The 2009 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 11-15, 2009 St. Louis, USA

Grounding of Word Meanings in Multimodal Concepts Using LDA

Tomoaki Nakamura[1], Takayuki Nagai[1], Naoto Iwahashi[2]
[1]Department of Electronic Engineering, The University of Electro-Communications
1-5-1 Chofugaoka Chofu-shi, Tokyo 182-8585, Japan
[2]NICT Knowledge Creating Communication Research Center
2-2-2 Hikaridai, Seika-cho Souraku-gun, Kyoto 619-0288, Japan
naka_t@apple.ee.uec.ac.jp, tnagai@ee.uec.ac.jp, naoto.iwahashi@atr.jp

***Abstract*— In this paper we propose LDA-based framework for multimodal categorization and words grounding for robots. The robot uses its physical embodiment to grasp and observe an object from various view points as well as listen to the sound during the observing period. This multimodal information is used for categorizing and forming multimodal concepts. At the same time, the words acquired during the observing period are connected to the related concepts using multimodal LDA. We also provide a relevance measure that encodes the degree of connection between words and modalities. The proposed algorithm is implemented on a robot platform and some experiments are carried out to evaluate the algorithm. We also demonstrate a simple conversation between a user and the robot based on the learned model.**

***Index Terms*— Multimodal categorization, symbol grounding, Latent Dirichlet Allocation**

I. INTRODUCTION

It is well known fact that the capability of categorizing objects is very important for our human-like intelligence[1]-[3]. The generated categories are the bases of our concepts and each word works as a label of a specific category. Therefore, the categorization is very important for the language understanding as well. These facts motivate us to pursue abilities of the categorization and the symbol grounding algorithm for intelligent robots.

In this paper we examine the algorithm for grounding of word meanings. To achieve this goal we take two-step procedure, that is, the categorization and the mapping of categories to linguistic labels. The categorization can be considered as a problem of unsupervised learning. Unsupervised learning of objects using only images has been extensively studied in the field of computer vision[4]-[8]. Such unsupervised framework enhances the pliability of object recognition systems in various environments. However, it is obvious that object categories do not depend only on visual information but also various one. In [9], we have proposed multimodal categorization that is based on the pLSA(probabilistic Latent Semantic Analysis). The multimodal categorization has been shown to be successful for categorizing objects in the same way as humans do. Then the robot can recognize the category of an unseen object. The validity of the method is, rather, to be able to infer properties of the objects from limited observations. For example, the robot can stochastically infer the sound and/or hardness of the object only from the visual information. This kind of inference is required in day-to-day situations. However, the pLSA requires heuristics to deal with novel input data, since it is point estimation[9]. In order to solve this problem, LDA (Latent Dirichlet Allocation)[10] is extended to multimodal LDA, which is applied to the multimodal categorization, in this paper. In contrast to the multimodal pLSA, the multimodal LDA requires no heuristics, since it is based on Bayesian learning. We also make a performance comparison between pLSA and LDA in this paper.

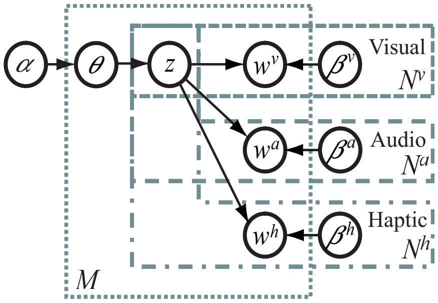


Fig. 1. The graphical model for multimodal LDA.

In the second step, association of words with corresponding categories is carried out. Since we assume that the robot has a vocabulary, the problem that we consider here is purely a correspondence problem. The multimodal LDA is also involved in the proposed framework, hence the robot can stochastically recall words from observations and vice versa. We also consider the problem of measuring the connection between words and modalities. For example, the word 'soft' represents haptic information, hence strong connection between the word and haptic channel can be observed by the proposed measure. This measurement is important when the robot describes object property regarding a certain modality.

Related works include unsupervised categorization of objects using visual information[4]-[8] as we mentioned earlier. Language acquisition is an active research field recently[11]-[12] and is closely related to this paper. However, the multimodal categorization is not involved in these works. In [13], the same problem, that is categorization and grounding, has been tackled. However, [13] does not utilize multimodal categorization. Moreover, it considers only object names as words to be connected, while this paper deals with adjectives as well as nouns.

II. MULTIMODAL CATEGORIZATION USING LDA

The proposed method consists of two steps. In this section, LDA-based multimodal categorization is described as the first step.

A. Overview of LDA based multimodal categorization

The robot can grasp an object and can observe it from different viewpoints. During the observation, identity of the object is guaranteed. This fact motivates us to use occurrence

978-1-4244-3804-4/09/$25.00 ©2009 IEEE

frequency of visual, audio and haptic information, which are collected over the observing period of a single object. This is nothing short of the 'bag of words' model when each feature is considered as 'word'. This idea makes it possible to discover object categories using multimodal information by robots.

The robot is equipped with cameras, microphones, an arm and a hand with pressure sensors. Therefore the robot can actually grasp each object and gain information such as the sequence of images, and signals from microphones and pressure sensors. Then the objects are categorized based on similarities in their appearance, sound that are made by objects in motion and hardness. The graphical model of the proposed LDA-based multimodal categorization is shown in Fig.1. In the figure, w^v, w^a and w^h represent visual, audio, haptic information and are assumed to be drawn from each multinomial distribution parameterized by β^v, β^a and β^h, respectively. z denotes the category and is chosen from multinomial distribution parameterized by θ that depends on Dirichlet prior distribution $Dir(\alpha)$.

B. Signal processing for the multimodal categorization

Each signal is pre-processed for the multimodal categorization as follows:

1) Visual information: The robot has a stereo-camera that is attached to its head and images are grabbed while it grasps and observes an object. Those images are used as visual information (100 images are used in the later experiment). For each image, 128-dimensional SIFT descriptors[14] are computed, then each feature vector is vector quantized using a codebook with 500 clusters. To cope with occlusion by the robot's own hand, images of the robot hand are collected in advance and features (codebook indices) are computed. This set of features is removed from the visual information all the time.

2) Audio information: As for audio information, the sound is recorded while the robot grasps and shakes an object. The audio signal is then divided into frames followed by the transformation into 13-dimensional MFCC (Mel-Frequency Cepstrum Coefficient) as feature vector. Finally, the feature vectors are vector quantized using the codebook with 50 clusters.

3) Haptic information: Haptic information is obtained through the two-finger robotic hand with four pressure sensors. When the robot grasps an object, sum of digitized voltages from these pressure sensors, which encodes hardness of the object, is obtained. During the two-finger grasp, the robot presses the object with two fingers and the amount of change in the angle between the base and left finger is measured. This change in angle can be considered as 'softness' of the object. Thus we obtain two dimensional feature vectors as haptic information. The feature vectors are finally vector quantized using the codebook with 5 clusters.

C. Multimodal LDA

The categorization is carried out as parameter estimation of the graphical model in Fig.1 using multimodal information observed by the robot. Parameters are estimated so that the log likelihood of the multimodal information under the model is maximized. Since the direct computation of the log likelihood is intractable, we apply variational inference, which provides us with a tractable lower bound on the log likelihood using Jensen's inequality[15]. For given multimodal information $\boldsymbol{w}^v$, $\boldsymbol{w}^a$ and $\boldsymbol{w}^h$, the log likelihood can be written as:

$$\begin{aligned}
&\log p(\boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^h | \alpha, \beta^v, \beta^a, \beta^h) \\
&\quad = \log \int \sum_{\boldsymbol{z}} \frac{p(\theta, \boldsymbol{z}, \boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^h | \alpha, \beta^v, \beta^a, \beta^h)}{q(\theta, \boldsymbol{z} | \gamma, \phi^v, \phi^a, \phi^h)} \\
&\qquad\qquad \times q(\theta, \boldsymbol{z} | \gamma, \phi^v, \phi^a, \phi^h) d\theta \\
&\quad \geq \int \sum_{\boldsymbol{z}} q(\theta, \boldsymbol{z} | \gamma, \phi^v, \phi^a, \phi^h) \\
&\qquad\qquad \times \log p(\theta, \boldsymbol{z}, \boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^h | \alpha, \beta) d\theta \\
&\qquad - \int \sum_{\boldsymbol{z}} q(\theta, \boldsymbol{z} | \gamma, \phi^v, \phi^a, \phi^h) \\
&\qquad\qquad \times \log q(\theta, \boldsymbol{z} | \gamma, \phi^v, \phi^a, \phi^h) d\theta, \qquad (1)
\end{aligned}$$

where $q(\theta, \boldsymbol{z} | \gamma, \phi^v, \phi^a, \phi^h)$ is the variational distribution, which approximates $p(\theta, \boldsymbol{z} | \boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^h,\ \alpha, \beta)$ and assumed to be the product of independent terms. ϕ^* denotes the variational parameter of multinomial distribution from which z is sampled and γ is the variational parameter of Dirichlet distribution from which the multinomial parameter θ is drawn. The variational EM algorithm for the proposed multimodal LDA is as follows:

[E-step]

Following procedures are repeated until convergence for each object d.

$$\phi^v_{dw^v k} \propto \beta^v_{kw^v} \exp\left(\psi(\gamma_{dk}) - \psi(\sum_{k'} \gamma_{dk'}) \right) \qquad (2)$$

$$\phi^a_{dw^a k} \propto \beta^a_{kw^a} \exp\left(\psi(\gamma_{dk}) - \psi(\sum_{k'} \gamma_{dk'}) \right) \qquad (3)$$

$$\phi^h_{dw^h k} \propto \beta^h_{kw^h} \exp\left(\psi(\gamma_{dk}) - \psi(\sum_{k'} \gamma_{dk'}) \right) \qquad (4)$$

$$\gamma_{dk} = \alpha_k + \sum_{w^v} \phi^v_{dw^v k} + \sum_{w^a} \phi^a_{dw^a k} + \sum_{w^h} \phi^h_{dw^h k} \qquad (5)$$

[M-step]

$$\beta^v_{kw^v} \propto \sum_d n_{dw^v} \phi^v_{dw^v k} \qquad (6)$$

$$\beta^a_{kw^a} \propto \sum_d n_{dw^a} \phi^a_{dw^a k} \qquad (7)$$

$$\beta^h_{kw^h} \propto \sum_d n_{dw^h} \phi^h_{dw^h k} \qquad (8)$$

$$\begin{aligned}
\frac{\partial L}{\partial \alpha_k} &= N \left(\psi(\sum_{k'} \alpha_{k'}) - \psi(\alpha_k) \right) \\
&+ \sum_d \left(\psi(\gamma_{dk}) - \psi(\sum_{k'} \gamma_{dk'}) \right), \qquad (9)
\end{aligned}$$

where $d(= 1, \cdots, N)$, k and n_{dw^*} represent index of the object, index of category, and occurrence count of a feature w^* for the object d, respectively. α_k is computed using Newton-Raphson method so that the log likelihood L is maximized.

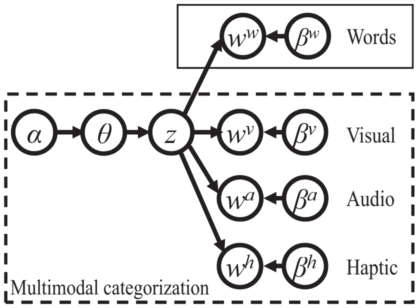


Fig. 2. The graphical model for word acquisition.

D. Category recognition for unseen objects

By using the learned model, the category of unseen object can be recognized. For given multimodal information of unseen object $\boldsymbol{w}^v_{obs}$, $\boldsymbol{w}^a_{obs}$ and $\boldsymbol{w}^h_{obs}$, its category can be determined as z that maximizes $p(z|\boldsymbol{w}^v_{obs}, \boldsymbol{w}^a_{obs}, \boldsymbol{w}^h_{obs})$. Therefore, the category can be found by computing:

$$\begin{aligned} \hat{z} &= \underset{z}{\operatorname{argmax}}\, p(z|\boldsymbol{w}^v_{obs}, \boldsymbol{w}^a_{obs}, \boldsymbol{w}^h_{obs}) \\ &= \underset{z}{\operatorname{argmax}} \int p(z|\theta)p(\theta|\boldsymbol{w}^v_{obs}, \boldsymbol{w}^a_{obs}, \boldsymbol{w}^h_{obs})d\theta, \quad (10) \end{aligned}$$

where $p(\theta|\boldsymbol{w}^v_{obs}, \boldsymbol{w}^a_{obs}, \boldsymbol{w}^h_{obs})$ is determined by recalculating α using variational EM algorithm described above, while learned β^v, β^a and β^h are kept fixed.

E. Inference among modalities

From visual information, we can infer hardness of the object, whether the object makes sound or not, and so on. Such inference among modalities is very important capability for robots as well as for us human. Let us think about the inference of auditory information w^a only from the observed visual information $\boldsymbol{w}^v_{obs}$:

$$p(w^a|\boldsymbol{w}^v_{obs}) = \int \sum_z p(w^a|z)p(z|\theta)p(\theta|\boldsymbol{w}^v_{obs})d\theta. \quad (11)$$

In the above equation, $p(\theta|\boldsymbol{w}^v_{obs})$ should be recomputed in the same way as before. It should be noted that the recomputation implies that the inference is carried out through categories, since the probability of generating category z from $\boldsymbol{w}^v_{obs}$ is recomputed. Furthermore, one can see that Eq.(11) performs the Bayesian inference, which is the essential difference between LDA and pLSA.

III. Grounding words in multimodal concepts

In the foregoing section, we have discussed the algorithm for forming multimodal concept using LDA. In this section, we propose the method for grounding words in the multimodal concepts which are formed by multimodal LDA. The multimodal LDA framework is also involved here as shown in Fig.2. Hence the symbol grounding comes to the problem of the following parameter estimation.

A. Parameter estimation

Figure 2 shows the proposed graphical model. The part in dashed line has been learned by multimodal categorization. In this model, w^w denotes words information which is represented by the 'Bag-of-words' model. Hence, similar to the perceptual information, words information is modeled by occurrence frequency and assumed to be chosen from multinomial distribution parameterized by β^w. At first the robot collects sentences, which are uttered by a user, during the observing period of multimodal information. Continuous speech recognition and morphological analysis are utilized for converting speech signals into sequences of words. Only nouns and adjectives are extracted from these sequences of words and represented as numerical ID. Finally, the robot obtains set of words corresponding to each object. The estimation of parameter β^w is straightforward, since the category z is not a latent variable at this moment.

$$\beta^w_z = \frac{n_{w,z}}{\sum_w n_{w,z}}, \quad (12)$$

where $n_{w,z}$ represents occurrence count of the word w for the category z.

B. Inference of words

Now, all of parameters in Fig.2 has been estimated. This means that the robot is ready for inferring word meanings, that is, the robot can recall highly probable perceptual information from input words. Conversely, the robot can also describe the input multimodal perceptual information (e.g. scene) using suitable words. These processes are realized by computing $p(w^*|\boldsymbol{w}^w_{obs})$ and $p(w^w|\boldsymbol{w}^*_{obs})$ in the same way as in Eq.(11).

Here, let us focus our attention on the example of inferring words w^w only from visual one $\boldsymbol{w}^v_{obs}$. Such inference can be carried out as follows:

$$p(w^w|\boldsymbol{w}^v_{obs}) = \int \sum_z p(w^w|z)p(z|\theta)p(\theta|\boldsymbol{w}^v_{obs})d\theta. \quad (13)$$

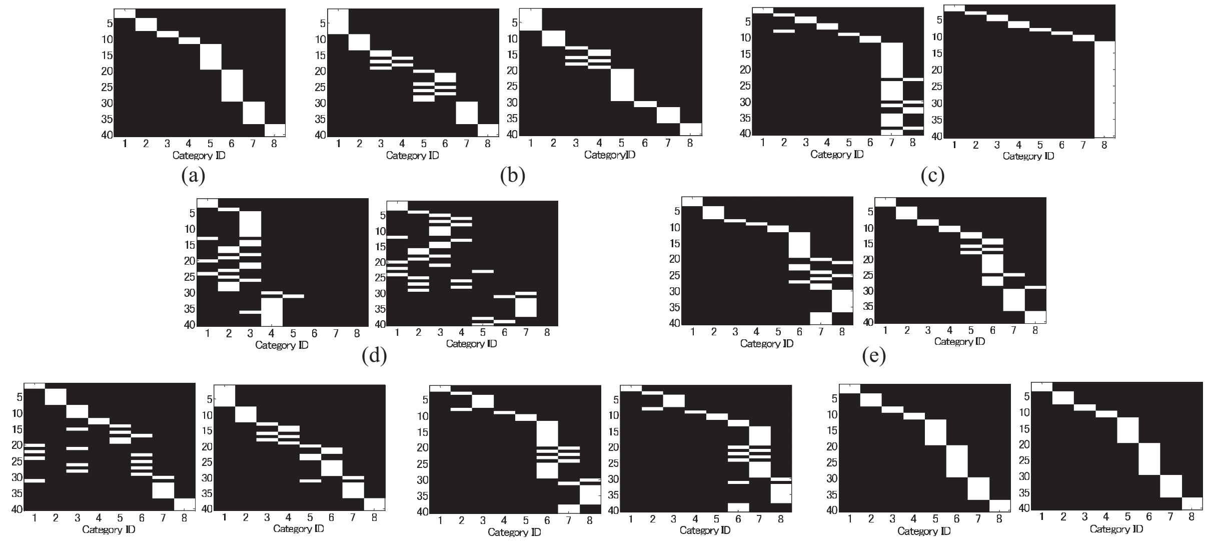
We get $p(\theta|\boldsymbol{w}^v_{obs})$ by recomputing α using variational EM algorithm as before.

C. Degree of connection between word and modality

There are words that represent rather abstract concepts such as 'round', 'soft', 'shape' and so on. Many of these words (e.g. adjectives) are strongly connected to a particular modality. For example, the word 'shape' is connected to visual information. If the robot is aware of the connection between the word and modality, it is possible for the robot to pay attention to the appropriate modality when a certain word is input.

Therefore, the problem here is to measure the degree of connection between words and modalities. In order to do this, we pay attention to the fact that features represented by these words are shared among some relevant categories. For instance, the word 'hard' is connected to haptic modality. Hence similar haptic information would appear in the categories which are connected to 'hard'. On the other hand, audio-visual information is not shared among these categories. For this reason, we propose relevance measure $C_m(w^w_{obs})$, which encodes the degree of connection between the word w^w_{obs} and modality m ($m \in \{audio, visual, haptic\}$), as follows:

$$\begin{aligned} C_m(w^w_{obs}) = \sum_z p(z|w^w_{obs}) \sum_i^{N_m} \min(p(\bar{w}^m_i|w^w_{obs}), \beta^m_{zi}) \\ - \frac{1}{N_m} \sum_i^{N_m} \min(p(\bar{w}^m_i|w^w_{obs}), \beta^m_{zi}), \quad (14) \end{aligned}$$



(a) (b) (c)

(d) (e)

(f) (g) (h)

Fig. 5. The results of categorization.(a)Hand categorization. (b)Visual only categorization. (c)Audio only categorization. (d)Haptic only categorization. (e)Audio-visual categorization. (f)Visual-haptic categorization. (g)Audio-haptic categorization. (h)Categorization using all modalities (audio, visual and haptic). (left:pLSA, right:LDA)

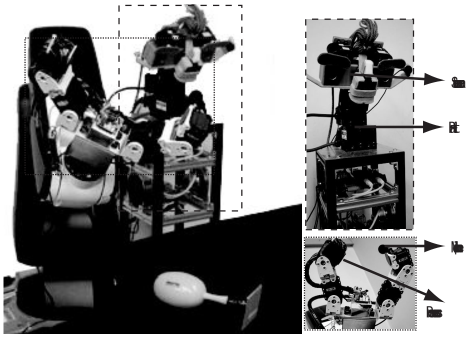
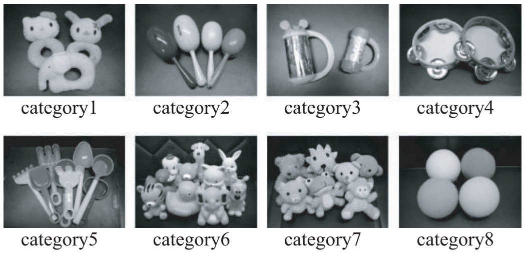


Fig. 3. The robot platform used in the experiments.



category1 category2 category3 category4

category5 category6 category7 category8

Fig. 4. Eight categories consisting of forty toys.

where $\bar{w}_i^m$ and N_m denote i-th component of modality m and number of dimensions of modality m, respectively. $\sum_i \min(a_i, b_i)$ represents a similarity measure between $\boldsymbol{a}$ and $\boldsymbol{b}$, called intersection, ranging from 0 to 1 when $\boldsymbol{a}$ and $\boldsymbol{b}$ are normalized to unity. The more similar $\boldsymbol{a}$ and $\boldsymbol{b}$ are, the closer to unity the output value becomes. C_m can be interpreted as a difference between average of the similarity measure in all categories and weighted average of the similarity considering $p(z|w^w_{obs})$.

IV. Experiment

The proposed algorithm has been implemented on the robot shown in Fig.3. The robot consists of a 6 DOFs arm, a 4 DOFs hand and a 2 DOFs head. There are four pressure sensors on its left fingertip as we mentioned earlier. One microphone is mounted on the right fingertip as well to capture audio information when the robot grasps and shakes an object.

Four experiments are carried out to evaluate the proposed algorithm using the robot. Before the tests, we asked eight subjects to classify the fifty toys according to their own criteria. Although the results differed from person to person, they had 8 categories with 40 objects in common. These are shown in Fig.4. Hence, we have tested the system using those 40 objects.

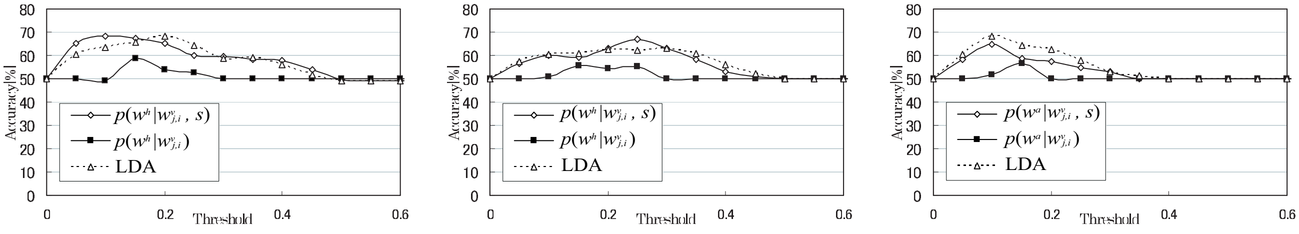
A. Results of categorization

The results of categorization using pLSA and LDA under various conditions are given in Fig.5. In these figures the horizontal and vertical axes indicate category and object indices, respectively. The white bar in the figure represents that the object is classified into the category. In the audio only categorization shown in Fig.5(c), one can see that pLSA generated false category with two different kinds(Fig.5(c) left, category ID 2). On the other hand, LDA successfully categorizes sounder objects without confusing different kinds of objects, although it oversegments the category. Moreover, pLSA classifies non sounder objects (Object11-39) into two classes because of noise, while LDA categorizes non sounder objects as one unified class correctly. The results given in Figs.5 (b) and (f) also indicate that the LDA-based categorization works more robust than pLSA with unreliable information.

In spite of these differences between pLSA and LDA, exactly the same and correct results are given by both pLSA and LDA when all of three modalities are used(Fig.5(h)). This means that three modalities are needed to classify the objects correctly.

B. Category recognition for unseen objects

To evaluate the performance of category recognition for unseen objects, leave-one-out cross validation is carried out using above mentioned 40 objects. All of 40 objects are classified correctly in both cases of pLSA and LDA.



(a)Soft **(b)Hard** **(c)Sounder**

Fig. 6. Results of inference among modalities.

C. Inference of object properties

Here we conduct an experiment of inferring object property only from visual cue. Thirty nine objects, which are chosen from 40 toys, are used to learn the model. The remaining object is put in front of the robot. The robot detects its region in the image and infers properties (i.e. hardness, softness, and sounder or not) from visual features within the region. Such inference is repeated three different locations for each object. Therefore, the total of one hundred twenty inferences are carried out. In the pLSA [9], the inference is carried out by calculating the following a_j^{soft} in each object region:

$$a_j^{soft} = \frac{1}{N_j} \sum_{i,\bar{w}^h} p(\bar{w}^h | w^v_{j,i}, \bar{s}_j) \quad \bar{w}^h \in soft, \tag{15}$$

where $\bar{s}_j$ represents ID of j-th ($j = 0, \cdots, 39$) object. N_j represents the total number of visual features within j-th object region and $w^v_{j,i}$ denotes i-th ($i = 0, \cdots, N_j - 1$) visual feature of j-th object. $\bar{w}^h \in soft$ is defined as co-occurring features among soft objects. In the pLSA, $p(\bar{w}^* | w^v_{j,i}, \bar{s})$ can be replaced by $p(\bar{w}^* | w^v_{j,i})$, which acts as direct inference of auditory and/or haptic properties from visual information. In the LDA, the inference is carried out by calculating the following p_j^{soft} in each object region:

$$p_j^{soft} = \sum_{\bar{w}^h} p(\bar{w}^h | \boldsymbol{w}^v_j) \quad \bar{w}^h \in soft. \tag{16}$$

In the above equation, $p(z|\boldsymbol{w}^v_j)$ is recomputed using visual features $\boldsymbol{w}^v_j$ of j-th object, which clearly indicates that the object category affects the inference.

The inference is carried out by setting a threshold for a_j^* and p_j^*. The accuracy is measured by the arithmetic average of true-positive and true-negative. Therefore the accuracy is 50% for the random selection. Figure 6 shows the relationship between accuracy and threshold. The inferences based on $p(\bar{w}^* | w^v_{j,i}, \bar{s})$ and LDA give better results than that of $p(\bar{w}^* | w^v_{j,i})$. The method based on $p(\bar{w}^* | w^v_{j,i}, \bar{s})$ uses fold-in heuristic to recompute $p(z)$ and $p(s|z)$, which results in the inference through object category. The LDA also uses category information. Hence these results clearly indicate the importance of categorization.

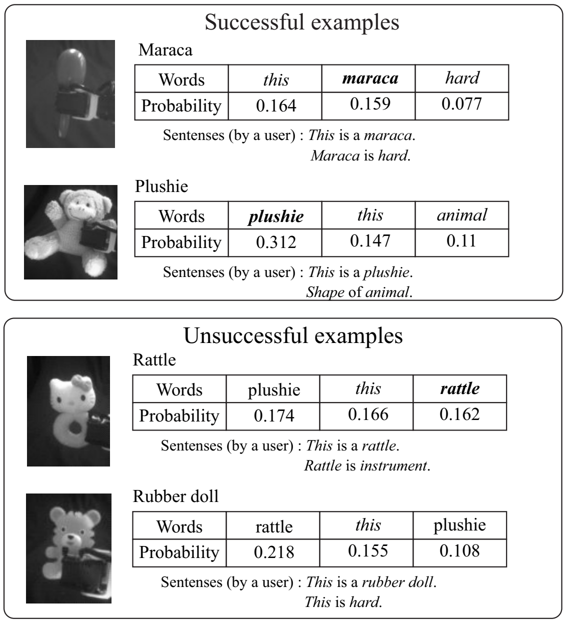
We don't have big difference between LDA and pLSA based on $p(\bar{w}^* | w^v_{j,i}, \bar{s})$. This is because both methods are the inference through object categories. In addition, we consider this is because the scale of experiment (number of objects and categories) is small. We need to verify this point by enlarging the scale of experiment.

D. Inference of words

We carried out an experiment to evaluate the proposed words grounding algorithm using forty objects shown in Fig.4. It should be noted that this experiment is conducted in Japanese. The user was asked to describe each object with a few sentences. The robot recognized these sentences and extracted nouns and adjectives. The word histogram was generated for each object. Then, the learning process was carried out using the method in section III. The words extracted in this experiment are shown in Tab.I.

TABLE I
THE WORDS USED IN THE EXPERIMENT.

this	rattle	sound	soft
instrument	animal	shape	maraca
hard	round	long	tambourine
sandbox	toy	rubber	doll
plushie	ball		



Successful examples

Maraca

Words	*this*	***maraca***	*hard*
Probability	0.164	0.159	0.077

Sentenses (by a user) : *This* is a *maraca*.
Maraca is *hard*.

Plushie

Words	***plushie***	*this*	*animal*
Probability	0.312	0.147	0.11

Sentenses (by a user) : *This* is a *plushie*.
Shape of *animal*.

Unsuccessful examples

Rattle

Words	plushie	*this*	***rattle***
Probability	0.174	0.166	0.162

Sentenses (by a user) : *This* is a *rattle*.
Rattle is *instrument*.

Rubber doll

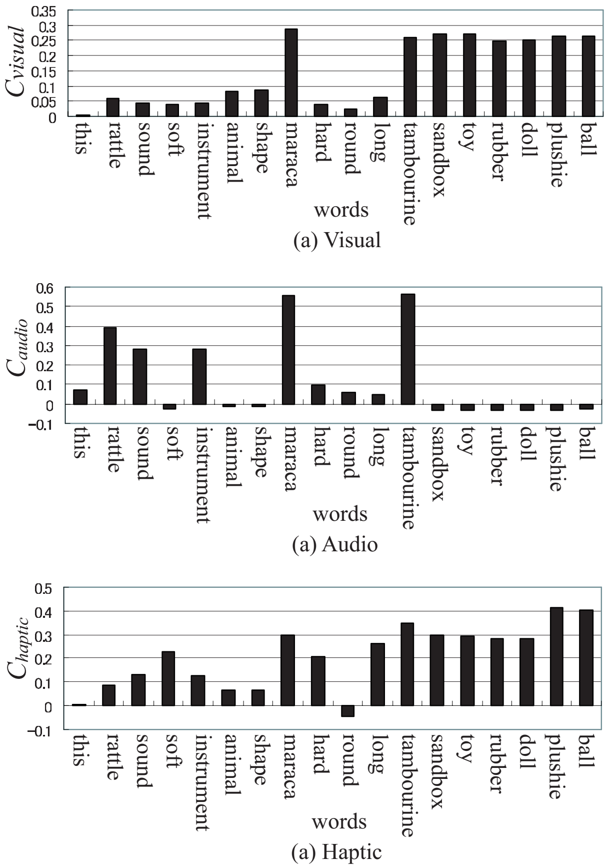
Words	rattle	*this*	plushie
Probability	0.218	0.155	0.108

Sentenses (by a user) : *This* is a *rubber doll*.
This is *hard*.

Fig. 7. Examples of recalled words.

In the test phase, the object was put in front of the robot and three words with three highest $p(w^w | \boldsymbol{w}^v_{obs})$ were inferred from visual information. We calculate the inclusion rate of these three words for the sentences used in the learning process. For forty objects, the inclusion rate was 77.5%, which implies the grounding algorithm worked well. Figure 7 shows some examples of words inference.

Next, the degree of connection between words and modalities C_m is calculated for each word according to Eq.(14). The results are shown in Fig.8. It can be seen that the correspondence is reasonably estimated. For example, the word representing an object category (e.g., rattle, sandbox, etc.) has strong connections with the modalities on which



(a) Visual

(a) Audio

(a) Haptic

Fig. 8. Connection between words and modalities.

U1:The user shows a plushie to the robot.
S1:('plushie', 'this', 'soft', and 'animal' are recalled.)
'This is a plushie.'
U2:'What does this look like?'
S2:'This looks like animal.'
U3:'How hard is this?'
S3:'This is soft.'
U4:'How does this sound?'
S4:'Nothing.'

Fig. 9. An example of the conversation.

the category relies. It is also convincing that $C_{audio}(rattle)$, $C_{audio}(sound)$ and $C_{audio}(instrument)$ are relatively high, since these are deeply related to auditory information. Moreover, $C_{haptic}(hard)$ and $C_{haptic}(soft)$ are higher than those of haptically unrelated words. Meanwhile, $C_{visual}(round)$ and $C_{visual}(long)$ are low even though these words relate to visual information. That is because roundness and/or length of the object cannot be represented by the local visual feature (SIFT in this case).

Finally, we realized a simple conversation between a user and the robot through some objects. The robot has the grounded words and concepts as described above. We set thresholds to each C_m such as $C_{visual} = 0.06$, $C_{audio} = 0.2$ and $C_{haptic} = 0.2$. Thus the robot can judge whether the word is related to the modality m or not according to $C_m('word')$. The robot responds to some interrogations and to pay attention to a certain modality according to keywords (e.g. 'look like'). Then the robot can answer the question using a sentence template and recalled words.

Figure 9 shows an actual example of the conversation. In the example, the robot saw a plushie at first and recalled some words('plushie', 'this', 'soft' and 'animal'). Since the word 'plushie' has very large C_{visual} and C_{haptic}, the robot inferred that the name of the object is 'plushie'. When the user asks a question regarding its appearance, the robot is designed to answer the recalled word which is related only to visual information. In the example, the robot answers 'This looks like animal.', since 'animal' is the word which is related only to visual information.

Although the current conversation system is simple enough, the robot is proven to be able to use the perceptually grounded words using the proposed framework.

V. CONCLUSION

In this paper multimodal object categorization has been explored. Then the symbol grounding problem has been examined using the concepts formed by the multimodal categorization. The proposed framework is an extension of LDA. Experimental results with 40 objects (8 categories) show that the proposed algorithm works better than the visual only categorization. We also demonstrated a possibility of the conversation between a user and the robot based on the grounded language. Now we are planning to expand the experimental scale (i.e. categories, objects, users etc.) to evaluate the proposed framework under severe environmental conditions.

VI. ACKNOWLEDGMENTS

This work is supported in part by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (C) 20500179 and the National Institute of Informatics joint study program.

REFERENCES

[1] F.G.Ashby, W.T.Maddox, "Human Category Learning", *Annu. Rev. Psychology*, 56, pp.149-178, Feb.2005

[2] D.J.Freedman, J.A.Assad, "Experience-dependent Representation of Visual Categories in Pariental Cortex", *Nature*, 443(7), pp.85-88, 2006

[3] P.Bloom, DESCARTES' BABY: How the Science of Child Development Explains What Makes Us Human, *Basic Books*, 2004

[4] J.Sivic, B.C.Russell, A.A.Efros, A.Zisserman and W.T.Freeman, "Discovering Object Categories in Image Collections", *AI Memo*, 2005-005, pp.1-12, Feb.2005

[5] R.Fergus, P.Perona and A.Zisserman, "Object Class Recognition by Unsupervised Scale-invariant Learning", *in Proc. of CVPR2003*, Vol.2, pp.264-271, June2003

[6] R.Fergus, P.Perona and A.Zisserman, " Using Multiple Segmentations to Discover Objects and Their Extent in Image Collections", *in Proc. of CVPR2006*, Vol.2, pp.1605-1614, June2006

[7] L.Fei-Fei and P.Perona, "A Bayesian Hierarchical Model for Learning Natural Scene Categories", *in Proc. of CVPR2005*, Vol.2, pp.524- 531, June2005

[8] L.Fei-Fei, R.Fergus, P.Perona, "One-Shot Learning of Object Categories", *IEEE Trans. on PAMI*, Vol.28, No.4, pp.594-611, Apr.2006

[9] T.Nakamura, T.Nagai, N.Iwahashi, "Multimodal object categorization by a robot", IROS2007, pp.2415-2420, Oct.2007

[10] D.Blei, A.Y.Ng, M.Jordan, "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, pp.993-1022, Jan.2003

[11] D.Roy, A.Pentland, "Learning Words from Sights and Sounds: A Computational Model", *Cognitive Science*, Vol.26, No.1, pp.113-146, 2002

[12] N.Iwahashi, "Robots That Learn Language: A Developmental Approach to Situated Human-Robot Conversations", *In N.Sankar ed. Human-Robot Interaction*, pp.95-118, I-Tech Education and Publishing, 2007

[13] C.Yu, D.Ballard, "On the Integration of Grounding Language and Learning Objects", *in Proc. of 19th National Conference on Artificial Intelligence(AAAI)*, pp.488-494, Jul.2004

[14] D.G.Lowe, "Distinctive image features from scale-invariant keypoints", *Int. Journal of Computer Vision*, 60(2), pp. 91-110, Nov.2004

[15] M.Jordan, Z.Ghahramani, T.Jaakkola, L.Saul, "Introduction to variational methods for graphical models", *Machine Learning*, 37, pp. 183-233, Nov.1999