent component analysis with variable number of classes and inter-categorical typicality analysis. Through experiments by using images of plural categories in an image database, it is shown that the model learns a probabilistic structure of intra-categorical composition of objects and context and inter-categorical difference.

I. INTRODUCTION

In crowded scenes, a human can identify a variety of objects in a few abstract levels of their category. However, those objects and their categories have not necessarily been learned one by one in a supervised manner but learned in a semi-supervised or unsupervised manner. In this process, attention and organized perception performs an important role since they select spatially circumscribed regions of relevant objects and their co-occurring context for learning and recognition. In this paper, we propose a model of probabilistic object learning in conjunction with attention-based organized perception. This model consists of the following two sub-models: the former is a model of attention-based perceptual organization of segments and the latter is a model of learning object composition of categories based on a local feature representation of segments. In attention-based perceptual organization of segments, concurrent figure-ground segmentation is performed on dynamically-formed Markov random fields (MRFs) around salient preattentive points and co-occurring segments of objects and their context are grouped in the neighborhood of selective attended segments [1]. In order to represent local feature of segments, the SIFT features [2] of salient points in co-occurring segments are clustered by the K-tree method [3] to generate a set of key features (a code book) and a bag of features (BoF) [4] of each segment is calculated by using this code book. In probabilistic learning of categorical object composition, multi-class classifiers are learned based on intra-categorical probabilistic latent component analysis (PLCA) [5], [6] with variable number of classes and inter-categorical typicality analysis. The intra-categorical PLCA of each category learns a probabilistic structure of categorical objects and their context from BoF histograms of segments in the category. The inter-categorical typicality analysis judges how typical each class is in each category and synthesizes a typical probabilistic structure for each category. These processes are performed on a collection of categorized image sets which contain categorical objects in crowded scenes. This model makes it possible to learn intra-categorical composition of objects and context and inter-categorical difference from cues that categorical objects exist somewhere in crowded scenes.

As for visual attention, there have been proposed a lot of computational models, in which a saliency map model [7] is well-known and have a great influence on later studies [8], [9]. Though a saliency map represents space-based attention, it is known that attention works in two distinct and complementary modes of a space-based mode and an object-based mode [10]. It is also known that segmentation is largely autonomous process which precedes and guides allocation of attention [11] and organized percept of segments tends to attract attention automatically [12]. Our model is unique as it links spatial preattention and object-based attention through figure-ground segmentation on dynamically-formed MRFs and groups segments in the neighborhood of selective attended segments. As for incorporating context into object categorization, it is known that context improves recognition of ambiguous objects in a scene [13] and several computational models are recently proposed [14], [15], [16], [17]. The main difference of our method from those existing ones is that it uses attended co-occurring segments for learning and it learns a probabilistic structure of typical and non-typical objects and their context in each category.

This paper is organized as follows. Section II presents attention-based organized perception. Section III describes probabilistic learning of object composition of categories. Experimental results are shown in section IV and we conclude our work in section V.

II. ATTENTION-BASED ORGANIZED PERCEPTION

The model of attention-based organized perception consists of a saliency map for preattention, a collection of dynamically-formed MRFs for figure-ground segmentation, a visual work-
ing memory for maintaining segments and perceptually organizing them around selective attention, and an attention system on a saliency map and a visual working memory [1]. Fig. 1 depicts the organization and the computational steps of the model which are explained in subsections II-A and II-B. The detail of the saliency map is described in [9] and a related model of attention-based segmentation is described in [18]. As features of an image, brightness, hue and their contrast are obtained on a Gaussian resolution pyramid of the image. Brightness contrast and hue contrast are respectively computed by convolbing brightness and hue with a LoG/Laplacian of a Gaussian) kernel. However, since a hue value represents a color category by an angle in $[0, 2\pi]$ on a continuous color spectrum circle, hue contrast is obtained by performing convolution for hue difference of each point with its neighboring points.

A saliency map is obtained by calculating saliency from brightness contrast and hue contrast on each level of the Gaussian resolution pyramid and integrating the multi-level saliency into one map. In the first step of attention-based perceptual organization, preattentive points are stochastically selected from a saliency map according to their degrees of saliency.

In the second step, figure-ground labeling is iterated by gradually expanding 2-dimensional MRFs of brightness and hue around preattentive points by a certain margin until figure segments converge or a specified iteration is reached. If plural figure segments satisfy a merge condition, they are merged into one segment. The figure-ground labeling on a MRF is formulated as follows. Let $L = \{1, -1\}$ be a set of segment labels where “1” represents a figure label and “−1” represents a ground label and let $\bar{z} = (b, h)$ be an observation of features where $b$ is brightness and $h$ is hue. Let $W$ be a domain of a MRF and let $l = (l_w; w \in W, l_w \in L)$ be segment labels on $W$. Then, for a given observed feature $z = \{z_w\}_{w \in W}$, the problem of estimating segment labels is solved by using the EM procedure with the mean field approximation [19]. The mean field local energy function using mean field approximation is defined by

$$U^{mf}(l_w) = \sum_{w' \in B_w} V(l_w, \{l_{w'}\}) = -\frac{\eta}{8} \sum_{w' \in B_w} (l_w \times \{l_{w'}\}) \quad (2)$$

where $V$ is potential of a pair-site clique, $B_w$ is the 8-neighborhood system, $\eta$ is an interaction coefficient which is preset in this study, $\{l_{w'}\}$ is an expectation of a segment label in the neighborhood, $t$ is the EM iteration number and $\Phi$ is a parameter set that provides distributions of $p(z/l, \Phi)$. Concretely, $\Phi$ is means and variances of multivariate Gaussian distributions of figure and ground features. Then, a posterior probability of a segment label is given by

$$p^{mf}(l_w|z_w, \Phi^{(t)}) \approx \frac{\exp(-U^{mf}(l_w|z_w, \Phi^{(t)}))}{\tilde{H}^{mf}}$$

where $\tilde{H}^{mf}$ is the partition function and an expectation of a segment label is obtained as

$$\langle l_w | z_w \rangle = \sum_{l_w \in L} (l_w \times p^{mf}(l_w|z_w, \Phi^{(t)})). \quad (4)$$

In the second step of attention-based perceptual organization, for each extracted segment, the attention degree of the segment is calculated from its saliency, closedness and attention bias for object-based attention. Saliency of a segment is defined by the maximum value of a figure segment and its surrounding ground segment. The degree of spot attention is defined by the distance between mean features (brightness and hue) of a face by simply using its hue and aspect ratio. Then, the attention degree of a segment $s$ is defined by

$$A(s) = p(s, \nu) \times (\kappa_s \times A_s(s) + \kappa_p \times A_p(s) + \kappa_0 \times A_0(s)) \quad (5)$$

where $A_s(s)$ is the degree of surface attention, $A_p(s)$ is the degree of spot attention, $A_0(s)$ is the attention bias, and $\kappa_s,$
\(\kappa_p, \kappa_b (\kappa_s + \kappa_p = 1, \kappa_b \geq 0)\) are weighting coefficients for them respectively. The function \(\rho(s, \nu)\) takes 1 if a segment \(s\) is closed and \(\nu\) otherwise, where \(\nu (0 \leq \nu < 1)\) is the decrease rate of attention when the segment isn’t closed.

In the forth step, from these segments, the specified number of segments whose attention degree are larger than others are selected as selective attended segments. In the fifth step of organized perception, each selective attended segment and its neighboring segments are grouped as a co-occurring segment. If two sets of co-occurring segments overlap, they are combined into one co-occurring segment. This makes it possible to group part segments of an object or group salient contextual segments with an object.

### III. Probabilistic Learning of Object Composition

The problem to be addressed is statistically learning object composition of each category from a cue that the categorical objects exist somewhere in crowded scenes. Concretely, given a collection of categorized image sets which contain the categorical objects in crowded scenes, it is required to learn a probabilistic structure of categorical objects and their context from attended co-occurring segments extracted in images of each category. In the proposed method, intra-categorical PLCA with variable number of classes is firstly applied to each category and learns a multi-class classifier from BoF histograms of co-occurring segments in the category. Then each class in each category is judged how typical in the category through inter-categorical typicality analysis. We call this learning method a probabilistic latent component forest (PLCF).

#### A. Object Representation by using a BoF

An object segment is represented by a BoF histogram [4] of local feature of its salient points. In order to calculate the BoF histogram, first of all, any points in a segment whose saliency are above a given threshold are extracted as salient points at each level of a multi-level saliency map. As a local feature, a 128-dimensional SIFT feature [2] is calculated for each salient point at its resolution level. Then, all the SIFT features of all segments are clustered by the K-tree method [3] 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.

As a result, feature of co-occurring segments is represented by BoF histograms.

In this paper, let \(s_{c,i,j}\) be a segment \(j\) extracted from an image \(i\) of a category \(c\), \(S_c\) be a set of segments extracted from any images of a category \(c\), and \(N_{c,i}\) be the number of segments in \(S_c\). Let \(f_n\) be a \(n\)-th element of key features \(F\), \(N_f\) be the number of key features and \(\{H(s_{c,i,j}, f_n)\}_{f_n \in F}\) be a BoF histogram of a segment \(s_{c,i,j}\). Let \(q_{c,r}\) be a latent class of a category \(c\), \(Q_c\) be a set of latent classes of a category \(c\), \(N_{c,q}\) be the number of classes of a category \(c\), \(C\) be a set of categories and \(N_c\) be the number of categories.

#### B. Learning Multi-class Classifiers of Object Categories

The problem of learning a multi-class classifier for segments in a category \(c\) 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 \(\{p(q_{c,r})|q_{c,r} \in Q_c\}, \{p(s_{c,i,j}|q_{c,r})|s_{c,i,j} \in S_{c,q}, q_{c,r} \in Q_c\}, \{p(f_n|q_{c,r})|f_n \in F, q_{c,r} \in Q_c\}\), and the number of latent classes \(N_{c,q}\) that maximize the following log-likelihood

\[L_c = \sum_{i,j} H(s_{c,i,j}, f_n) \log p(s_{c,i,j}, f_n).\]  
(6)

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**

\[p(q_{c,r}|s_{c,i,j}, f_n) = \frac{[p(q_{c,r})p(s_{c,i,j}|q_{c,r})p(f_n|q_{c,r})]^{\beta}}{\sum_{q_{c,r'}}[p(q_{c,r'})p(s_{c,i,j}|q_{c,r'})p(f_n|q_{c,r'})]^{\beta}}.\]  
(7)

**M-step**

\[p(f_n|q_{c,r}) = \frac{\sum_{s_{c,i,j}} H(s_{c,i,j}, f_n)p(q_{c,r}|s_{c,i,j}, f_n)}{\sum_{s_{c,i,j}} \sum_{f_n} H(s_{c,i,j}, f_n)p(q_{c,r}|s_{c,i,j}, f_n)}\]  
(8)

\[p(q_{c,r}) = \frac{\sum_{s_{c,i,j}} \sum_{f_n} H(s_{c,i,j}, f_n)p(q_{c,r}|s_{c,i,j}, f_n)}{\sum_{s_{c,i,j}} \sum_{f_n} H(s_{c,i,j}, f_n)}\]  
(9)

where \(\beta\) 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,

\[\delta_{q_{c,r}} = \frac{\sum_{s_{c,i,j}} (\sum_{f_n} p(f_n|q_{c,r}) - D(s_{c,i,j}, f_n))) \times p(s_{c,i,j}|q_{c,r})}{N_f \times N_{c,q}}\]  
(11)

where

\[D(s_{c,i,j}, f_n) = \frac{H(s_{c,i,j}, f_n)}{\sum_{f_n'} H(s_{c,i,j}, f_n')}\]  
(12)

for \(\forall q_{c,r} \in Q_c\). Then a class whose degree of scatter takes a maximum value among all classes is divided into two classes. This iterative process is continued until \(\delta_{q_{c,r}}\)-values for all classes become less than a certain threshold. 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 \(s_{c,i,j} = \arg\max_i \{p(s_{c,i,j}|q_{c,r_0})\}\) which has the maximum conditional probability and its BoF histogram \(H(s_{c,i,j}, f_n) = [h_{c,i,j}(1), ..., h_{c,i,j}(N_f)]\), one class \(q_{c,r_1}\) is set by specifying its conditional probability of key features, conditional probabilities of segments and a class probability as

\[p(f_n|q_{c,r_1}) = \frac{h_{c,i,j}(n) + \alpha}{\sum_{n'}(h_{c,i,j}(n') + \alpha)} \quad \forall f_n \in F\]  
(13)

\[p(s_{c,i,j}|q_{c,r_1}) = p(s_{c,i,j}|q_{c,r_0}) \quad \forall i_j \in S_c\]  
(14)
\[ p(q_{c,r}) = \frac{p(q_{c,r})}{2} \]  

respectively where \( \alpha \) is a positive correction coefficient. Another class \( q_{c,r} \) is set by specifying its conditional probability of key features \( \{ p(f_n|q_{c,r}) | f_n \in F \} \) at random, conditional probabilities of segments \( \{ p(s_{c,r} | q_{c,r}) | s_{c,r} \in S_c \} \) as 0 for \( s_{c,i} \) and equal for other segments, and a class probability as \( \frac{p(q_{c,r})}{2} \). As a result of subsequent class division, latent classes form a binary tree, which we call a probabilistic latent component tree (PLCT).

The temperature coefficient \( \beta \) 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.

C. Inter-categorical Typicality Analysis

Inter-categorical typicality analysis evaluates each latent class of each category whether or not it is typical in the category and calculates a conditional probability of key features for the category by synthesizing those probabilities of its typical classes. The typicality of a class is judged based on whether it appears in its category with high frequency but does not appear in other categories only with low frequency. Typical classes consist of classes of object segments and co-occurring contextual segments which distinguish a category and a synthesized conditional probability of key features encodes characteristics of the category. Here, in general, co-occurring contextual segments are objects of other categories or background. By the way, exceptional object segments, as well as typical object segments, are also encoded by conditional probabilities of key features of some non-typical classes.

For the inter-categorical typicality analysis, let the distance between classes \( q_{c_1,r_1} \in Q_{c_1} \) and \( q_{c_2,r_2} \in Q_{c_2} \) of any different categories \( c_1 \) and \( c_2 \) be

\[ L_1(q_{c_1,r_1}, q_{c_2,r_2}) = \frac{\sum_{f_n} |p(f_n|q_{c_1,r_1}) - p(f_n|q_{c_2,r_2})|}{2} \]  

by using their conditional probabilities of key features. Then the distance of a class \( q_{c,r} \in Q_c \) of a category \( c \) from classes of other categories is defined as

\[ d(c, q_{c,r}) = \frac{\sum_{c' \in C - c} \sum_{q_{c',r'} \in Q_{c'}} (L_1(q_{c,r}, q_{c',r'}) \times p(q_{c',r'}) \times p(q_{c,r}))}{N_c - 1} \]  

Now, for the mean distance \( \bar{d}(c) \) is given by

\[ d(c, q_{c,r}) = \frac{\sum_{q_{c,r} \in Q_c} d(c, q_{c,r})}{N_c} \]  

for all classes of a category \( c \). The deviation of a distance \( d(c, q_{c,r}) \) is given by \( \Delta(c, q_{c,r}) = d(c, q_{c,r}) - d(c) \). Then the typicality index of a class \( q_{c,r} \) of a category \( c \) is defined as

\[ \gamma(q_{c,r}) = \frac{1}{1 + \exp \left( -\mu \left( p(q_{c,r}) - \left( \frac{1}{N_c} - \Delta(c, q_{c,r}) \right) \right) \right)} \]  

where \( \mu \) is a gain coefficient. This index is called the degree of categorical class.

The conditional probability for a category \( c \) is defined for a set of typical classes \( Q_c^t = \{ q_{c,r} | \gamma(q_{c,r}) \geq \theta, q_{c,r} \in Q_c \} \) as

\[ p(f_n|Q_c^t) = \frac{\sum_{q_{c,r} \in Q_c^t} (\lambda(q_{c,r}) \times p(f_n|q_{c,r}))}{\sum_{q_{c,r} \in Q_c^t} p(q_{c,r})} \]  

where \( \theta \) is a threshold which determines whether or not a class is typical.

IV. EXPERIMENTS

A. Experimental Framework

To evaluate probabilistic learning of object composition of categories through attention-based organized perception, experiments were conducted by using images of the Caltech-256 image database [20]. For each of 20 categories, 4 images, each of which contains one or a few categorical objects in a crowded scene, were selected and used for experiments. Fig. 2 shows some categorical images.

Main parameters are set as follows. The number of levels of an image pyramid is 5. As for attention-based organized perception, an interaction coefficient \( \eta \) is 1.5, weighting coefficients and a decrease rate for the attention degree of segments in the expression (5) are \( \kappa_s = 0.5, \kappa_p = 0.5, \kappa_s = 1.0 \) and \( \nu = 0.2 \) respectively, and the upper bound number of selective attention is 4. As for learning of multi-class classifiers, a threshold for salient points is 0.1, a threshold of class division is 0.07 and a correction coefficient \( \alpha \) in the expression (13) is 2.0. In the tempered EM, a temperature coefficient \( \beta \) was decreased by multiplying it by 0.95 at every 20 iterations until it became 0.8. As for inter-categorical typicality analysis, a gain coefficient \( \mu \) in the expression (18) is 5.0 and a threshold \( \theta \) for determining typical classes is 0.47.

In learning object composition of categories, the number of salient points (that is, SIFT features) which were extracted...
from all segments is 76019. The codebook size of key features which were obtained by the K-tree method is 438. The BoF histograms were calculated for 181 segments whose number of salient points were more than 100.

B. Experimental Results

Learning was performed for a set of co-occurring segments extracted from 20 categories by the attention-based organized perception. Fig. 3 shows co-occurring segments and their labels for some categorical images which were extracted by the attention-based organized perception. There were observed three types of co-occurring segments. The first type of co-occurring segments represents organized perception in which an object consists of one segment and it is grouped with its contextual segments. Examples of “telephone-box” and “hibiscus” in Fig. 3 show organized perception of this type. The second type of co-occurring segments represents organized perception in which each co-occurring segment is a part of an object and the object consists of those segments. Examples of “people” and “school-bus” in Fig. 3 show organized perception of this type. The third type of co-occurring segments represents organized perception in which an object consists of plural segments and it is also grouped with its contextual segments. Examples of “chimp” and “butterfly” in Fig. 3 show organized perception of this type.

Learning results were analyzed for class composition of categorical multi-class classifiers and characterization of categories by conditional probability of key features.

Fig. 4 shows PLCTs of multi-class classifiers for some categories. In Fig. 4, a typical segment of a class \( r \) of each category \( c \) is a segment \( s_{c,i,j} \) that maximizes \( p(q_{c,r} | s_{c,i,j}) \). Also, a typical co-occurring segment of each category \( c \) is a co-occurring segment \( s_c = \{ s_{c,i,k} | k \in K \} \) that maximizes the following typicality index \( R(s_c) = \sum_{k \in K} \max_{g_{c,r} \in Q_c} p(g_{c,r} | s_{c,i,k}) \), where \( Q_c \) is a set of typical classes of the category \( c \). The mean number of classes and typical classes per PLCT for 20 categories are 7.55 and 4.05 respectively. As shown in Fig. 4, typical classes mainly distinguish segments of categorical objects but classes of frequent co-occurring contextual segments also become typical ones. For example, all the 3 typical classes of a “butterfly” category distinguish segments of the categorical object, that is, segments of butterflies. In a “hibiscus” category, 4 classes out of 5 typical classes distinguish segments of the categorical object and another one distinguishes contextual segments. Also in a “helicopter” category, 2 classes out of 3 typical classes distinguish segments of the categorical object and another one distinguishes contextual segments. According to the expression (18) of the degree of categorical class, a class becomes typical if it has a high class probability and its feature does not appear in other categories. In a “hibiscus” category, two classes of categorical objects with low class probabilities are selected as typical classes because their feature does not appear in other categories. Two classes of contextual segments with high class probabilities remain non-typical because their feature is shared in many categories. On the other hand, in a “helicopter” category, there exists a non-typical class of a categorical object since its feature is exceptional in the category.

A conditional probability of key features for a category is a weighted sum of conditional probabilities of key features for its typical classes with their class probabilities. Fig. 5 shows conditional probabilities of key features for all categories and Fig. 6 shows distance between each pair of them which is defined by the expression (16). Each category has a different
distribution of conditional probability of key features and the mean distance of all pairs of categories is 0.62. It is possible to distinguish each category from others by their conditional probabilities of key features since the distance between them is large and they encode mainly features of typical categorical objects through selective attention and typicality analysis.

![Conditional probabilities of key features for all categories.](image)

Fig. 5. Conditional probabilities of key features for all categories.

The main future work is to build an integrated category and object recognizer which makes full use of this probabilistic structure.

**REFERENCES**


**V. CONCLUSIONS**

We have proposed a probabilistic model of learning categorical object composition through attention-based organized perception. In this model, object composition of categories is learned from a set of BoF of attended co-occurring segments based on intra-categorical probabilistic latent component analysis with variable number of classes and inter-categorical typicality analysis. Through experiments by using images of plural categories in the Caltech-256 image database, it was shown that, by the attention-based organized perception, the model extracted a set of co-occurring segments which represented objects and their context and that, from those co-occurring segments, the model learned a probabilistic structure which distinguished intra-categorical composition of objects and context and inter-categorical difference.

![Distance between conditional probabilities of categorical key features.](image)

Fig. 6. Distance between conditional probabilities of categorical key features.