2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
November 3-7, 2013. Tokyo, Japan

Formation of Hierarchical Object Concept Using Hierarchical Latent Dirichlet Allocation

Yoshiki Ando, Tomoaki Nakamura, Takaya Araki, Takayuki Nagai

Abstract— **In recent studies, it has been revealed that robots can form concepts and understand the meanings of words through inference. The key idea underlying these studies is "multimodal categorization" of a robot's experience. However, previous studies considered only nonhierarchical categorization methods, which led to nonhierarchical concept structures. Our concepts have a hierarchical structure, thus ensuring that the resulting inferences are more efficient and accurate. In this paper, we propose a novel hierarchical categorization method. The method involves extending multimodal latent Dirichlet allocation (MLDA) to hierarchical MLDA using the nested Chinese restaurant process, which makes it possible for robots to acquire concepts in a hierarchical structure. We show that a robot can form a hierarchical concept structure based on self-obtained multimodal information. Moreover, by focusing on the common features of each category in the hierarchy, the robot is able to infer unobserved information including word meanings.**

I. Introduction

Categorization of things plays an important role in human cognition [1]. By forming a category, humans can obtain more information with a minimum reference to their experiences [2]. The importance of categorization is prediction using experience-based categories. Humans predict unknown things. In addition, we consider "concept" categories that have been classified by self-organization, and concept–word links formed by the categorization have led to the understanding of a word's meaning [3]. As such, it is considered prediction based on the categorization is the bedrock of human intelligence flexibility. Therefore, it is important that intelligent robots have such a capability [3].

A method based on latent Dirichlet allocation (LDA) [4], which is one of the statistical models in the field of natural language processing, has been proposed [5], [6]. In these studies, via LDA-based object clustering using multimodal information such as visual, auditory, and tactile, it was shown that robots can categorize (conceptualize) objects such as tambourines, maracas, and stuffed animals in line with human senses. However, those authors considered only the formation of object categories that do not capture hierarchical relationships, which is an inadequate model for representing the concept of a human. Each category formed by humans is not necessarily independent, and such categories form an interrelated hierarchical structure. Using this hierarchy, humans can make predictions using an appropriate granularity category even if the objects are unknown. For example, objects such as maracas and tambourines are part of the category "percussion." In addition, they belong to a category called "instruments" with many other objects. If a robot learns a hierarchical relationship, it can predict the properties and functions of objects even though they may not have features similar to those of maracas and tambourines, but have the characteristic features of an instrument. By contrast, the LDA-based, conventional, nonhierarchical clustering can form a subordinate instrument concept by combining the concepts of maracas and tambourines. However, it may not be able to predict the nature and features of the instrument from an object that has features different from those of maracas and tambourines. In other words, the members of each category within a hierarchy of a hierarchically conceptual structure have some common features. By focusing on these common features, a robot can acquire concepts with a range of granularity.

In this paper, we propose a method for robots to form a hierarchical concept via hierarchical multimodal LDA (hMLDA). This is an extension of hierarchical LDA (hLDA) [11], a method in which the nested Chinese restaurant process (nCRP) is applied to LDA. In [11], hLDA was applied for the clustering of documents by topic, and each topic was expressed as a path in a tree structure. In this model, a word is generated according to the degree of sharing of each node. That is, a word generated at a higher-ranked node is shared among two or more topics and, therefore, represents a more extensive category. In the proposed hMLDA, considering documents as objects, topics as categories, and words as features generated from objects, it is possible to categorize the objects in an unsupervised manner.

There have been studies on the unsupervised learning of object categories using only visual information [12], [13], [14], [15]. Furthermore, in recent years, studies on categorization without a teacher have been conducted using point clouds acquired with a laser range finder or a time-of-flight (TOF) camera [16]. However, in these studies, only visual information is used, and a hierarchical structure is not considered. Moreover, those studies aimed to find and recognize an object category, while this study endeavors to predict unobservable information. Therefore, it is a very important point that an understanding of an object as well as that of the meaning of a word are realizable using robots. [17], [18], [19], [20] are mentioned as research on layered category structure. These studies did not aim to understand word meanings and form concepts through categorization,

Yoshiki Ando, Tomoaki Nakamura, Takaya Araki, and Takayuki Nagai are with Dept. of Mechanical Engineering and Intelligent Systems, The University of Electro-Communications, 1-5-1 Chofugaoka Chofu-shi, Tokyo 182-8585, Japan { y4422, naka_t, taraki, tnagai} @apple.ee.uec.ac.jp

but pursued category recognition from the computer vision viewpoint. Therefore, this study aims to improve recognition performance by taking into account hierarchical category structures. As stated before, the authors have been studying object categorizaton using multimodal information, and it has been indicated that multimodal information acquired by the robot is important for object concept formation. However, layered category structures have not been considered in our previous studies. This study shows that highly precise concept-based prediction can be achieved through the formation of concepts having layered structures. For robots, Ogata et al. proposed a method for learning the movement–sound relationship using recurrent neural network (RNN) [23]. However, hierarchical conceptual structures are not considered in these studies. Moreover, practically, RNN is considered to have scalability issues. For example, in [23] only five objects are treated. Consequently, it is unclear as to how complicated things can be treated. In contrast, the statistical model used in this paper is conventionally applied to the clustering of several thousands of documents. Currently, robots can handle several tens to hundreds of objects. Given the number of objects used in everyday life, a model that can treat tens of thousands of objects should be considered in the future.

II. Multimodal Categorization and Concept Formation

Here, we outline categorization, the robot's concept formation, and the relationship of understanding of the meaning of a word. For details, please refer to [3]. In this section, we outline MLDA and describe the layered structure of a concept.

A. Categorization and Concept Formation

In this paper, each category is formed via clustering of the robot-collected multimodal information, and these categories are considered as concepts. Concepts have been expressed as clusters in the feature space, and it is possible to predict unobservable information from some input using the cluster. In addition, language-related information is part of the feature space, and the concept-based prediction mechanism serves as an understanding of the meaning of a word. In order to realize such a clustering and prediction framework, we use multimodal LDA (MLDA), as described below.

B. Multimodal LDA

MLDA [3] is an extension of LDA that can classify multimodal information. It is represented using the graphical model shown in Fig. 1. In the figure, w^v, w^a, w^h, and w^w denote visual, auditory, haptic, and word information, respectively. In addition, β^* is determined from the Dirichlet prior distribution with parameter θ. z represents object category, as generated from a multinomial distribution with parameter θ. Similarly, θ is determined from the Dirichlet prior distribution with a parameter α. The categorization problem involves estimating the model parameters using observed multimodal information. It can be seen from Fig. 1 that MLDA offers the framework for stochastically predicting unobservable information. This serves as the bases of prediction-based understanding.

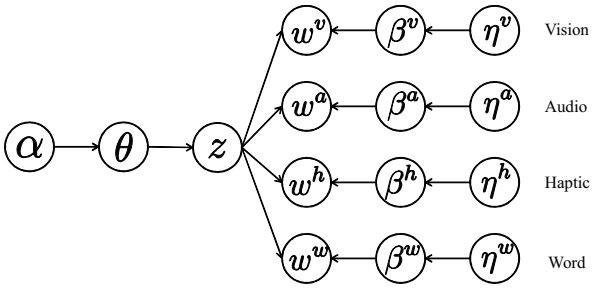

Fig. 1. Graphical model of MLDA.

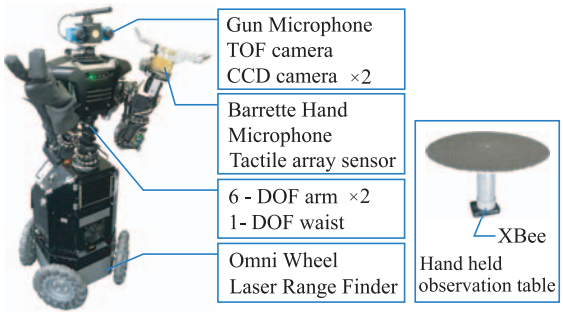

Fig. 2. Robot platform used in this paper.

C. Hierarchical Conceptual Structure

To ensure that the abovementioned MLDA classifies feature space uniformly, the formed concept is designed as a nonhierarchical structure.

Therefore, it is difficult to predict the property of such an abstract concept. For example, the concept of a maraca or a tambourine is included within a dominant concept such as a musical instrument. The dominant concept is formed by embedding narrower concepts in it using MLDA. However, the concept that suitably represents a musical instrument is not necessarily formed through combinations of subconcepts within the dominant concept, such as that of a maraca and tambourine. Therefore, it becomes difficult to recognize an unknown object that has the features of a musical instrument but differs from those of a maraca or tambourine. This problem can be solved by the concept model that considers a hierarchical structure. In section IV, we extend MLDA to hierarchical MLDA.

III. Multimodal information

Fig. 2 shows the robot platform used in this experiment. A robot finds an object and acquires multimodal information autonomously. Here, we describe the acquisition of multimodal information and its processing.

1) Visual Information: The target object is segmented out in each image frame; thereafter, 128-dimensional DSIFT [21] descriptors are computed. In a later experiment, 36 image frames of each object are captured. Three hundred to 400

feature vectors are extracted from each image, resulting in about 10000–15000 features for each object. Each feature vector is vector quantized using a codebook with 500 clusters. The codebook is generated beforehand using a k-means algorithm. Finally, a 500-dimensional histogram is built as the bag-of-features representation.

2) Auditory Information: Sound is recorded while the robot grasps and shakes an object. The sound data are then divided into frames and transformed into 13-dimensional mel-frequency cepstral coefficients (MFCCs) as feature vectors. Finally, the feature vectors are vector-quantized using a codebook with 50 clusters, and a histogram is constructed.

3) Haptic Information: Haptic information is obtained from the three-finger robotic hand equipped with a tactile array sensor. A total of 162 time series of sensor values were obtained by grasping an object. Each time series was approximated using a sigmoid function, the parameters of which encode the object's tactile information [9]. Hence, a total of 162 feature vectors are obtained by grasping an object. Again, the bag-of-features model is applied to the data so that any variation resulting from changes in the grasping point can be absorbed. The feature vectors are vector-quantized using a codebook with 15 clusters, and the corresponding histogram is constructed.

4) Word Information: The user teaches object features to the robot through speech. The robot recognizes speech using continuous speech recognition and divides the recognized speech into words using morphological analysis. Finally, the word information is treated as the bag of words.

It should be noted that a dictionary of words is required for speech recognition and morphological analysis. In this study, it is assumed that the robot that has a vocabulary in advance, and we provide a framework that can understand meaning by connecting words with concepts. Therefore, the problem that we tackle in this study is not vocabulary acquisition but word grounding. However, it is possible to simultaneously acquire vocabulary as a phoneme sequence by applying unsupervised morphological analysis. In this study, the formation of a hierarchical conceptual structure is the main aim; therefore, we do not consider the simultaneous acquisition of vocabulary, but will do so in future.

IV. Hierarchical Category Clustering

Blei et al. have used the nested Chinese Restaurant Process (nCRP) as a prior distribution of the LDA. nCRP is an extension of the Chinese Restaurant Process (CRP), one of the Dirichlet processes. Here, we extend hLDA to hMLDA, which can form hierarchical concept structures by classifying robot-gathered multimodal information.

A. Chinese Restaurant Process [22]

CRP is the marginal distribution on partitions induced by the Dirichlet process that generates infinite-dimensional multinomial distributions by considering a Chinese restaurant with an infinite number of tables. When $n-1$ customers are already at K tables, the table z_n at which the n-th subsequent customer sits is drawn from the following distribution:

$$P(z_n = k|\gamma) = \begin{cases} \frac{N_k}{\gamma+n-1} & (k = 1, \cdots, K) \\ \frac{\gamma}{\gamma+n-1} & (k = K+1) \end{cases}, \qquad (1)$$

where N_k is the number of customers who sit at table k, and γ is a CRP parameter. An example of a Chinese restaurant is shown in Fig. 3. In this figure, ten customers are seated, and a new customer chooses a table according to the number of customers.

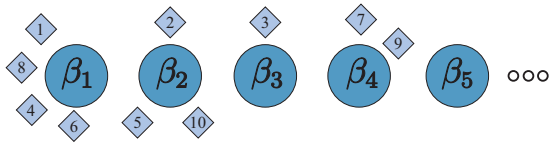

Fig. 3. Illustration of CRP.

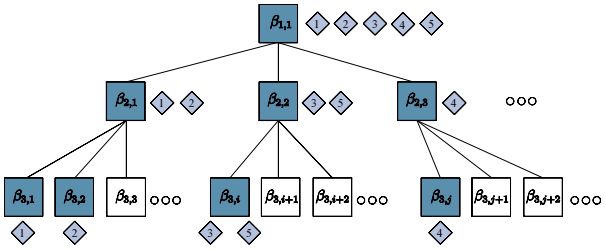

Fig. 4. Nested CRP

B. Nested CRP [11]

nCRP is an extension of CRP. nCRP can be defined by the following scenario. We suppose that there is an infinite number of Chinese restaurants, each of which has an infinite number of tables, in a city. One of these restaurants is the root restaurant, and there is a card that indicates a name of another restaurant on each table in each restaurant within the city. In addition, there is a card that refers to another restaurant on each table in the restaurant referred in the root restaurant, and this structure repeats infinitely. However, each restaurant is referred once. Thus, an infinitely-branched tree is organized.

A tourist arrives in the city, enters the root Chinese restaurant, and selects a table using Eq. (1) on the first evening. On the second evening, he goes to the restaurant referred to on the card placed on the table at which the tourist sat last night. The tourist repeats this process for L days. At the end of the trip, the tourist has sat at L restaurants, which constitute a path from the root to a restaurant at the L-th level in the infinite tree structure. Fig. 4 shows an example of the path in the case of five tourists and $L = 3$. In this figure, each box represents a Chinese restaurant, and each Chinese restaurant has a probability distribution with a parameter $\beta_{\ell,i}$, which generates the data.

C. Hierarchical Multimodal Latent Dirichlet Allocation

hLDA is a model that can classify documents hierarchically by introducing nCRP into the topic model. We extend

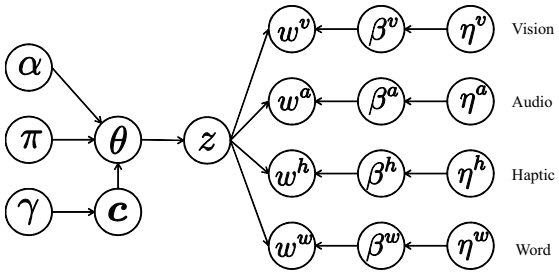

Fig. 5. Graphical model of hierarchical multimodal LDA.

hLDA to hMLDA to allow for the hierarchical clustering of multimodal information. The graphical model of hMLDA is shown in Fig. 5. In this figure, $\boldsymbol{c}$ is a path on the tree structure generated using γ-parameterized nCRP. In addition, z is an object category generated by the π- and α-parameterized stick-breaking process. w^v, w^a, w^h, and w^w denote visual, auditory, haptic, and word information and are generated from a multinomial distribution with parameter β^*. β^* is determined from the Dirichlet prior distribution using a parameter η^*. Object generation using hMLDA is given as follows:

1) For each modality($m \in \{v, a, h, w\}$), the multinomial distribution parameter β_k^m, which represents the probability of generating multimodal information in table ($k \in T$), is determined. (T represents a set of tables.)

$$\beta_k^m \sim \mathrm{Dirichlet}(\eta^m) \tag{2}$$

2) The following process is iterated for each object $d(\in \{1, 2, \cdots, D\})$.
   i) Path $\boldsymbol{c}_d$ in a tree structure is determined using nCRP.

$$\boldsymbol{c}_d \sim \mathrm{nCRP}(\gamma) \tag{3}$$

   ii) Parameter θ_d of multinomial distribution is generated using the stick-breaking process.

$$\theta_d \sim \mathrm{GEM}(\alpha, \pi) \tag{4}$$

   iii) The following is repeated for each feature n of modality m.
      a) Category $z_{d,n}^m$ of the n-th feature of modality m is determined.

$$z_{d,n}^m \sim \mathrm{Mult}(\theta_d) \tag{5}$$

      b) Feature $w_{d,n}^m$ is generated from the category $z_{d,n}^m$ on the path $\boldsymbol{c}_d$.

$$w_{d,n}^m \sim \mathrm{Mult}(\beta_{\boldsymbol{c}_d}[z_{d,n}^m]) \tag{6}$$

D. Hierarchical Category Clustering

Object categorization is equivalent to learning the model parameters shown in Fig. 5 using multimodal information. In this paper, Gibbs Sampling is used to learn the model parameters. The hMLDA parameters are estimated by sampling the category $z_{d,n}^m$ and path $\mathbf{c}_d$ from the posterior distribution.

1) Sampling Category: Given the current path assignments, a category $z_{d,n}^m$ of the n-th feature of modality m in object d is sampled from distribution as follows:

$$\begin{aligned} &p(z_{d,n}^m | \mathbf{z}_{-(d,n)}^m, \mathbf{c}, \mathbf{w}^m, \alpha, \pi, \eta^m) \propto \\ &\quad p(z_{d,n}^m | \mathbf{z}_{d,-n}^m, \alpha, \pi) p(w_{d,n}^m | \mathbf{z}, \mathbf{c}, \mathbf{w}_{-(d,n)}^m, \eta^m), \end{aligned} \tag{7}$$

where $\mathbf{c}$ and $\mathbf{w}^m$ denote a set of paths assigned to all the objects and a set of object feature of modality m, respectively. In addition, $\mathbf{z}_{-(d,n)}^m$ and $\mathbf{w}_{-(d,n)}^m$ denote the vectors of category allocations and observed features except for $z_{d,n}^m$ and $w_{d,n}^m$, respectively. Moreover, $\mathbf{z}_{d,-n}^m$ is the remainder except for the category $z_{d,n}^m$ assigned to the n-th feature from the set of categories $\mathbf{z}_d^m$ assigned to all features of modality m of object d. The first term of Eq. (7) denotes a multinomial distribution generated using the stick-breaking process, and represents probability that k is assigned to a category of n-th feature of modality m of d-th object.

$$\begin{aligned} &p(z_{d,n}^m = k | \mathbf{z}_{d,-n}^m, \alpha, \pi) \\ &= E\left[V_k \prod_{j=1}^{k-1} (1 - V_j) | \mathbf{z}_{d,-n}^m, \alpha, \pi \right] \\ &= E\left[V_k | \mathbf{z}_{d,-n}^m, \alpha, \pi \right] \prod_{j=1}^{k-1} E\left[1 - V_j | \mathbf{z}_{d,-n}^m, \alpha, \pi \right] \\ &= \frac{(1-\alpha)\pi + \#[\mathbf{z}_{d,-n}^m = k]}{\pi + \#[\mathbf{z}_{d,-n}^m \geq k]} \prod_{j=1}^{k-1} \frac{\alpha\pi + \#[\mathbf{z}_{d,-n}^m > j]}{\pi + \#[\mathbf{z}_{d,-n}^m \geq j]}, \end{aligned} \tag{8}$$

where $\#[.]$ counts the elements of an array satisfying a given condition. The second part of Eq. (7) is the probability that a feature quantity will be generated from path $\mathbf{c}_d$ and category $z_{d,n}^m$. The following formulas can be obtained under the assumption that the multinomial distribution parameter that generates the feature quantity is in itself generated from Dirichlet distribution with hyper-parameter η^m.

$$\begin{aligned} &p(w_{d,n}^m | \mathbf{z}, \mathbf{c}, \mathbf{w}_{-(d,n)}^m, \eta^m) \propto \\ &\quad \#[\mathbf{z}_{-(d,n)}^m = z_{d,n}^m, \mathbf{c}_{z_{d,n}}^m = c_{d,z_{d,n}^m}, \mathbf{w}_{-(d,n)}^m = w_{d,n}^m] + \eta^m \end{aligned} \tag{9}$$

This equation expresses the number of times that category $z_{d,n}^m$ was assigned to feature quantity $w_{d,n}^m$ on path $\mathbf{c}_d$.

2) Sampling Paths: Given the category allocation variables, path sampling is carried out as follows:

$$\begin{aligned} &p(\mathbf{c}_d | \mathbf{w}^v, \mathbf{w}^a, \mathbf{w}^h, \mathbf{w}^w, \mathbf{c}_{-d}, \mathbf{z}, \eta^v, \eta^a, \eta^h, \eta^w, \gamma) \\ &\quad \propto p(\mathbf{c}_d | \mathbf{c}_{-d}, \gamma) \\ &\qquad \times p(\mathbf{w}_d^v | \mathbf{c}, \mathbf{w}_{-d}^v, \mathbf{z}^v, \eta^v) p(\mathbf{w}_d^a | \mathbf{c}, \mathbf{w}_{-d}^a, \mathbf{z}^a, \eta^a) \\ &\qquad \times p(\mathbf{w}_d^h | \mathbf{c}, \mathbf{w}_{-d}^h, \mathbf{z}^h, \eta^h) p(\mathbf{w}_d^w | \mathbf{c}, \mathbf{w}_{-d}^w, \mathbf{z}^w, \eta^w), \end{aligned} \tag{10}$$

where $\boldsymbol{c}_{-d}$ denotes the remainder excluding $\boldsymbol{c}_d$ from $\boldsymbol{c}$. $p(\mathbf{w}_d^m | \mathbf{c}, \mathbf{w}_{-d}^m, \mathbf{z}, \eta^m)$ is the probability that the feature quantity of modality m will be generated from a specific path and

$p(\mathbf{c}_d|\mathbf{c}_{-d},\gamma)$ is the prior probability generated by nCRP. In each modality m, the probability that the feature quantity is generated as follows by marginalizing the parameters of the multinomial distribution.

$$p(\mathbf{w}_d^m|\mathbf{c},\mathbf{w}_{-d}^m,\mathbf{z}^m,\eta^m) = \prod_{\ell=1} \frac{\Gamma(\sum_w \#\left[\mathbf{z}_{-d}^m=\ell,\mathbf{c}_{-d,\ell}=c_{d,\ell},\mathbf{w}_{-d}^m=w\right]+V^m\eta^m)}{\prod_w \Gamma(\#[\mathbf{z}_{-d}^m=\ell,\mathbf{c}_{-d,\ell}=c_{d,\ell},\mathbf{w}_{-d}^m=w]+\eta^m)} \times \frac{\prod_w \Gamma(\#[\mathbf{z}^m=\ell,\mathbf{c}_\ell=c_{d,\ell},\mathbf{w}^m=w]+\eta^m)}{\Gamma(\sum_w \#[\mathbf{z}^m=\ell,\mathbf{c}_\ell=c_{d,\ell},\mathbf{w}^m=w]+V^m\eta^m)} \tag{11}$$

3) Learning by Gibbs Sampling: Given a random initial value $\mathbf{c}_1 \sim \mathbf{c}_D$ and $\mathbf{z}_1 \sim \mathbf{z}_D$, the following steps are iterated until convergence.

1) For each object $d \in \{1,\dots D\}$
   i) Sampling path

$$\mathbf{c}_d \sim p(\mathbf{c}_d|\mathbf{w}^v,\mathbf{w}^a,\mathbf{w}^h,\mathbf{w}^w,\mathbf{c}_{-d},\mathbf{z},\eta^v,\eta^a,\eta^h,\eta^w,\gamma) \tag{12}$$

   ii) Sampling a category for the n th feature quantity of modality m for each object

$$z_{d,n}^m \sim p(z_{d,n}^m|\mathbf{z}_{-(d,n)}^m,\mathbf{c},\mathbf{w}^m,\alpha,\pi,\eta^m) \tag{13}$$

Finally, the path and category of all objects converge to $\hat{\mathbf{c}}$ and $\hat{\mathbf{z}}$, respectively, by repeating this algorithm.

E. Unknown Object Recognition

Unknown object recognition becomes possible using the learned hMLDA. When the multimodal information $\mathbf{w}_{\bar{d}}^v$, $\mathbf{w}_{\bar{d}}^a$, $\mathbf{w}_{\bar{d}}^h$ and $\mathbf{w}_{\bar{d}}^w$ of novel object $\bar{d}$ are given, learned parameters $\hat{\mathbf{c}}$ and $\hat{\mathbf{z}}$ are fixed, and path sampling, as mentioned in the above algorithm, and category sampling are performed only for the novel object $\bar{d}$. However, the following equation is used instead of Eq. (11).

$$p(\mathbf{w}_{\bar{d}}^m|\hat{\mathbf{c}},\mathbf{w}^m,\hat{\mathbf{z}}^m,\mathbf{c}_{\bar{d}},\mathbf{z}_{\bar{d}}^m,\eta^m) = \prod_{\ell=1} \frac{\Gamma(\sum_w \#\left[\hat{\mathbf{z}}^m=\ell,\hat{\mathbf{c}}_\ell=c_{\bar{d},\ell},\mathbf{w}^m=w\right]+V^m\eta^m)}{\prod_w \Gamma(\#[\hat{\mathbf{z}}^m=\ell,\hat{\mathbf{c}}_\ell=c_{\bar{d},\ell},\mathbf{w}^m=w]+\eta^m)} \times \frac{\prod_w \Gamma(\#[\hat{\mathbf{z}}^m=\ell,\hat{\mathbf{c}}_\ell=c_{\bar{d},\ell},\mathbf{w}^m=w] + \#[\mathbf{z}_{\bar{d}}^m=\ell,\mathbf{w}_{\bar{d}}^m=w]+\eta^m)}{\Gamma(\sum_w \#[\hat{\mathbf{z}}^m=\ell,\hat{\mathbf{c}}_\ell=c_{\bar{d},\ell},\mathbf{w}^m=w] + \#[\mathbf{z}_{\bar{d}}^m=\ell,\mathbf{w}_{\bar{d}}^m=w]+V^m\eta^m)}, \tag{14}$$

where $\hat{\mathbf{z}}^m$ is a set of categories that assigned to the feature quantity of modality m at the time of learning.

F. Predicting Unobserved Information

The validity of hMLDA lies in the prediction of unobserved information. That is, by looking at an object, the robot can predict its hardness or whether the object emits any sound. Conversely, if a word is given, the robot can understand its meaning by predicting multimodal information from observed information. Here, we consider predicting word information $\mathbf{w}_{\bar{d}}^w$ from visual information $\mathbf{w}_{\bar{d}}^v$ of an unknown object $\bar{d}$. First, object category recognition is performed from the given information, as described in the previous section. However, the following equation is used instead of Eq. (10).

$$p(\mathbf{c}_{\bar{d}}|\mathbf{w}^v,\hat{\mathbf{c}},\hat{\mathbf{z}},\mathbf{w}_{\bar{d}}^v,\mathbf{z}_{\bar{d}}^v,\eta^v,\gamma) \propto p(\mathbf{c}_{\bar{d}}|\hat{\mathbf{c}},\gamma)p(\mathbf{w}_{\bar{d}}^v|\hat{\mathbf{c}},\mathbf{w}^v,\hat{\mathbf{z}}^v,\mathbf{c}_{\bar{d}},\mathbf{z}_{\bar{d}}^v,\eta^v) \tag{15}$$

In addition, category sampling is performed only for visual information. Through repetition of the above-described procedure until samplings converge, an object category can be determined using only a part of the information. For a category that has been estimated, we can determine the probability of a word's occurrence using the following equation.

$$p(w_{\bar{d}}^w|\hat{\mathbf{z}},\hat{\mathbf{c}},\mathbf{w}^w,\mathbf{w}^v,\mathbf{c}_{\bar{d}},\mathbf{w}_{\bar{d}}^v,\alpha,\pi,\eta^w,\eta^v) = \sum_{z_{\bar{d}}} p(w_{\bar{d}}^w|z_{\bar{d}},\hat{\mathbf{z}}^w,\hat{\mathbf{c}},\mathbf{w}^w,\eta^w)p(z_{\bar{d}}|\hat{\mathbf{z}}^v,\hat{\mathbf{c}},\mathbf{w}^v,\mathbf{c}_{\bar{d}},\mathbf{w}_{\bar{d}}^v,\alpha,\pi,\eta^v) \tag{16}$$

This refers solely to word prediction; other predictions can possibly be made following a similar method.

TABLE I
EXAMPLES OF TEACHER'S UTTERANCES (THE EXPERIMENT WAS CARRIED OUT IN JAPANESE.)

This is a Bear stuffed animal.	It is white and light.
This is a stuffed animal with tinkling sounds.	This thing is soft.
A blue spray can.	It is likely to sound.
This is a drink.	This is a plastic bottle.
This drink is tea.	This is food.

Fig. 6. Sixty-seven objects used in experiment. (objects in the rectangle are used for recognition experiments in V-B and V-C as unseen objects.)

V. EXPERIMENTS

Experiments were carried out using information such as visual, audio, haptic, and word information acquired by the robot, as shown in Fig. 2. As word information, five subjects taught characteristics of an object to the robot using the speech recognition. Examples of the teacher's utterances are shown in TABLE I. In addition, the 67 objects shown in Fig. 6 were used, we set the number of layers to 4 and performed clustering, recognition, and prediction using hMLDA. Hyper-parameters γ, α and π of the model were set to the values given in the literature [11], and η^* wrere determined empirically.

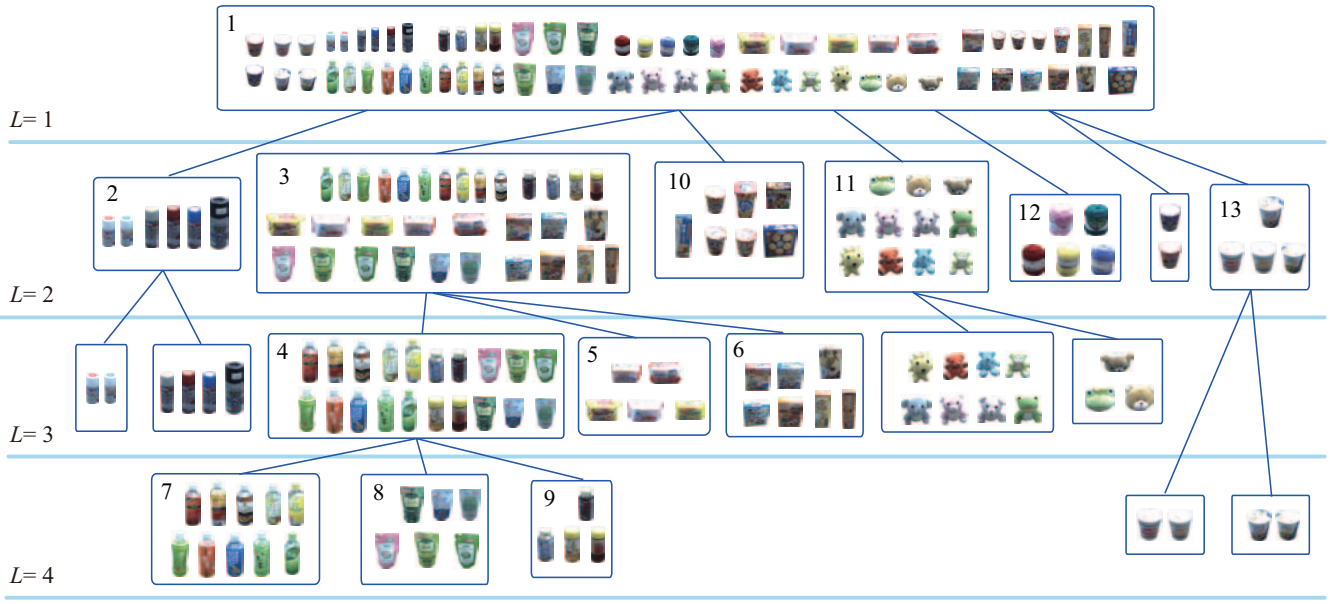

Fig. 7. Categorization result using hMLDA.

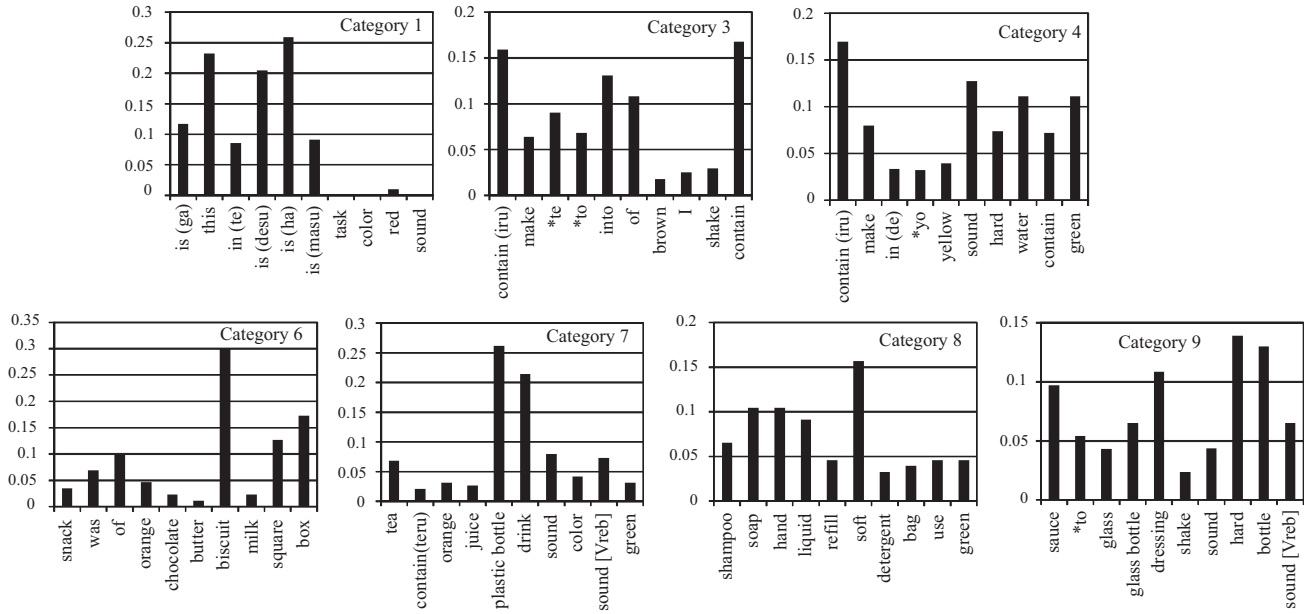

Fig. 8. Inference results of words in each category: * represents a function word in Japanese, while words inside parentheses denote different word in Japanese that have the same meaning in English.

A. *Hierarchical Categorization*

First, hMLDA was used for hierarchical categorization of 67 objects, the result of which is shown in Fig. 7. In the layers corresponding to $L = 2$ in the Fig. 7, Category 2 was composed of only the spray can. Below it, in the layers corresponding to $L = 3$, individual categories were formed based on spray can size. Similarly, Category 13 was composed of only the instant noodle. Below it, in the layers corresponding to $L = 3$, individual categories were formed based on instant noodle type. In addition, in the hierarchy of $L = 2$, category 3 consisted of plastic bottles, glass bottles, shampoo, flooring wiper, and cookies, and in the hierarchy below it, flooring wiper and cookies formed individual categories. In addition, in the hierarchy of $L = 4$, plastic bottles and shampoo were classified correctly. In the hierarchy of $L = 2$, category 11 comprised rattles and plushies. In the hierarchy of $L = 3$, rattles and, plushies were classified correctly. As mentioned above, some incorrect categories, such as category 10, which comprises cookies and snacks, exist. However, using hMLDA, the robot was able to pick up clues regarding similarities in terms of visual, audio, haptic, and word information, and form both extensive and

concrete categories automatically.

Furthermore, the top ten words with high probability were extracted from each category, and the probabilities of those words are shown in Fig. 8. Because categories 1 and 3 consist of two or more types of objects, words that occur in any categories such as "this" and "is" have high probability in these categories. Furthermore, category 4 has a high probability of occurrence of words "water" and "contain", which indicate the liquid contains, and of a word "sound", which indicates sounder. Finally, from categories 6, 7, 8, and 9, which form individual object categories, the words "biscuit," "bottles," "shampoo," and "dressing", which express each category, are generated with high probability. Thus, it is correctly connected with the word that expresses each category in lower layer.

B. Unknown Object Recognition

Next, 67 objects were divided into objects for recognition and learning, hMLDA was learned by the objects for learning, and the objects for recognition were recognized as unknown objects. In Fig. 6, the objects in the rectangle are used for recognition. The result of recognition is shown in Fig. 9, and objects surrounded by a rectangular were the recognized objects. From this figure, a plushie was misclassified as yarns in the right category of $L = 4$; however, the other objects were classified correctly.

C. Word Prediction

Then, using the visual, auditory, and tactile information of the recognition objects, word information is predicted by hMLDA. Figs. 10 (a)–(d) show the results of word prediction, and represent the probabilities of the occurrence of the top five words with high probability. The same prediction was performed using MLDA for comparison (Fig. 10 (e)–(h)). However, because it was necessary to define the number of categories beforehand, we set the number of categories to 11. Fig. 10 shows that hMLDA can accurately predict a word that expresses an object category name and its features. However, using MLDA, words contained in every object such as "is" and "this" are predicted, but the word representing a category is not predicted. Because MLDA does not consider hierarchy, it predicts the words that occurred frequently during learning. However, considering the hierarchy in hMLDA, we were able to circumvent this problem.

VI. Conclusion

In this paper, we have proposed hMLDA that can form hierarchical concept structure. Concepts were formed by classifying the visual, audio, tactile, and word information acquired by a robot. hMLDA is extended from hLDA, which was a method proposed for document clustering, to classify multimodal information, and, therefore, the robot can form the categories of various granularity, which are from concrete categories to extensive categories. Although in a limited situation, it was experimentally shown that such a layered structure can actually be formed. Furthermore, the connection between a concept and a word is obtained. In the nonhierarchical conceptual structure formed using MLDA, functional words "this" and "is" are connected with many concepts. Therefore, when predicting a word from sensory information, a heuristic method that removes such words is required. By contrast, in hMLDA, the higher layer of a hierarchical structure absorbs such functional words. Furthermore, because the proposed hMLDA is based on the Bayesian nonparametric method, it does not required the number of categories in advance.

We are planning to apply online learning to hMLDA for future research. Moreover, we believe that it is necessary to conduct a large-scale experiment with a considerably greater number of objects. Furthermore, it is necessary to consider concept formation for adjectives. For example, the concept of a red object will be formed and it will be connected with the adjective "red." However, in order to form a category that essentially means red, it is necessary to perform category clustering considering only color modality. Such clusterings can be realized by model selection introducing weight to modality [24]. This idea will be applied to hMLDA, and we will realize the formation of various concepts having hierarchical structures.

References

[1] F.G. Ashby, and W.T. Maddox, "Human category learning," Annual Review of Psychology, vol.56, pp.149–178, 2005.

[2] E. Rosch, "Principles of categorization," Concepts: core readings, pp.189–206, 1999.

[3] T. Nakamura, T. Nagai, and N. Iwahashi, "Bag of Multimodal LDA Models for Concept Formation," in Proc. IEEE Int. Conf. on Robotics and Automation, pp.6233–6238 2011.

[4] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent dirichlet allocation," Journal of Machine Learning Research, vol.3, pp.993–1022, 2003.

[5] T. Nakamura, T. Nagai, and N. Iwahashi, "Multimodal Object Categorization by a Robot," in Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp.2415–2420 2007.

[6] T. Nakamura, T. Nagai, and N. Iwahashi, "Grounding of word meanings in multimodal concepts using LDA," in Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp.3943–3948, 2009.

[7] T. Nakamura, T. Nagai, and N. Iwahashi, "Multimodal categorization by hierarchical Dirichlet process," in Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp.1520–1525, 2011.

[8] T. Araki, T. Nakamura, T. Nagai, K. Funakoshi, M. Nakano, N. Iwahashi, "Online object categorization using multimodal information autonomously acquired by a mobile robot," Advanced Robotics, Vol.26, Issue 17, pp.1995–2020, 2012.

[9] T. Araki, T. Nakamura, T. Nagai, K. Funakoshi, M. Nakano, N. Iwahashi, "Autonomous acquisition of multimodal information for online object concept formation by robots," in Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp.1540–1547, 2011.

[10] T. Araki, T. Nakamura, T. Nagai, S. Nagasaka, T. Taniguchi, and N. Iwahashi, "Online learning of concepts and words using multimodal LDA and hierarchical Pitman-Yor language model," in Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp.1623–1630, 2012.

[11] D. Blei, T. Griffiths, and M. Jordan, "The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies," Journal of the ACM, vol.57, no.2, article 7, 2010.

[12] J. Sivic, B.C. Russell, A.A. Efros, A. Zisserman, and W.T. Freeman, "Discovering object categories in image collections," in Proc. IEEE Conf. on Computer Vision, pp.17–20, 2005.

[13] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition, vol.2, pp.264–271, June 2003.

[14] L. Fei-Fei, "A bayesian hierarchical model for learning natural scene categories," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp.524–531, 2005.

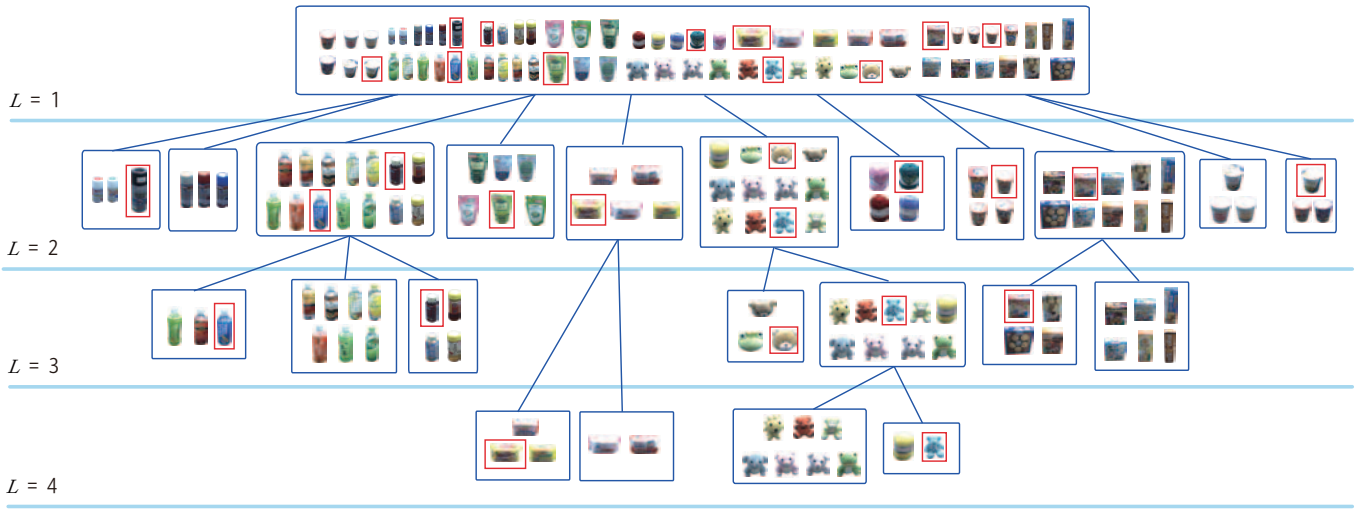

Fig. 9. Unseen object recognition using hMLDA.

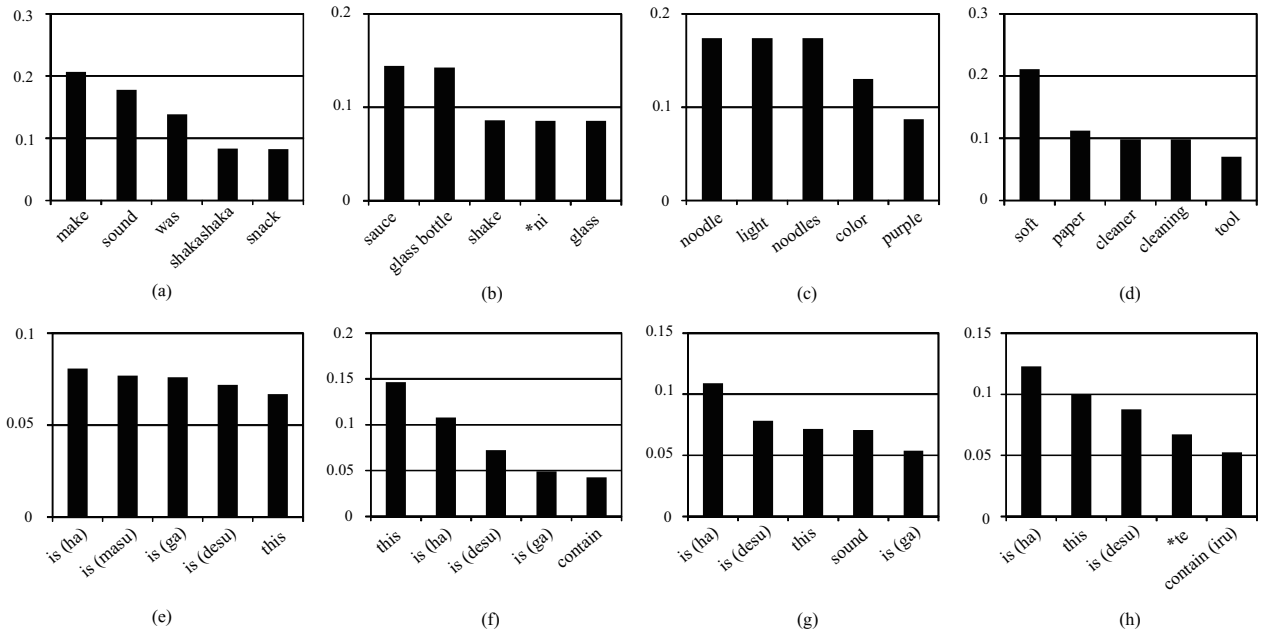

Fig. 10. Inference result of words using hMLDA and MLDA: (a)Inference from chips using hMLDA, (b)Inference from glass bottle using hMLDA, (c)Inference from noodles using hMLDA, (d)Inference from flooring wiper using hMLDA, (e)Inference from chips using MLDA, (f)Inference from glass bottle using MLDA, (g)Inference from noodles using MLDA and, (h)Inference from flooring wiper using MLDA. * represents a function word in Japanese, while the word inside parentheses denotes different words in Japanese that have the same meaning in English.

[15] C. Wang, D. Blei, and L. Fei-Fei, "Simultaneous image classification and annotation," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition, vol.0, pp.1903–1910, 2009.

[16] F. Endres, C. Plagemann, C. Stachniss, and W. Burgard, "Unsupervised discovery of object classes from range data using latent dirichlet allocation," Robotics: Science and Systems, 2009.

[17] M. Marszałek, and C. Schmid, "Constructing category hierarchies for visual recognition," Computer Vision–ECCV 2008, pp.479–491, 2008.

[18] L. Li, C. Wang, Y. Lim, D. Blei, and L.Fei-Fei, "Building and using a semantivisual image hierarchy," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp.3336–3343, 2010.

[19] Z. Jianhua, Z. Jianwei, S. Chen, H. Ying, and G. Haojun, "Constructing dynamic category hierarchies for novel visual category discovery," in Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp.2122–2127, 2012.

[20] J. Sivic, B. C. Russell, A. Zisserman, W. T. Freeman, and A. A. Efros, "Unsupervised Discovery of Visual Object Class Hierarchies," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp.1–8, 2008.

[21] A. Vedaldi, and B. Fulkerson, "Vlfeat: An open and portable library of computer vision algorithms," ACM Int. Conf. on Multimedia, pp.1469–1472, 2010.

[22] D. Aldous, "Exchangeability and related topics," École d'Été de Probabilités de Saint-Flour XIII-1983, pp.1–198, 1985.

[23] T. Ogata, S. Nishide, H. Kozima, K. Komatani, and H. Okuno, "Inter-modality mapping in robot with recurrent neural network," Pattern Recognition Letters, Vol.31, No.12, pp.1560–1569, 2010.

[24] T. Nakamura, T. Nagai. N. Iwahashi, "Bag of Multimodal Hierarchical Dirichlet Processes: Model of Complex Conceptual Structure for Intelligent Robots," in Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, 3818–3823, 2012