Abstract—The formation of categories, which constitutes the basis of developing concepts, requires multimodal information with a complex structure. We propose a model called the bag of multimodal hierarchical Dirichlet processes (BoMHDP), which enables robots to form a variety of multimodal categories. The BoMHDP model is a collection of a large number of MHDP models, each of which has a different set of weights for sensory information. The weights work to realize selective attention and enable the formation of various types of categories (e.g., object, haptic, and color). The BoMHDP model is an extension of the HDP, and categorization is unsupervised. However, categories that are not natural for humans are also formed. Therefore, only the significant categories are selected through interaction between the user and the robot. At the same time, words obtained during the interaction are connected to the categories. Finally, categories, which are represented by words, are selected. The BoMHDP model was implemented on a robot platform and a preliminary experiment was conducted to validate it. The results revealed that various categories can be formed with the BoMHDP model. We also analyzed the formed conceptual structure by using multidimensional scaling. The results indicate that the complex conceptual structure was represented reasonably well with the BoMHDP model.

I. INTRODUCTION

Categorization of objects plays an important role in human intelligence. Human understanding is based on the inference of unobservable information through categories, which can be seen as constituting the basis of concepts. Therefore, the ability to categorize is also important for intelligent autonomous robots.

Against this background, we proposed a multimodal categorization method [1] [2], which is an extension of probabilistic latent semantic analysis (pLSA) [3] and latent Dirichlet allocation (LDA) [4]. We also proposed a multimodal categorization method [5] based on the hierarchical Dirichlet process (HDP) [6] to solve the problem of pLSA and LDA requiring the number of categories in advance. In these reports, we showed that multimodal information can be used to form categories that seem natural to humans. We also showed that it is possible to recognize the category of an unknown object and to infer its unobservable properties by using the learnt model. For example, the proposed model enables a robot to stochastically infer how hard an object is or whether it produces a sound only from its visual characteristics. This is an inference of the object’s function through its category—something that comes naturally to humans. We have implemented such an ability on a robot and thus showed the validity of the model.

Tomoaki Nakamura and Takayuki Nagai are with the Department of Electronic Engineering, The University of Electro-Communications, 1-5-1 Chofugaoka Chofu-shi, Tokyo 182-8585, Japan naka@apple.ee.uue.ac.jp

Naoto Iwahashi is with the NICT Knowledge Creating Communication Research Center, 2-2-2 Hikaridai, Seika-cho Souraku-gun, Kyoto 619-0288, japan naoto.iwahashi@atr.jp

These previous studies aimed at constructing categories based on specific object types, for example, plush toys and musical instruments. However, humans use not only object categories but also other types of categories, such as color, shape, and haptic characteristics. Rather than a hierarchical structure, these categories, which can be thought of as concepts, have very complex structures. To form such categories, we extended the MLDA to the bag of multimodal LDA (BoMLDA) [7], in which weights for multimodal information are introduced. This model consists of a large number of MLDA models, each of which has a different set of weights for multimodal information. However, since the BoMLDA model is based on LDA, it also suffers the serious problem of requiring the number of categories in advance.

To overcome this problem, we propose here the multimodal HDP (MHDP) model, which automatically determines the number of categories, as shown in Fig. 1. First, multimodal information that a robot acquires is classified using many MHDP models with a variety of weights for multimodal perceptual information. We call this collection of MHDP models the bag of MHDP (BoMHDP) model. Although this collection of models represents various types of categories that correspond to complex conceptual structures, many are irrelevant to both humans and robots. This is because all possible combinations of weights are applied.
to the model and significant models should be selected. This is a model selection problem and can be solved through a dialogue between the user and the robot. Finally, the robot can obtain the categories, which are formed on the basis of multimodal information, as well as the words that represent the categories and the degree of connection between the categories and modalities.

The proposed BoMHDP model is an extension of the HDP model and enables the robot to infer the unobservable properties of unseen objects. Therefore, the robot can infer haptic or audio information of an unseen object only from its visual information. The robot can also recall sensory information from words because words are connected to the categories through the dialogue. Moreover, the robot can obtain the connection between a modality and a word. Accordingly, it can focus on a particular modality.

We attempted visualization of the BoMHDP to grasp the conceptual structure formed by the robot. Multidimensional scaling (MDS) was used for this purpose. A very complex model structure can be represented in a low-dimensional space based on the distances among MHDP models.

Many studies on the categorization of only visual information have been conducted [8]–[11]. However, many categories are formed not only by visual information but also by many types of perceptual information. As a different method of sensing, several studies have explored categorization of the objects on the basis of sounds generated when a robot touches them [12], [13]. These studies showed that sound can be used to form categories when the same is difficult with only visual information. However, one purpose of this research is to form and recognize object categories, and inferring unobserved information is not considered. In addition, the present research is related to previous work on language acquisition [14]–[16]. However, these studies did not use haptic information and did not take into account inference among modalities. Studies have also been conducted on stochastic model selection [17]–[19]. In these studies, one model was selected according to stochastic criteria. The models proposed in these studies can also be used instead of the HDP model. For our research, we used an extension of the HDP model because it is simple and easy to implement. Furthermore, HDP was used by Canini et al. to model categorization by humans. They showed that HDP can model a human’s transfer learning [20] and form hierarchical categories that are similar to the human’s categories. This is another reason why we use the extension of HDP for multimodal categorization.

II. MULTIMODAL INFORMATION

In this section, we explain multimodal information (visual, audio, and haptic) obtained from a robot (Fig. 2) observing objects.

Visual Information A stereo camera is mounted on the head of the robot (Fig. 2), and images captured with it are used as visual information. We used 50 images in the experiment. Principal component analysis scale-invariant feature transform (PCA SIFT) [21] features are then extracted from these images. PCA SIFT is a local feature that uses the principal component of image patches as the descriptor instead of the SIFT descriptor [22]. It is believed that PCA SIFT is better than SIFT as a local image feature. We obtained many PCA SIFT descriptors from one image. It is difficult to use these descriptors for describing an image because their number differs in each image. To solve this problem, we applied the bag of keypoints approach. Each feature is vector quantized by 500 representative vectors and used as a 500-dimensional histogram.

Furthermore, we use the hue, saturation, value (HSV) color histogram as another visual feature. There are ten bins of both hue and saturation. Therefore, the color histogram is a 100-dimensional vector.

Audio information We use sound captured when the robot shakes an object as audio information. The sound is divided into frames, and each frame is transformed to a feature vector. We use a 13-dimensional Mel-frequency cepstrum coefficient (MFCC) as the feature vector. Finally, each feature vector is vector quantized by 50 representative vectors as well as image information, and the sound is transformed into a 50-dimensional histogram.

Haptic information We use a tactile array sensor composed of 162 sensors mounted on the robot’s hand. Time-series sensor values are captured while the robot grasps an object, and haptic information is computed from these data. First, we approximate the sensor values, and use the parameters of the approximation as the feature vector of each sensor [7]. Then, each feature vector is vector quantized. Finally, a 15-dimensional histogram is used as the haptic information.

III. BAG OF MULTIMODAL HDP: BoMHDP

As shown in the graphical representation of the model in Fig. 3, categories (concepts), denoted by $k$, are generated from the multinomial distribution parameterized by $\xi$. Then, multimodal information $x^v$, $x^c$, $x^a$, and $x^h$ and words
$x^w$ are generated independently from each category $k$. $\theta^*$ denotes parameters of the multinomial distribution, which generate multimodal information and words. The learning process of this graphical model is divided into category formation (the part inside the solid-lined rectangle in Fig. 3) and word grounding (the part inside the dashed-lined rectangle in Fig. 3). First, multimodal information is classified into various categories by using the BoMHDP model. The classification corresponds to learning the parameters $\theta^*$ and $\xi$, which are estimated using the BoMHDP model. In this model, $\theta^*$ denotes the parameter of distribution that generates $x^w$, and $\xi$ denotes the parameter of distribution that generates the category $k$. In the HDP model, $\xi$ is decided by the Dirichlet process.

As shown in the detailed graphical model in Fig. 4, the BoMHDP model is an extension of the MHDP model [5]. In this figure, $x^s_{jn}, x^c_{jn}, x^a_{jn},$ and $x^h_{jn}$ denote the $n$-th feature of the $j$-th object’s SIFT, color, audio, and haptic information, respectively. The modality $m \in \{s, c, a, h\}$ information $x^m_{jn}$ is generated from multinomial distribution parameterized by $\theta^m_k$. The multinomial distribution is generated from the Dirichlet prior distribution, whose parameter is $\alpha_0^m$. Furthermore, $w^m_k$ denotes the weights for modality $m$, and $\phi$ denotes the set of weights. Many types of categories (object, color, and haptic) can be formed using these weights. We can explain the generation process for the MHDP model by using the Chinese restaurant franchise [6] as well as the HDP model, considering an object as a Chinese restaurant, and modality information as a customer.

A. Chinese Restaurant Franchise

Considering that a modality $m$ is a type of customer (child, man, woman, etc.), and a dish $k$, which includes foods for each type of customer, is served at a table. Each customer $x^m_{jn}$ eats the food for his/her type at their table $x^m_{jt}$. We can divide this process into two Chinese restaurant processes (CRPs) [23]. One is the table selection process and the other is the dish selection process.

**Table selection process:** With the CRP and the weights for each modality $m$, the expression for the prior probability that a customer $x^m_{jn}$ of the $j$-th restaurant sits at a table $t$ is as follows:

$$
P(t^m_{jn} = t | \lambda, \phi) = \begin{cases} \frac{\sum_m w^m K^m_{jn}}{\gamma + \sum_m w^m K^m_{jn}} & (t = 1, \cdots, T_j) \\ \frac{\sum_m w^m T^m_{jn}}{\gamma + \sum_m w^m T^m_{jn}} & (t = T_j + 1) \end{cases},$$

where $T_j$ denotes the number of tables in the $j$-th restaurant and $N^m_{jn}$ denotes the number of customers of the $j$-th restaurant who sit at table $t$, which is also considered a measure of the popularity of that table. A customer is more likely to sit at a popular table. In addition, the probability that customer $x^m_{jn}$ prefers dish $k$ is as follows:

$$
P(x^m_{jn} | X^m_k, \phi) = \frac{w^m_k N^m_{kj} + \alpha^m_0}{w^m K^m_{kj} + d^m \alpha^m_0},$$

where $N^m_{kj}$ and $N^m_{kj}$ denote the number of customers who eat dish $k$ and the number of the customers who eat dish $k$ among all the customers $x^m_{jn}$. Further, $d^m$ denotes the dimension of modality $m$. $X^m_k$ denotes a set of $m$-type customers who eat dish $k$. Therefore, the posterior probability that customer $x^m_{jn}$ sits at table $t$ by using Bayes’ theorem is as follows:

$$
P(t^m_{jn} = t | X, \lambda, \phi) \propto P(x^m_{jn} | X_{k=kj}, \phi) P(t^m_{jn} | \lambda, \phi)$$

$$
\propto \begin{cases} P(x^m_{jn} | X^m_{k=kj}, \phi) \frac{\sum_m w^m N^m_{jn}}{\gamma + \sum_m w^m N^m_{jn}} & (t = 1, \cdots, T_j) \\ P(x^m_{jn} | X^m_{k=kj}, \phi) \frac{\lambda}{\gamma + \sum_m w^m T^m_{jn}} & (t = T_j + 1) \end{cases},$$

where $X^m$ denotes the set of $m$-type customers in all restaurants. Each customer selects a table to sit according to this probability.

**Dish selection process:** The prior probability that dish $k$ is served on table $t$ by using CRP is as follows:

$$
P(k_{jt} = k | \gamma, \phi) = \begin{cases} \frac{M_k}{\gamma + M_k} & (k = 1, \cdots, K) \\ \frac{\gamma}{\gamma + M_k} & (k = K + 1) \end{cases}.$$
follows:
\[
P(k_{jt} = k | \mathbf{X}, \gamma, \phi) \\
\propto P(\mathbf{X}_{jt} | \mathbf{X}_k, \phi) P(k_{jt} = k | \gamma, \phi) \\
= \begin{cases} 
P(\mathbf{X}_{jt} | \mathbf{X}_k) M_{jt}^{\gamma + M - 1} (k = 1, \ldots, K) \\ 
P(\mathbf{X}_{jt} | \mathbf{X}_k) M_{jt}^{\gamma + M - 1} (k = K + 1) \end{cases} 
\]  
(5)

Each dish is selected according to this probability. To select dish \( k = K + 1 \) means generating a new dish.

B. Multimodal Categorization

Categorization of objects is achieved by table assignment and dish assignment using Gibbs sampling. In Gibbs sampling, table \( t_{jn}^m \) is sampled from the following probability given \( X^{-mjn} \), which represents all customers except customer \( x_{jn}^m \).

\[ t_{jn}^m \sim P(t_{jn}^m | X^{-mjn}, \lambda) \]  
(6)

Dish \( k_{jt} \), which is served on table \( t \), is sampled from the following probability given \( X^{-jt} \), which represents all customers except customers \( X_{jt} \) who sit at table \( t \).

\[ k_{jt} \sim P(k_{jt} | X^{-jt}, \gamma) \]  
(7)

The parameters can be estimated through repeated sampling using Eq. (6) for all customers and all the restaurants and by sampling using Eq. (7) for all restaurants and all tables in each restaurant. Finally, the probability that the \( j \)-th object is classified into category \( k \) is written as follows:

\[ P(k | \mathbf{X}_j, \phi) = \frac{\sum_t T_j \delta(k_{jt}) \sum_m w^m N_{jt}^m}{\sum_m w^m N_{jt}^m} \]  
(8)

where \( \mathbf{X}_j \) denotes all features of the \( j \)-th object; \( N_{jt}^m \) and \( k_{jt} \) are converged values of \( N_{jt}^m \) and \( k_{jt} \), respectively, by the above sampling iteration; and \( \delta_a(b) \) is the delta function, which is 1 if \( a = b \) but 0 otherwise.

Various categories are formed by varying parameter \( \phi \) of the BoMHDP model. Therefore, categories that are connected to multiple modalities and a particular modality are formed.

IV. WORD GROUNDING AND CATEGORY SELECTION

As mentioned in the previous section, the part inside the solid-lined rectangle in Fig. 3 was learnt. Therefore, the robot can then form various categories unsupervised by using the BoMHDP model. Next, words were grounded on the formed categories on the basis of the learning dependency between the words and categories (the part inside the dashed-lined rectangle in Fig. 3). Further, a category represented by each word was selected.

A. Word Grounding

Word grounding is achieved by estimating parameter \( \theta^w \) of the multinomial distribution from which words are generated. Word information is treated as a bag-of-words model. Therefore, words are modeled by occurrence frequency regardless of their position. First, a human teaches a robot words that represent an object by showing the object to the robot. The robot can estimate the object’s category by using the learnt models. Therefore, we can directly calculate the probability that words occur from each category in the model \( \phi \) from the occurrence frequency of words and estimated categories as follows:

\[ P(x^w | k, \phi) \propto \sum_j P(k | \mathbf{x}_j^w, \mathbf{x}_j^a, \mathbf{x}_j^b, \phi) n(x^w, j), \]  
(9)

where \( \mathbf{x}_j^w \) denotes the multimodal information of the \( j \)-th object and \( n(x^w, j) \) denotes the occurrence frequency of word \( x^w \), which is assigned to the \( j \)-th object. The probability \( P(k | \mathbf{x}_j^w, \mathbf{x}_j^a, \mathbf{x}_j^b, \phi) \) is a result of category estimation from the learnt model, whose parameter is \( \phi \).

B. Category Selection

A vast number of categories are formed using the BoMHDP model, which is composed of MHDP models that have different \( \phi \). Then, the words are grounded on the categories; that is, categories represented by each word are selected. We introduce mutual information as the degree of connection between a category and a word, and mutual information between word \( x^w \) and category \( k \) in model \( \phi \) is written as follows:

\[ I(x^w, k | \phi) = P(w^w, k | \phi) \log \frac{P(x^w, k | \phi)}{P(x^w | \phi) P(k | \phi)}. \]  
(10)

The mutual information is the amount of information that is shared with two stochastic variables and is considered as a degree of mutual dependence. Hence, if there is a large amount of mutual information between a word and a category, the word is considered to represent the category. Finally, a category \( k_{x^w} \) and a model \( \phi_{x^w} \) represented by the word \( x^w \) are selected according to the following equation:

\[ (k_{x^w}, \phi_{x^w}) = \arg \max_{k,\phi} I(x^w, k | \phi) \]  
(11)

V. EXPERIMENTS

Two experiments were conducted with 39 objects (e.g., plush toys, cups) to examine the validity of the proposed BoMHDP.

A. Multimodal Categorization

First, the BoMHDP model was learnt by varying the weights for each type of information \( w^a \). The weights used in the experiment were 0, 1, 2, 5, 10, 20, and 30, which were empirically chosen to form the various categories. Therefore, the BoMHDP model consisted of 2400 (= \( 7^4 - 1 \)) MHDP models, except the models for which all weights were 0. Then, 20 of the 39 objects were chosen randomly, and a user assigned words to each of the 20 objects. Finally, the categories that have the maximum amount of mutual information with the words were selected. Fig. 6 shows the results of category selection. The object categories, which represent “plush toy,” “maraca,” and so on, were correctly formed. Moreover, the category “instrument” was considered to have been correctly formed because it consists of only those objects that produce sound. Furthermore, the categories connected to a particular modality, such as color and haptic, were reasonably formed.
Fig. 5. 2D plots of BoMHDP structure represented using MDS. The horizontal and vertical axes respectively represent the first and second significant dimensions extracted by MDS. Each point represents an MHDP model and its intensity reflects the following values. (a) Weight for visual information (SIFT), (b) weight for visual information (color), (c) weight for audio information, (d) weight for haptic information, (e) mutual information between words “plush toy” and models, (f) mutual information between word “yellow” and models, (g) mutual information between word “maraca” and models, and (h) mutual information between word “soft” and models.

Fig. 6. Examples of formed categories.

B. Visualization of Relationship among Models

Next, we plotted the models in 2D space by using MDS to visualize the relationships among the MHDP models. Multidimensional scaling is a method of the multivariate analysis technique and enables high dimensional information to be visualized in low dimensional space only from distances between the MHDP models. However, we cannot simply calculate distances between two models because each MHDP model has a different model structure. Therefore, we use the Kullback-Leibler distance between the probabilities

\[ P(x^w|\mathbf{x}^s_i, \mathbf{x}^c_i, \mathbf{x}^a_i, \mathbf{x}^t_i, \phi) \]

in which words \( x^w \) are generated from the multimodal information \( x^a \) of the \( j \)-th object, as the distance between models. Therefore, the distance between models \( \phi_1 \) and \( \phi_2 \) is defined as follows:

\[ D(\phi_1, \phi_2) = \frac{D_{KL}(\phi_1|\phi_2) + D_{KL}(\phi_2|\phi_1)}{2}, \]  

where

\[ D_{KL}(\phi_1|\phi_2) = \sum_j \sum_{x^w} P(x^w|\mathbf{x}^s_j, \mathbf{x}^c_j, \mathbf{x}^a_j, \mathbf{x}^t_j, \phi_1) \times \log \frac{P(x^w|\mathbf{x}^s_j, \mathbf{x}^c_j, \mathbf{x}^a_j, \mathbf{x}^t_j, \phi_1)}{P(x^w|\mathbf{x}^s_j, \mathbf{x}^c_j, \mathbf{x}^a_j, \mathbf{x}^t_j, \phi_2)}. \]  

Fig. 5(a)-(d) show the results of 2D plotting of each model, and the intensity of each point reflects the weights of SIFT, color, audio, and haptic information. We can see that models strongly connected to SIFT are distributed in the center, those strongly connected to color information are distributed at the bottom left, those strongly connected to audio information are distributed at the bottom right, and those strongly connected to haptic information are distributed at the top in the 2D space. Fig. 5(e) also shows the plots of the models in 2D space, and the intensity of each point represents the mutual information between each model and the words “Plush toy.” From this figure, models that have a larger amount of mutual information with “Plush toy” are distributed at the top right in the 2D space. This means that the category “Plush toy” consists of SIFT, audio, and haptic information. In the case of the word “Yellow,” models connected to it are distributed at the bottom left in the 2D space. This means that models connected to color information are selected as models that represent “Plush toy.” Similarly, models connected to audio information are selected as models that represent “Maraca” because a maraca produces sound, and models connected
to haptic information are selected as models that represent “Soft.” Similarly, we can see the characteristics of each category by plotting models in low-dimensional space by using MDS.

C. Relationship among Words

We then plotted the words in 3D space by using MDS. We used the top 50 models with a larger amount of mutual information with word $x^w$ in order to calculate its coordinate. The coordinate of word $x^w$ is the average of the coordinates of these models that are weighted by their mutual information. Fig. 7 shows the results of the 3D plotting of each word. We can see that “Hard” and “Soft” are located close to each other in the 3D space. Moreover, words related to audio information (“Instrument,” “Maraca,” and “Plastic bottle”), words representing objects without sound (“Plush toy,” “Cup,” “Rubber doll,” and “Ball”), and words representing color (“Blue,” “Red,” “Pink,” “Green,” and “Yellow”) are located close to each other. These results show that the distances between models in which similar types of words are represented are short, and that the relationship of each word is reasonably learnt.

VI. CONCLUSION

We proposed the BoMHDP model, which consists of multiple MHDP models, and a word-grounding method involving dialogue between a human and a robot for selecting categories based on the mutual information. The experimental results showed that a large number of categories were formed and reasonably selected. Moreover, we plotted many MHDP models in two- or three-dimensional space by using MDS. From this plot, we could see the relationships among the models and the words and determine the validity of category selection. In future work, we intend to pursue quantitative evaluation of the formed categories as well as an extension of the number of objects.

VII. ACKNOWLEDGMENTS

This work was supported by KAKENHI (23-10330, 23500240, 20500179) and a Grant-in-Aid for Scientific Research on Innovative Areas (The study on the neural dynamics for understanding communication in terms of complex hetero systems).

REFERENCES