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

Integrated Concept of Objects and Human Motions Based on Multi-layered Multimodal LDA

Muhammad Fadlil, Keisuke Ikeda, Kasumi Abe, Tomoaki Nakamura, and Takayuki Nagai

***Abstract*— The human understanding of things is based on prediction which is made through concepts formed by the categorization of experience. To mimic this mechanism in robots, multimodal categorization, which enables the robot to form concepts, has been studied. On the other hand, segmentation and categorization of human motions have also been studied to recognize and predict future motions. This paper addresses the issue on how these different kinds of concepts are integrated to generate higher level concepts and, more importantly, on how the higher level concepts affect the formation of each lower level concept. To this end, we propose the multi-layered multimodal latent Dirichlet allocation (mMLDA), which is an expansion of the MLDA to learn and represent the hierarchical structure of concepts. We also examine a simple integration model and compare it with the mMLDA. The experimental results reveal that the mMLDA leads to a better inference performance and, indeed, forms higher level concepts which integrate motions and objects that are necessary for real-world understanding.**

I. Introduction

In recent years, intelligent robots have been studied and developed extensively. One of the key technologies for such robots is the categorization of perceptual information. This is because the categorization generates concepts, which enable robots to infer unobservable information. It is obvious that such an inference mechanism helps the robots to operate flexibly in unknown environments. We strongly believe that this is the basis of "true" understanding. Many categorization methods concerning various perceptual information have been studied in the literature [1]–[4]. We have proposed a framework of object concept formation based on multimodal categorization by robots using a statistical model called multimodal LDA (MLDA). The MLDA has been proven to enable robots to categorize objects in the same way as humans do [5], [6].

Meanwhile, the segmentation and categorization of human motions have been studied by researchers in the past. For example, [7] segmented and hierarchically categorized human motions to generate the motion symbol tree. The authors have proposed to predict a human motion in the future using the motion symbols. In [9], the double articulation analyzer of human motions has been proposed to do segmentation and categorization of human motions. These works are significant enough to pursue since they provide a key to connect human motions and symbols (language). However, one thing we have to point out regarding these works is the lack in considering the object categories. Since many human motions are deeply related to objects, it is inevitable to learn higher concepts relating object and motion concepts for intelligent robots.

Muhammad Fadlil, Keisuke Ikeda, Kasumi Abe, Tomoaki Nakamura, and Takayuki Nagai are with the Department of Mechanical Engineering and Intelligent Systems, The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu-shi, Tokyo 182-8585, Japan, {mfadlil, citrus-ki, k_ishii, naka_t}@apple.ee.uec.ac.jp, tnagai@ee.uec.ac.jp

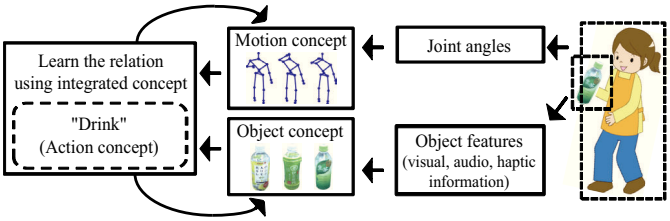

Fig. 1. Schematic of the integrated concepts.

In this paper, we focus on how to form a higher level concept which will represent the relationship between the object concepts and the motion concepts. Fig. 1 conceptualizes the purpose of this paper. In this figure, there are two kinds of lower level concepts: "juice" (object concept) and "take something to mouth" (motion concept). The integration of these concepts provides a higher level concept "drink" (action concept). The important aspect of this model is that inference in various level is possible. In the above example, the robot can recollect "take something to mouth" motion given visual information and/or the sound that the plastic bottle made. Inferring "juice" from the "take something to mouth" motion is also possible, which can be considered as "gesture understanding". It is also worth noting that higher level concepts affect the lower level categorization. For instance, bottle-shaped objects with totally different textures may belong to different object categories; however, if these objects are used with the similar motion "take something to mouth", then the integrated higher concept "drink" affects the lower level object categorization, which leads to a single object category of "juice" (object concept). Another possibility is that the objects with a small difference can be categorized into different classes if these objects are connected to different motions.

To this end, we propose multi-layered multimodal LDA (mMLDA). The mMLDA consists of the bottom-layers as the object and motion concepts, and the top-layer as the integrated action concepts. The robot observes human actions and objects used in the actions. Then, object concepts are formed by the MLDA using multimodal, i.e. visual, auditory, and haptic, information that the robot obtained regarding the objects. The motion concepts are also formed by the MLDA based on the human joint angles obtained using the Kinect loaded on the robot. At the same time, the top-layer tries to capture the relationship between lower level concepts.

978-1-4673-6357-0/13/$31.00 ©2013 IEEE

It should be noted that the categorization processes of all layers are mutually interdependent. Comparison of the proposed mMLDA with a simple approximation model, which has feed-forward connections between MLDAs, reveals the validity of the proposed model.

Recently, there are a great deal of studies on categorization based on various sensor information [1]-[4]. In addition, considerable research on the modeling of human motions has also been conducted [7]-[9]. Although this paper focuses on the categorization of perceptual information, the difference lies in the fact that the proposed model aims at simultaneously categorizing different kinds of concepts and learning their relationships. Therefore, the proposed model enables robots to infer among concepts, e.g. the most likely motion from a visual input, probabilistically.

In [10], a system for mapping between different sensor modalities using Recurrent Neural Network with Parametric Bias (RNNPB) has been proposed. This system enables robots to generate motions expressing auditory signals and sounds that is generated by the object movements. The aim of the paper is to make the system learn direct mapping between the signals from different kinds of sensors. Therefore, categories (concepts) and their interdependence are not considered explicitly. Moreover, the RNNPB may have a problem in the scalability. In [10], only 5 objects were used in the experiment.

Regarding the sensory-motor mapping, affordance learning [11], [12] has been studied in the area of robotics recently. They use Bayesian networks to model the relationship among the objects, actions, and effects. Unfortunately, the model is too simple to represent a complex concept structure, which will be discussed in this paper. Furthermore, the actions are fixed in advance, which means the robots cannot learn novel concepts, but learn only the relationship among the given concepts.

In the area of computer vision, the idea of human-object interaction (HOI) has been proposed in [13]. They showed that the model using the idea of HOI significantly improves the performance of both object detection and human pose estimation. Although [13] focused on the relationship between objects and human motions, the goal of the paper was to improve the recognition performance. Thus the method in [13] is based on a supervised learning, while we are interested in unsupervised learning.

II. Multimodal LDA

To understand the proposed method better, the categorization and inference based on the MLDA will be explained in this section. Latent Dirichlet Allocation (LDA) is a generative probabilistic model for collections of discrete data. The MLDA is an extended version of the LDA that can handle multimodal input signals as shown in Fig. 2. In the figure, $\{x^1, x^2, \cdots\}$ denotes a set of multimodal observations. Each observation is assumed to be drawn from multinomial distribution parameterized by β^*, which is chosen by the Dirichlet distribution for parameter ϕ^*. z is a latent variable corresponding to category, that is drawn from multinomial distribution parameterized by θ. α is the hyperparameter to characterize the Dirichlet prior distribution for θ.

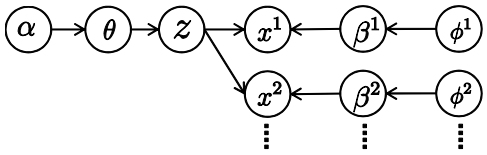

Fig. 2. Graphical model of MLDA.

Gibbs sampling is used to infer parameters of the MLDA. In Gibbs sampling, the category z_{nij}, which is assigned to the i-th data of modality $n \in \{1, 2, \cdots\}$ in the j-th object, is sampled from the following conditional probability:

$$p(z_{nij} = k | \boldsymbol{z}^{-nij}, \boldsymbol{x}^n, \alpha, \pi^n) \propto (N_{kj}^{-nij} + \alpha) \frac{N_{nx^nk}^{-nij} + \pi^n}{N_{nk}^{-nij} + W^n \pi^n}, \quad (1)$$

where W^n denotes the dimensionality of modality n. N_{nx^nkj} represents the frequency count of assigning x^n to the category k for the modality n of the j-th object. Here, N_{nx^nk}, N_{kj}, and N_{nk} can be calculated as follows:

$$N_{nx^nk} = \sum_j N_{nx^nkj}, \quad (2)$$

$$N_{kj} = \sum_{n,x^n} N_{nx^nkj}, \quad (3)$$

$$N_{nk} = \sum_{x^n,j} N_{nx^nkj}. \quad (4)$$

N_{nx^nk}, N_{kj}, and N_{nk} respectively represent the number of assigning x^n to the category k for all objects of the modality n, the number of times of assigning all modalities of the j-th object to the category k, and the frequency of assigning modality n of all objects to the category k. The superscript with the minus sign in Eq. (1) donates exception, e.g., $\boldsymbol{z}^{-nij}$ represents the assigned categories except for z^{nij}.

The category assigned to the i-th data of the modality n of the j-th object is sampled according to Eq. (1). This process is repeated until N_* converges to $\hat{N}_*$. After the convergence, the final estimation of parameters $\hat{\beta}^n_{x^nk}$ and $\hat{\theta}_{kj}$ can be written as follows:

$$\hat{\beta}^n_{x^nk} = \frac{\hat{N}_{nx^nk} + \phi^n}{\hat{N}_{nk} + W^n \phi^n}, \quad (5)$$

$$\hat{\theta}_{kj} = \frac{\hat{N}_{kj} + \alpha}{\hat{N}_j + K\alpha}, \quad (6)$$

where K represents the number of categories.

III. Model of integrated concepts

In this paper, the object and motion concepts, each of which is represented by the MLDA, are integrated to generate higher level concepts. Fig. 3 shows an extension of MLDA to multi-layered MLDA (mMLDA) that can form integrated concepts by capturing the relationship between lower level concepts. In the top-layer (left side of Fig. 3), z represents the integrated category (concept), and the top-layer can be seen as a generative model that generates z^O (object concept) and z^M (motion concept) according to the integrated concept

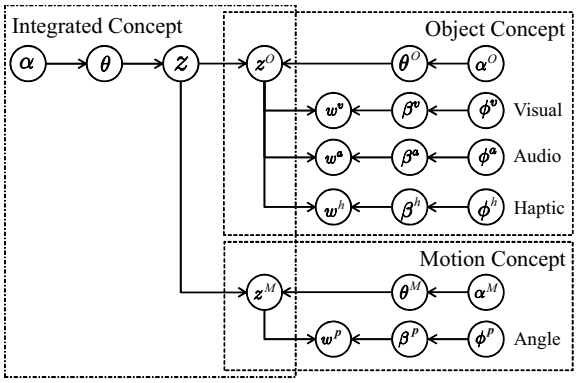

Fig. 3. Graphical model of multi-layered MLDA.

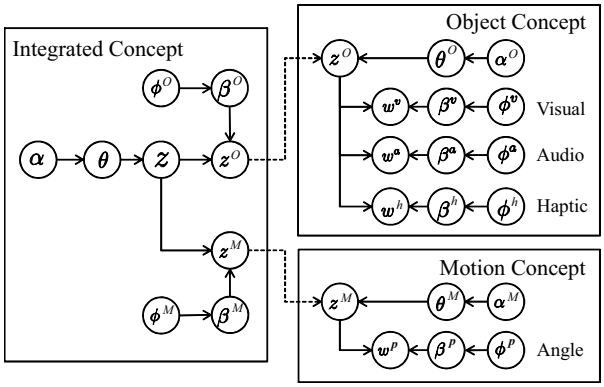

Fig. 4. Approximated model of integrated concept.

z. In the bottom-layer (right side of Fig. 3), visual w^v, auditory w^a, and haptic w^h signals are generated from z^O, and the motion information w^p is generated from z^M. Here, z, z^O and z^M are latent variables, which cannot be observed directly. Therefore, the inference algorithm for latent variables from observations w^* is required.

On the other hand, the easiest way to integrate multiple concepts is to connect multiple MLDAs using simple feed-forward connections as shown in Fig. 4. We call this the approximated model, in which the categorization process is carried out independently.

A. Object concept

The robot forms the object concept based on the categorization of observed multimodal information. The object concepts are represented by the MLDA as illustrated in Fig. 2. In the figure, x^1, x^2, and x^3 correspond to visual, auditory, and haptic information, respectively. Fig. 5 (a) shows the robot platform DiGORO [14] used in this study. The details of each multimodal information are explained below.

Visual information The robot uses the CCD and TOF cameras to obtain visual information. The robot grasps the target object and places it on the observation table (Fig. 5 (b)) that is held in the other hand as shown in Fig. 5 (c). The observation table equipped with an XBee wireless controller enables the robot to rotate the table freely to capture images of the object from various viewpoints. The target object

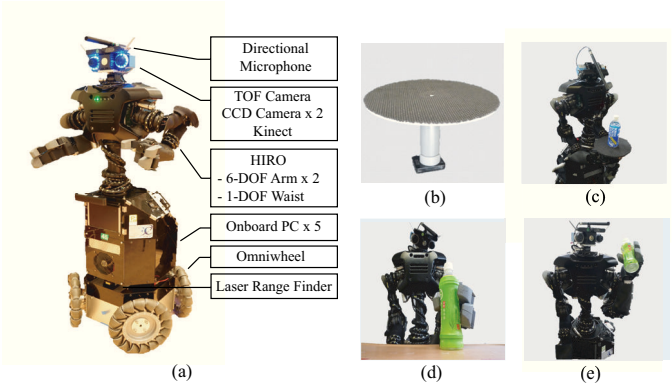

Fig. 5. Robot and acquisition of multimodal information: (a) robot platform used in this study, (b) hand-held observation table, (c) acquisition of visual information, (d) acquisition of haptic information, and (e) acquisition of auditory information.

is segmented out in each image frame, and then 128-dimensional DSIFT [15] descriptors are computed. Thirty-six images are captured for each object. Each feature vector is vector quantized using a codebook with 500 clusters. The codebook is generated by a k-means algorithm in advance. Finally, a 500-dimensional histogram is built as the bag-of-features (BoF) representation.

Haptic information Haptic information is obtained from the three-finger robotic hand with tactile array sensors as shown in Fig. 5 (d). A total of 162 time series of sensor values are obtained by grasping an object. Again, the BoF model is applied to the data, so that any variation resulting from the changes in the grasping point can be absorbed. The feature vectors are vector quantized using a codebook with 15 clusters and a histogram is constructed.

Auditory information The sound is recorded while the robot is grasping and shaking an object as shown in Fig. 5 (e). The sound data are then divided into frames and transformed into 13-dimensional MFCCs as feature vectors. Finally, the feature vectors are vector quantized using a codebook with 50 clusters, and then, a histogram is constructed.

B. Motion concept

The motion concepts are formed by observing human motions which correspond to x^1 in Fig. 2 similar to the above-mentioned object concept formation. The robot captures joint angles of the person in motion using the Kinect placed on its head. There are 11 joints to track and the robot captures them continuously from beginning to end of the motion. We assume that the motion can be segmented according to the object used. Then, a sequence of 11-dimensional vectors is captured for each motion. In order to input this motion information to the MLDA, the BoF representation is preferable. Thus, the feature vectors are vector quantized using a codebook with 70 clusters, and then, a 70-dimensional histogram is constructed. The BoF representation of the motion has been proposed in [16], and it is shown to be useful in the motion recognition task.

C. Integrated concept

The integrated concepts can be formed by learning the relationship between the object and motion concepts using the mMLDA as shown in Fig. 3.

1) Multi-layered MLDA: In the proposed model (Fig. 3), z, z^O, and z^M, each of which represents concept, are latent variables and learned from the observable data w^v, w^a, w^h, and w^p. This can be done by estimating parameters which are sampled from posterior probability. w^v, w^a, w^h, and w^p are assumed to be drawn from each multinomial distribution parameterized by β^v, β^a, β^h, and β^p, respectively. ϕ^v, ϕ^a, ϕ^h, and ϕ^p denote parameters of Dirichlet distributions for β^*. z, z^O, and z^M are assumed to be drawn from each multinomial distribution parameterized by θ, θ^M, and θ^O, respectively. α, α^M, and α^O denote parameters of Dirichlet distributions for θ^*. These parameters are estimated using Gibbs sampling as follows:

$$\begin{aligned} &P(z, z^M, z^O, w^v, w^a, w^h, w^p| \\ &\qquad \boldsymbol{z}, \boldsymbol{z}^O, \boldsymbol{z}^M, \boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^h, \boldsymbol{w}^p) \\ &= P(z)P(z^M|z)P(z^O|z)P(w^p|z^M)P(w^a|z^O) \\ &\qquad P(w^v|z^O)P(w^h|z^O), \end{aligned} \tag{7}$$

$$P(z|\boldsymbol{z}) = \frac{\alpha + N_{jz}}{K\alpha + N_j}, \tag{8}$$

$$P(z^*|z, \boldsymbol{z}, \boldsymbol{z}^*) = \frac{\alpha^* + N_{zz^*}}{L^*\alpha^* + N_z}, \tag{9}$$

$$P(w^m|z^*, \boldsymbol{z}^*, \boldsymbol{w}^m) = \frac{\phi^m + N_{z^*w^m}}{W^m\phi^m + N_{z^*}}, \tag{10}$$

where, N_{jz} represents the number of times assigning all modalities of object j to the higher category z. $N_{z^*w^m}$ represents the frequency of assigning w^m to the lower category z^* for all observed information of the modality m. The categories of the top-layer z and the bottom-layer z^O, z^M, which are assigned to the i-th data of the modality m of the j-th object are sampled according to Eqs. (8)–(10).

The learning process is started from the concept formation in the bottom-layer z^*, which is illustrated in the right side of Fig. 3. At this time, each concept is formed by sampling the value of $z^* \in \{z^O, z^M\}$ using the equation below:

$$\begin{aligned} z^*_{jmi} &\sim P(z^*_{jmi}|w^m_{ji}, \boldsymbol{w}^m_{-ij}, \boldsymbol{z}^*_{-jmi}, \boldsymbol{z}_{-jmi}) \\ &\propto \sum_z P(z|\boldsymbol{z}_{-jmi})P(z^*_{jmi}|\boldsymbol{z}_{-jmi}, \boldsymbol{z}^*_{-jmi}, z) \\ &\quad P(w^m_{ji}|\boldsymbol{w}^m_{-ji}, \boldsymbol{z}^*_{-jmi}, z^*_{jmi}). \end{aligned} \tag{11}$$

After the bottom-layer concepts z^O and z^M were formed, a whole layer of mMLDA is learned in order to form the integrated concept z. Using Gibbs sampling, the value of z can be sampled as

$$\begin{aligned} z_{jmi} &\sim P(z_{jmi}|w^m_{ji}, \boldsymbol{w}^m_{-ij}, \boldsymbol{z}^*_{-jmi}, \boldsymbol{z}_{-jmi}) \\ &\propto \sum_{z^*} P(z_{jmi}|\boldsymbol{z}_{-jmi})P(z^*|\boldsymbol{z}_{-jmi}, \boldsymbol{z}^*_{-jmi}, z_{jmi}) \\ &\quad P(w^m_{ji}|\boldsymbol{w}^m_{-ji}, \boldsymbol{z}^*_{-jmi}, z^*). \end{aligned} \tag{12}$$

Algorithms 1 and 2 denote the learning processes in the bottom-layer and a whole layer, respectively. K and L^* represent respectively the number of categories of the top- and the bottom-layer. It should be noted that the Algorithm 1 is necessary for obtaining good initial values for the Algorithm 2. We found that the Algorithm 2 does not provide a good solution if we start from random initial values. The learning process is repeated until N_* converges to a certain value. After the convergence, the final estimation of parameters $\hat{\beta}^m_{w^m z^*}$, $\hat{\theta^*}_{zz^*}$, and $\hat{\theta}_{jz}$ can be written as follows:

$$\hat{\beta}^m_{w^m z^*} = \frac{N_{z^*w^m m} + \phi^m}{N_{z^* m} + W^m\phi^m}, \tag{13}$$

$$\hat{\theta^*}_{zz^*} = \frac{N_{zz^* m} + \alpha^*}{N_{zm} + L^*\alpha^*}, \tag{14}$$

$$\hat{\theta}_{jz} = \frac{N_{jz} + \alpha}{N_j + K\alpha}, \tag{15}$$

where W^m represents the dimensionality of modality m, and $N_{z^*w^m m}$ represents the frequency of assigning w^m to the lower category z^* for all observed information of the modality m.

Using the learned model, the robot can infer unobservable information. For example, the most probable motion $\hat{z}^M$

Algorithm 1 Multi-layered MLDA (bottom-layer)

for all i, j, m **do**
  $u \leftarrow$ draw from Uniform [0,1]
  for $l \leftarrow 1$ to L^* **do**
    $P[l] \leftarrow P[l-1] + P(z^*_{jmi} = l|w^m_{ji}, \boldsymbol{w}^m_{-ji}, \boldsymbol{z}^*_{-jmi}, \boldsymbol{z}_{-jmi})$
  end for
  for $l \leftarrow 1$ to L^* **do**
    if $u < P[l]/P[L^*]$ **then**
      $z^*_{ij} = l$, break
    end if
  end for
end for

Algorithm 2 Multi-layered MLDA (whole layer)

for all $i, j, *, m$ **do**
  for $k \leftarrow 1$ to K **do**
    $P[k] \leftarrow P[k-1] + P(z_{jmi} = k|w^m_{ji}, \boldsymbol{w}^m_{-ji}, \boldsymbol{z}^*_{-jmi}, \boldsymbol{z}_{-jmi})$
  end for
  $u \leftarrow$ draw from Uniform [0,1]
  for $k \leftarrow 1$ to K **do**
    if $u < P[k]/P[K]$ **then**
      $z_{ij} = k$, break
    end if
  end for
  for $l \leftarrow 1$ to L^* **do**
    $P[l] \leftarrow P[l-1] + P(z^*_{jmi} = l|w^m_{ji}, \boldsymbol{w}^m_{-ji}, \boldsymbol{z}^*_{-jmi}, \boldsymbol{z}_{-jmi})$
  end for
  $u \leftarrow$ draw from Uniform [0,1]
  for $l \leftarrow 1$ to L^* **do**
    if $u < P[l]/P[L^*]$ **then**
      $z^*_{ij} = l$, break
    end if
  end for
end for

can be recollected using only the visual information of the object $\boldsymbol{w}^v$, and vice versa. Of course, another perceptual information, such as $\boldsymbol{w}^a$ and $\boldsymbol{w}^h$, can be incorporated for inference. Such an inference can be made by the following equation:

$$\hat{z}^M = \underset{z^M}{\operatorname{argmax}} \sum_z \sum_{z^O} P(z)P(z^M, z^O|z) \times P(\boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^t|z^O). \quad (16)$$

In the same way, the most probable object $\hat{z}^O$ can be inferred for a given motion $\boldsymbol{w}^p$:

$$\hat{z}^O = \underset{z^O}{\operatorname{argmax}} \sum_z \sum_{z^M} P(z)P(z^O, z^M|z)P(\boldsymbol{w}^p|z^M). \quad (17)$$

2) Approximated model: As we mentioned earlier, the easiest way to integrate multiple concepts is to connect independent MLDAs in a feed-forward way. In Fig. 4, the integrated concept is represented by z that is to be learned from the object concept z^O and motion concept z^M in order. Since we will compare the mMLDA with the approximated model in the evaluation, the mechanism behind the approximated model will be discussed briefly.

Here, z^O and z^M are drawn from multinomial distributions $P\left(z^O \mid \boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^h\right)$ and $P\left(z^M \mid \boldsymbol{w}^p\right)$ respectively in the bottom-layer. As we have already explained, the object concept z^O and motion concept z^M are generated using independent MLDAs. In the top-layer of the approximated model shown in the left side of Fig. 4, z^O and z^M are assumed to be x^1 and x^2 in Fig. 2, respectively. Thus, the relationship between two kinds of concepts is learned by the model, in which the latent variable z represents integrated concepts (actions) of the objects and motions.

Using the learned model, unobservable information can be inferred. For example, the most likely motion for given observations regarding the unseen object $\boldsymbol{w}^m_{obs}$ can be inferred as:

Step 1 : Infer the category of the object using

$$\hat{z}^O \sim P\left(z^O \mid \boldsymbol{w}^v_{obs}, \boldsymbol{w}^a_{obs}, \boldsymbol{w}^h_{obs}\right). \quad (18)$$

Step 2 : Infer the most likely category of motion $\hat{z}^M$ by calculating $P(\hat{z}^M|\hat{z}^O)$ using the following equation;

$$P(\hat{z}^M|\hat{\boldsymbol{z}}^O) = \int \sum_z P(\hat{z}^M|z)P(z|\theta)P(\theta|\hat{\boldsymbol{z}}^O)d\theta, \quad (19)$$

where z represents the category of integrated concept. Objects that are related to an observed unseen motion can also be inferred in the same manner.

As we will see in the experiments, qualitative difference between the mMLDA and the approximated model is clear. Although the approximated model has some good points, such as simple and easy to implement, there is an obvious drawback. In the model, the error that is occurred in the bottom-layer is propagated, which may degrade the performance of the model considerably. This is because the learning of each MLDA carries out independently and there is no chance to make the categorization better considering the different kinds of concepts. In contrast, the mMLDA tries to capture the structure of all perceptual information as a whole, which makes the model more powerful to do categorization and inference.

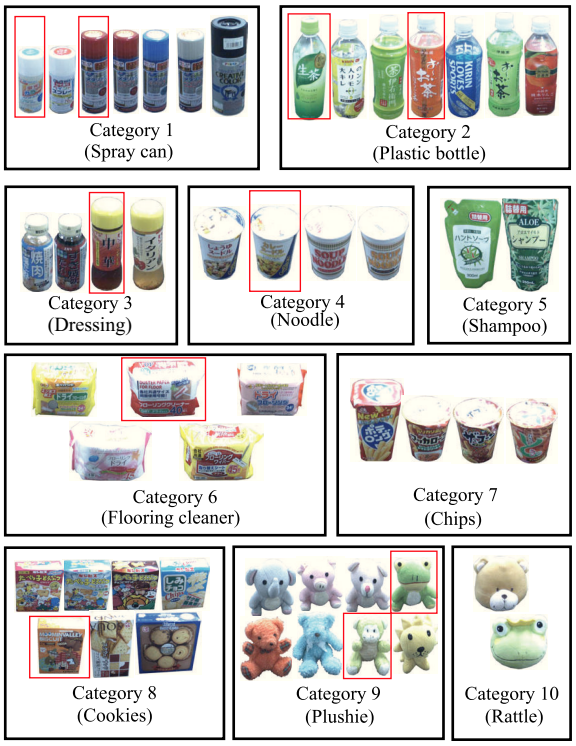

Fig. 6. The fifty objects used in the experiments; the red rectangles show the objects used in the recognition test.

TABLE I
MOTION PERFORMED ON THE OBJECTS (NUMBER IN THE PARENTHESIS REPRESENTS CATEGORY INDEX)

Motion	Object	Motion	Object
Put on (1)	Dressing (3)	Place (7)	Noodle (4)
Shake (2)	Spray can (1)		Chips (7)
	Plastic bottle (2)		Cookies (8)
	Dressing (3)	Throw (8)	Plushie (9)
	Rattle (10)	Open (9)	Spray can (1)
Drink (3)	Plastic bottle (2)	Open (10)	Plastic bottle (2)
Eat (4)	Noodle (4)	Open (11)	Flooring cleaner (6)
	Chips (7)	Pour (12)	Shampoo (5)
	Cookies (8)	Hug (13)	Plushie (9)
Wipe (5)	Flooring cleaner (6)	Pet (14)	Plushie (9)
Paint (6)	Spray can (1)		

IV. EXPERIMENTS

The experiments were carried out to evaluate the proposed model. Fig. 6 shows the 50 objects used in the experiments. The objects were manually classified into 10 categories, which provide the ground truth. The 40 objects without red rectangles in Fig. 6 were used for the categorization experiments, while the other 10 objects with red rectangles were used as the test set (unseen objects) in the inference experiment. The object concepts are formed based on the multimodal information that the robot obtained by observing each object. As for the human motions, the robot captured human motions in the use of objects using the Kinect. Table I shows the ground truth of the correspondence between

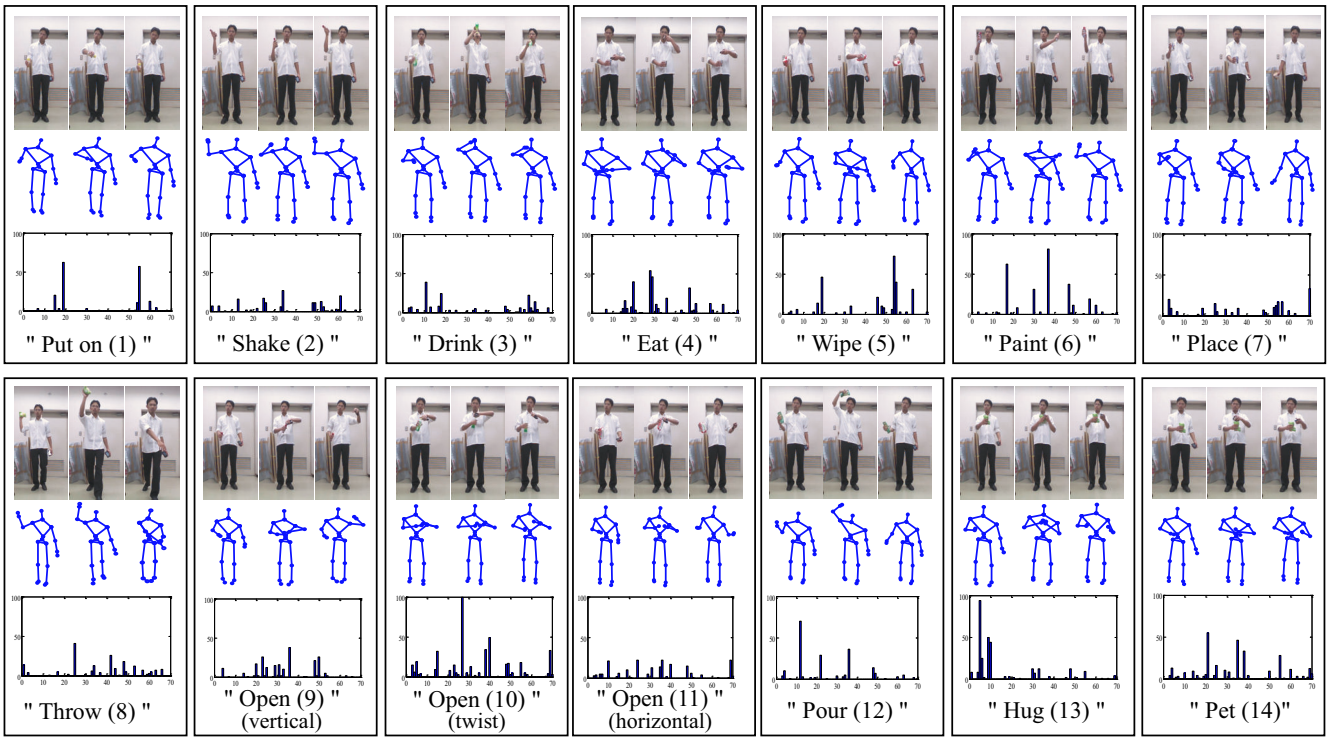

Fig. 7. Examples of the acquired motion information for each motion: (from top to bottom) actual images, acquired images from Kinect, and 70 dimensions of motion histogram (number in the paranthesis represents category index).

objects and human motions. Fig. 7 shows the examples of the captured images and the motion data. We set 14 motion categories as the ground truth, since the motions were classified into 14 categories by hand.

How to decide the number of categories is an important problem for the LDA. The mMLDA suffers from the same problem as it is based on the LDA. Here, we decided to use the number of categories defined by the ground truth. There have been many efforts on this issue, and we think it is possible to apply one of these methods, e.g. nonparametric Bayesian method. Worse still, the mMLDA requires the number of categories for the top-layer, which does not have the ground truth. We conducted the experiments several times with different number of higher level categories and decided to use 9, since it gave relatively good results in our experiment. One thing we have to mention is that the performance was not so sensitive to that number in our experiment. Although we could not find any theoretical reason to use that number, this problem can be solved by using the nonparametric Bayesian method that finds the number of categories from the input data automatically.

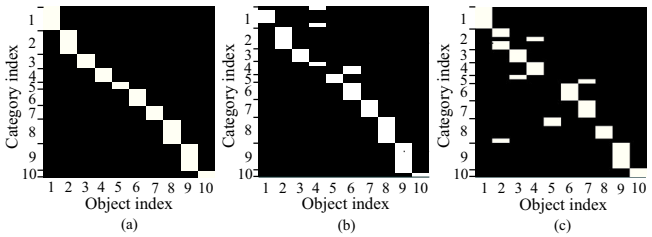

Fig. 8. Object categorization result: (a) the ground Truth, (b) the mMLDA, and (c) the approximated model.

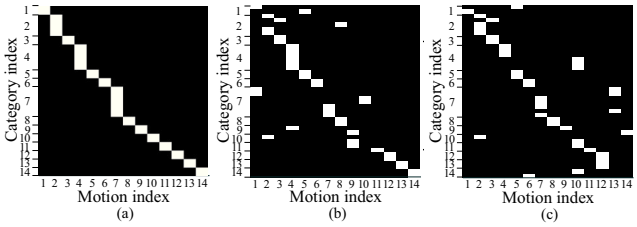

Fig. 9. Motion categorization result: (a) the ground Truth, (b) the mMLDA, and (c) the approximated model.

A. Object concept formation

The results of object concept formation are shown in Fig. 8 as the confusion matrices: (a) the ground truth, (b) the mMLDA, and (c) the approximated model. The vertical axis represents the index of each object category, and the horizontal axis represents resultant categories. The categorization accuracy of the mMLDA is 87.5%, while the approximated model provides 85.0%.

From the result of the approximated model (Fig. 8 (c)), one can see that "Cookies (8)" are divided into 3 categories, since they do not share a common visual information (texture). This kind of difference in perceptual data can easily divide the category into two or more, even though the other features are common in objects that belong to the same category. This is caused by the independent categorization process in the approximated model.

On the other hand, the mMLDA correctly categorized

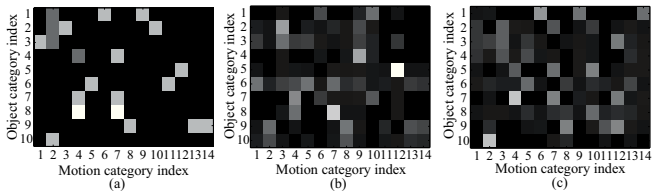

Fig. 10. Integrated concept: (a) ground truth (based on Table I), (b) the mMLDA and (c) the aprroximated model.

"Cookies (8)" into a single category. This is because "Cookies (8)" share the same motions "Eat (4)" and "Place (7)". In the proposed mMLDA, the motion concept formation also affects the object concept formation. This mutual interdependence among concepts helps to form a single "Cookies (8)" concept in this experiment, and the result clearly indicates the importance of such mutual interdependence.

B. Motion concept formation

The motion categorization results are shown in Fig. 9: (a) the ground truth, (b) the mMLDA, and (c) the approximated model. The vertical axis represents the index of actual category of the motion, and the horizontal axis represents the index of the classification result. The accuracy is 72.5% for the mMLDA, and it drops to 62.5% for the approximated model.

The difference can be seen in the categorization result of "Pour (12)" and "Hug (13)". The approximated model categorized "Pour (12)" and "Hug (13)" into a single category, while the mMLDA, as shown in Fig. 9 (b), correctly classified them into two different categories. In fact, the histograms of "Pour (12)" and "Hug (13)" were somewhat similar and confusing sometimes. Therefore, it is likely that these two motions are put together in a single category by the LDA.

In the mMLDA, the objects "Shampoo (5)" and "Plushie (9)" affected the motion concept formation, resulting in the correct categorization of "Pour (12)" and "Hug (13)". Again, one can see the advantage of the proposed mMLDA over the approximated model.

C. Integrated concept

Here we examine the integrated concepts, which were formed at the top-layer.

1) Joint probability of motion and object: The joint probabilities $P(z^O, z^M)$ were calculated using the mMLDA and the approximated model. Besides, by counting the number of training samples, we can compute $P(z^O, z^M)$ as a ground truth. Please note that the ground truth cannot be obtained in practice unless the object and motion categories are perfectly recognized using observations.

Fig. 10 (a) shows the ground truth, which was generated according to Tab. I. Figs. 10 (b) and (c) represent results of the mMLDA and the approximated model, respectively. The vertical and horizontal axis in the figure indicate the indexes of objects and motions, respectively. From these figures, one can see that the result of the mMLDA is closer to the ground truth. In fact, the KL-distances between the approximated model and the ground truth, and between the mMLDA and the ground truth are 50.25 and 46.50, respectively.

If we look at the details of the mMLDA's result, it can be seen that "Cookies (8)" occurred together with "Eat (4)" and "Place (7)" more often, which is a similar tendency to the ground truth. In the approximated model, "Cookies (8)" have dispersed joint probabilities over all motions. The same thing can be observed for "Shampoo (5)", which is only related to "Pour (12)" in the ground truth and the mMLDA, while the probabilities are spread over "Eat (4)", "Place (7)", and "Open (10)" in the approximated model.

2) Integrated concept formation: In the top-layer of the proposed mMLDA, the integrated concepts, which encode the relationships among the object and motion concepts, are formed. Here we will see that the integrated (higher-level) concepts were actually generated.

In the mMLDA, one of the integrated concepts consists of the motion "Eat (4)" and three object categories "Noodle (4)", "Chips (7)", and "Cookies (8)". Obviously, this concept represents "Eat something" behaviour. "Drink (3)" and "Plastic bottle (2)" are classified into a single category, which can be considered as "Drink" concept.

As for the approximated model, a category including "Place (7)", "Chips (7)", and "Cookies (8)" was formed, which can be considered as "Place something". However, due to the misclassification of "Hug (13)" and "Pour (12)" in the bottom-layer of the approximated model, a category of "Pour (12)", "Hug (13)", "Shampoo (5)", and "Plushie (9)" was formed in the top-layer. This example clearly indicates a drawback of the simple feed-forward model; the errors that occurred at the bottom-layer, unconditionally propagate through the layers.

D. Inference of unseen Information

Next, we have performed experiments on the inference of unseen information to evaluate the model. The experiments were conducted using the test set objects in Fig. 6 (marked with red rectangles). The inference of motion concept z^M was performed by observing multimodal information, $\boldsymbol{w}^v$, $\boldsymbol{w}^a$, and $\boldsymbol{w}^h$, regarding the target object. The object concept z^O was also inferred by observing a human motion.

Firstly, the mMLDA inferred the motion concept z^M from the observed object information as accurate as 80.0%. Meanwhile, the approximated model yielded z^M of 70.0%. Fig. 11 shows the inference results of motions from an unseen object using the mMLDA (left) and the approximated model (right). The graph represents the probability of each motion which is inferred from the observations of the novel "Noodle (4)" by the robot. The vertical and horizontal axis represent the probability and the index of each motion, respectively. As shown in Fig. 11 (a), the mMLDA, correctly inferred the motion "Eat (4)" with the highest probability. On the other hand, the approximated model inferred the motion "Wipe (5)" with the highest probability. In the approximated model, this kind of false inference is due to the categorization errors at the bottom-layer. In the above example, a part of "Noodle (4)" concepts and "Flooring cleaner (6)" concepts were mixed together as one category. This category is responsible for the confusion of the "Eat (4)" with "Wipe (5)" motions when the robot saw the "Noodle (4)".

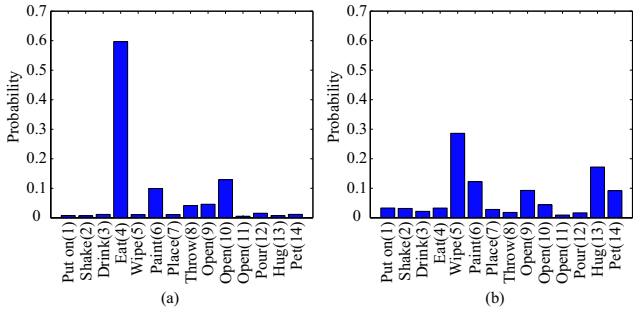

Fig. 11. The probability of each motion when "Noodle (4)" is observed as the unseen information using (a) the mMLDA and (b) approximated model.

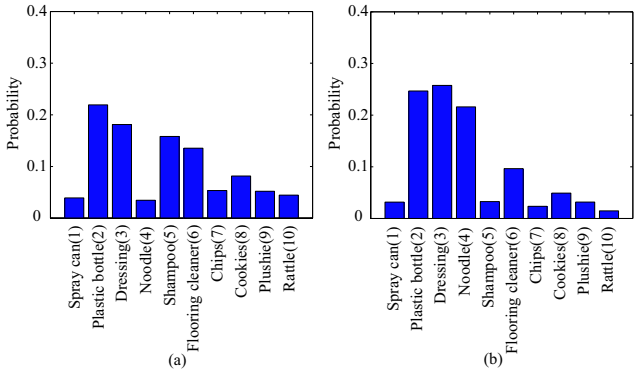

Fig. 12. The probability of each object when "Drink (3)" is observed as the unseen information using (a) the mMLDA and (b) approximated model.

The false inference occurred in the mMLDA when a "Spray can (1)" was given then the "Throw (8)" motion was inferred. It happened that the "Throw (8)" and "Shake (2)" motions share similar features which result in a single higher-level concept consisting of these two motion concepts. Such a concept caused the false inference; indeed, the probability of "Shake (2)" motion (correct motion) was the second highest.

When the robot observed only human motions, the object concept z^O could be inferred using the mMLDA with an accuracy of 70.0%. The accuracy of the approximated model was 60.0%. Fig. 12 shows the probability of each object, which was inferred from the motion "Drink (3)". The vertical and horizontal axis represent the probability and the index of object category, respectively. This result shows one of the successful cases of observing "Drink (3)" motion; the object "Plastic bottle (2)" is correctly inferred, as it has the highest probability as shown in Fig. 12 (a) using the mMLDA. The approximated model inferred the "Plastic bottle (2)" with high probability, yet not the highest. From Fig. 12 (b), one can see that "Dressing (3)" has the highest probability. This false inference is also caused by the categorization errors at the bottom-layer.

An example of the false inference by the mMLDA is the inference of "Flooring cleaner (5)" from "Place (7)" motion. The reason for this false inference is the similarity between "Place (7)" and "Wipe (5)" motions. These two motions form a single higher-level concept that gave a negative effect on the inference.

V. Conclusion

In this paper, we proposed the multi-layered MLDA, which bundles lower-level concepts in the bottom-layers into higher-level concepts at the top-layer. More precisely, the object and motion concepts are integrated to generate the action concepts. We evaluated the mMLDA through some experiments that validated the proposed model. The mMLDA was also compared with a simple approximated model and it was revealed that the mutual interdependence is the key feature in the formation process of the multi-layered concept.

Clearly, the mMLDA can be generalized to as many layers as we want, starting with the integration of a large number of different concepts at the bottom-layer. We are currently working on this direction to integrate a great variety of different concepts using the mMLDA. Furthermore, the issue of determining the number of categories autonomously should be solved. This can be achieved using the nonparametric Bayesian method, such as hierarchical Dirichlet process (HDP), instead of the LDA.

References

[1] R. Fergus, P. Perona, and A. Zisserman, "Object Class Recognition by Unsupervised Scale-Invariant Learning", in Proc. of CVPR 2003, vol.2, pp.264–271, 2003

[2] J. Sivic, B.C. Russell, A.A. Efros, A. Zisserman, and W.T. Freeman, "Discovering Object Categories in Image Collections", in Proc. of ICCV 2005, pp.370–377, 2005

[3] E. Torres-Jara, L. Natale, and P. Fitzpatrick, "Tapping into Touch," Lund University Cognitive Studies, pp.22–24, 2005

[4] J. Sinapov, and A. Stoytchev, "Object Category Recognition by a Humanoid Robot Using Behavior-grounded Relational Learning", in Proc. of ICRA 2011, pp.184–190, 2011

[5] T. Nakamura, T. Nagai, and N. Iwahashi, "Multimodal Object Categorization by a Robot", in Proc. of IROS 2007, pp.2415–2420, 2007

[6] T. Nakamura, T. Nagai, and N. Iwahashi, "Grounding of Word Meanings in Multimodal Concepts Using LDA", in Proc. of IROS 2009, pp.3943–3948, 2009

[7] W. Takano, H. Imagawa, and Y. Nakamura, "Prediction of Human Behaviors in the Future through Symbolic Inference", in Proc. of ICRA 2011, pp.1970–1975, 2011

[8] W. Takano, and Y. Nakamura, "Bigram-Based Natural Language Model and Statistical Motion Symbol Model for Scalable Language of Humanoid Robots", in Proc. of ICRA 2012, pp.1232–1237, 2012

[9] T. Taniguchi, and S. Nagasaka, "Double Articulation Analyzer for Unsegmented Human Motion using Pitman-Yor Language Model and Infinite Hidden Markov Model", in Proc. of SII 2011, pp.250–255, 2011

[10] T. Ogata, S. Nishide, H. Kozima, K. Komatani, and H. Okuno, "Intermodality Mapping in Robot with Recurrent Neural Network", Pattern Recognition Letters, vol.31, no.12, pp.1560–1569, 2010

[11] L. Montesano. M. Lopes, A. Bernardino, and J. S.-Victor, "Learning Object Affordances: From Sensory-Motor Coordination to Imitation", IEEE Trans. on Robotics, vol.24, no.1, Feb. 2008

[12] B. Moldovan, P.Moreno, M. Otterlo, J. S.-Victor, and L. D. Raedt, "Learning Relational Affordance Models for Robots in Multi-Object Manipulation Tasks", in Proc. of ICRA 2012, pp.4373–4378, 2012

[13] B. Yao, and L. Fei-Fei, "Recognizing Human-Object Interactions in Still Images by Modeling the Mutual Context of Objects and Human Poses", IEEE Trans. on PAMI, vol.34, pp.1691–1703, Sep. 2012

[14] T. Nakamura, K. Sugiura, T. Nagai, N. Iwahashi, T. Toda, H. Okada, and T. Omori, "Learning Novel Objects for Extended Mobile Manipulation", Journal of Intelligent and Robotic Systems, vol.30, pp.1–18, 2011

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

[16] O. Mangin, and P.-Y. Oudeyer, "Learning to Recognize Parallel Combinations of Human Motion Primitives with Linguistic Descriptions using Non-negative Matrix Factorization", in Proc. of IROS 2012, pp.3268–3275, 2012