Online Object Categorization Using Multimodal Information Autonomously Acquired by a Mobile Robot

Takaya Araki[1,*], Tomoaki Nakamura[1], Takayuki Nagai[1],
Kotaro Funakoshi[2], Mikio Nakano[2] and Naoto Iwahashi[3]

[1] *Department of Mechanical Engineering and Intelligent Systems, The University of Electro-Communications, 1-5-1 Chofugaoka Chofu-shi, Tokyo 182-8585, Japan*

[2] *Honda Research Institute Japan Co., Ltd., 8-1 Honcho, Wako-shi, Saitama 351-0188, Japan*

[3] *NICT Knowledge Creating Communication Research Center, 2-2-2 Hikaridai, Seika-cho Souraku-gun, Kyoto 619-0288, Japan*

{*taraki, naka t, tnagai*}*@apple.ee.uec.ac.jp*, {*funakoshi, nakano*}*@jp.honda-ri.com, naoto.iwahashi@atr.jp*

Abstract

In this paper, we propose a robot that acquires multimodal information, i.e., visual, auditory, and haptic information, fully autonomously using its embodiment. We also propose batch and online algorithms for multimodal categorization based on the acquired multimodal information and partial words given by human users. To obtain multimodal information, the robot detects an object on a flat surface. Then, the robot grasps and shakes it to obtain haptic and auditory information. For obtaining visual information, the robot uses a small hand-held observation table with an XBee wireless controller to control the viewpoints for observing the object. In this paper, for multimodal concept formation, multimodal latent Dirichlet allocation (LDA) using Gibbs sampling is extended to an online version. This framework makes it possible for the robot to learn object concepts naturally in everyday operation in conjunction with a small amount of linguistic information from human users. The proposed algorithms are implemented on a real robot and tested using real everyday objects to show the validity of the proposed system.

keywords: Multimodal categorization, Concept formation, Online learning, Autonomous acquisition

1 Introduction

In recent years, a great deal of research has been conducted on intelligent robots that coexist with people on a daily basis. However, many problems still remain unsolved. One of the most important problems is the management of objects by the robot. Since vast numbers of objects exist in the real world and the objects that the robot should deal with are unknowable in advance, it is impractical to register all object information beforehand. This fact shows that the online learning of object concepts, which consist of generalized perceptual information and associating linguistic labels, is indispensable to

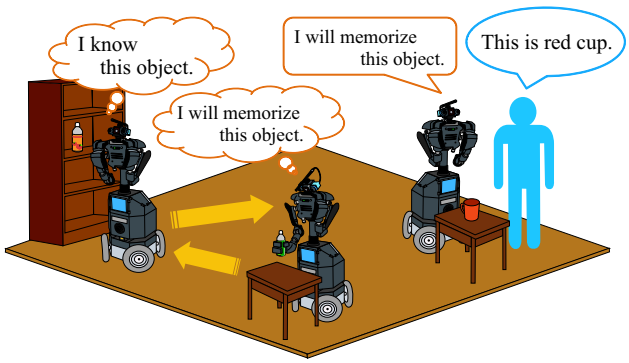

Figure 1: Autonomous learning scenario of object concepts by a robot.

the robot in a real environment. At this time, the multimodal perceptual information that the robot obtains is of importance for learning. Moreover, it is desirable for the robot to observe the object and obtain multimodal information in a fully autonomous fashion without any human help so that the robot develops object concepts by itself.

Unsupervised learning of objects using a huge number of images has been extensively studied previously [1]–[3]. Research on object categorization using sounds that occur when tapping the object [4][5] or haptic information when grasping an object [6][7] has been conducted. However, it is obvious that object category does not depend only on a single piece of information but also on various inputs. We have proposed a framework for object concept formation based on multimodal categorization by robots using statistical models such as probabilistic latent semantic analysis (pLSA) and LDA [8][9]. We showed that multimodal categorization makes it possible for the robot to categorize objects in the same manner as humans do [10]. This means that suitable object concepts can be formed through multimodal categorization, and such concepts are useful for predicting unobservable properties of unseen objects for the robot. We strongly believe that this kind of ability establishes a basis of "true understanding" and is a very important factor for human–robot coexistence.

To achieve this type of learning in real environments autonomously, the robot has to obtain multimodal data such as visual, auditory, and haptic information by itself. Hence, an autonomous multimodal information acquisition mechanism is required for the robot. Nonetheless, few researchers have addressed such systems in the past. For visual information, autonomous acquisition of object representations has been implemented on some humanoid robot platforms [11]–[14]. For example, in [11], the authors propose a method for obtaining multiview object representations by handing over the target object to the robot. The authors of Ref. [12] propose a robot that can grasp and obtain visual information of objects autonomously. In [15], the robotic systems can autonomously acquire three-dimensional (3D) information of an object by going around it. We have proposed a method in which a human user shows a target object to the robot for learning novel objects in [16]. It is worth noting that in none of these works the acquisition of multimodal information is considered.

The goal of this paper is to develop a robot that learns object concepts online by autonomously obtaining multimodal information on a daily basis in any way as illustrated in Fig. 1.

Linguistic labels are also important for object concept formation. In [9], we have shown that the robot can learn meanings of words by connecting multimodal concepts, which are formed by multimodal categorization, and corresponding words. Here, we take a step further to consider the learning process of entire object concepts, including word meanings at once using both words and multimodal perceptual information. The word information must carry useful cues for human-like categorization. This fact motivates us to include linguistic information for our multimodal categorization. Of course, the formed object concepts can be used for inferring suitable words for unseen objects. It should be noted that the word information must be given by a human user. However, it is not practical for a human user to always accompany the robot to provide linguistic information. As mentioned earlier, the learning process should be as autonomous as possible. The robot expects to have linguistic information only when a human user is available. Therefore, the robot is required to have the ability to form object concepts using perceptual information and partially given (incomplete) words.

In this paper, a robot that acquires multimodal information in a fully autonomous way using mounted sensors is proposed. The robot can acquire visual information from a 3D visual sensor [17], auditory information by shaking the object, and haptic information by grasping it. We also propose an online algorithm of multimodal categorization based on autonomously acquired multimodal information and partial words that are given by the human user. For this purpose, we first discuss a batch-type learning algorithm based on multimodal LDA using Gibbs sampling. Then, the proposed batch algorithm is extended to an online version so that the robot can discard the data after using them for learning. This needs to be done because the perceptual information consumes a large amount of memory and batch-type learning is inefficient.

In [19]–[21], some online algorithms for LDA have been proposed. An online variational Bayes (VB) algorithm for LDA is shown to converge to a local optimum of the VB objective function in [19]. We extend Gibbs-sampling-based LDA instead of VB-based LDA because Gibbs sampling yields better results in categorization [22]. Moreover, Gibbs-sampling-based LDA has an easy-to-implement property. An online LDA using a Gibbs sampler called o-LDA has been proposed in [20]; a batch initialization phase is required for o-LDA. In [21], the incremental Gibbs sampler has been proposed to improve the performance of o-LDA. The basic idea is to rejuvenate old assignments, which implies old data cannot be discarded.

2 Overview of the object concept formation

An overview of the proposed concept formation by a robot is shown in Fig. 2. The robot forms object concepts based on both perceptual information, which is acquired autonomously, and partially given linguistic information. The robot is assumed to have both a lexicon and grammatical skills, which allow the robot to decompose recognized sentences into words. Here nouns, adjectives, and adjective verb are considered. These are extracted from recognized sentences by using a morphological analysis. All perceptual information is vector quantized and represented as a "bag-of-features" model.

The object concepts are represented by a multimodal LDA (MLDA), as illustrated in Fig. 2. In the

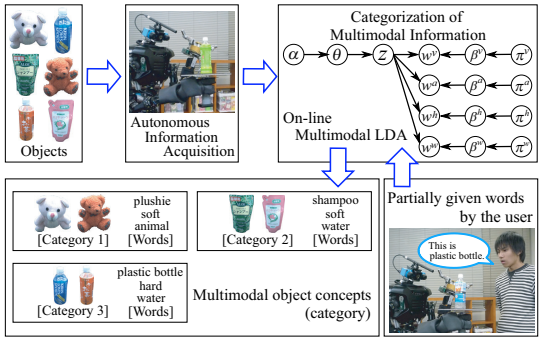

Figure 2: Overview of proposed system.

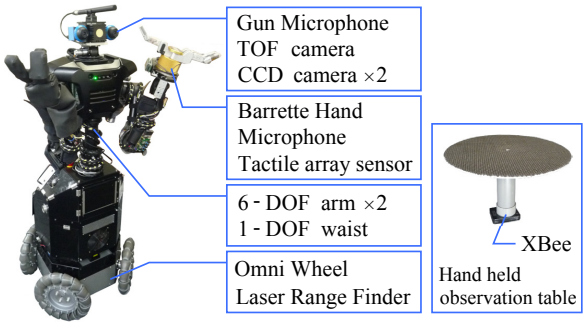

Figure 3: Robot platform and hand-held observation table.

figure,w^v, w^a, w^h, and w^w represent visual, auditory, haptic, and word information and are assumed to be drawn from each multinomial distribution parameterized by β^v, β^a, β^h, and β^w, respectively. π^v, π^a, π^h, and π^w denote parameters of Dirichlet prior distributions for β^*. z represents the category and is assumed to be drawn from a multinomial distribution parameterized by θ, which depends on the Dirichlet prior distribution parameterized by α.

The model makes it possible for the robot not only to recognize categories of unseen objects but also to infer unobserved properties of the object and words that are suitable for describing it. Understanding the meaning of words is also possible through the MLDA model. In this paper, Gibbs-sampling-based LDA [22] is extended to multimodal LDA. In addition, online MLDA is also proposed to enable incremental learning. The important point of the proposed system is that the robot can learn object concepts incrementally and autonomously in conjunction with partially given word information. In the next section, the autonomous acquisition of multimodal information is described.

3 Multimodal information acquisition and signal processing

3.1 Autonomous acquisition of multimodal information

Figure 3 shows the robot platform used in this paper. The robot is equipped with a 3D visual sensor [17], arms capable of six degrees of freedom (DOFs), and 4-DOF hands with a tactile array sensor consisting of 162 elements and a microphone. Moreover, omnidirectional wheels and a laser range finder (LRF) enable the robot to move freely in the living room. These sensors are used for obtaining multimodal information: color images from multiple viewpoints, 3D information from the time of flight (TOF) camera, near-infrared (NIR) reflectance intensities, pressure information from the tactile sensors, and sound that is made by shaking the object. In addition, word information obtained from the user's utterances can also be acquired while capturing multimodal information.

To capture such information autonomously, three problems have to be solved: (1) novel object detection, (2) grasping of novel objects, and (3) a method for observing perceptual information. In this paper, the first problem is simply resolved by plane-based object detection as in [17] and [23]. After object detection, object recognition is carried out to check whether the object is known or unknown. If the object is novel to the robot, the object observing action is activated. Three-dimensional information

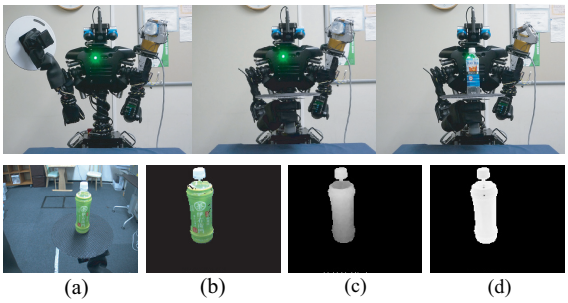

Figure 4: Acquisition of visual information and an example of acquired visual information: (a) color, (b) color-mapped 3D images, (c) depth, and (d) NIR intensity.

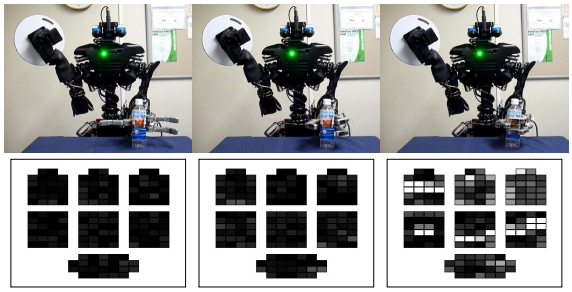

Figure 5: Acquisition of haptic information and tactile array sensor output.

can be a key to solve the second problem; that is, the pose and grasping points of the object can be computed based on the observed 3D point clouds [24].

The most important issue confronting the third problem is acquisition of visual information. Since the object should be observed from various viewpoints, the object has to be held in the hand of the robot during the acquisition process. This situation may cause a deformity and/or severe occlusion from its own fingers. To cope with this problem a small hand-held observation table, as shown in Fig. 3, is introduced. The robot grasps the target object and places it on the observation table that is held in the other hand. The paths of the arms can be planed using dual-arm rapidly-exploring random tree-connect (RRT) [25] by considering the shape of the object, which is grasped by the hand. The observation table with an XBee wireless controller enables the robot to circle the table freely to capture information on an object from various viewpoints. Finally, the color, texture, 3D shape, and NIR intensities of the object can be obtained from each view. In later experiments, 36 images are collected from each object. Scenes of the visual information acquisition process are shown in Fig. 4, the lower part of which shows an example of visual information obtained by the actual robot. Haptic information is acquired by performing the grasping actions in Fig. 5 five times. During the grasping action, the fingers are controlled to close at a constant velocity and stop when the torque reaches a predefined threshold value. The lower part of Fig. 5 shows actual sensor values of the tactile array sensor.

For auditory information, a microphone mounted on the robot hand is used to capture the sound produced when the robot shakes the object. A problem here is the motor noise of the robot arm, since the arm moves relatively fast, so the object must make an audible enough sound. Here, any frame that has energy less than a predetermined threshold is discarded since it can be considered as motor noise. Figure 6 shows a sequence of actions for obtaining auditory information.

3.2 Signal processing for multimodal categorization

In this section, signal processing methods for information obtained by the robot autonomously are described.

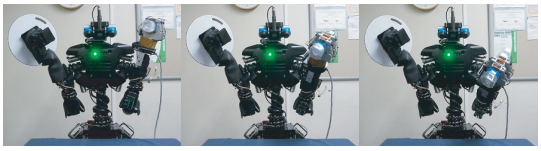

Figure 6: Acquisition of auditory information.

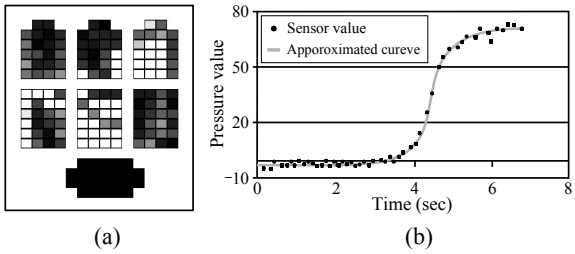

Figure 7: Example of tactile array sensor output: (a) output (6 s) and (b) values of the sensor and the approximated curve.

3.2.1 Visual information

The target object is segmented out in each image frame, and then, 128-dimensional DSIFT descriptors [26] are computed. In a later experiment, 36 image frames are captured for each object. Three hundred to 400 feature vectors are extracted at each image, resulting in about 10000–15000 features for each object. Each feature vector is vector quantized using a codebook with 500 clusters. The codebook is generated by a k-means algorithm in advance. Finally, a 500-dimensional histogram is built as the bag-of-features representation.

3.2.2 Auditory information

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

3.2.3 Haptic information

Haptic information is obtained from the three-finger robotic hand with a tactile array sensor. A total of 162 time series of sensor values are obtained by grasping an object. Each time series is approximated as

$$p(t) = a \tan^{-1}(b(t+c)) + d, \tag{1}$$

where $p(t)$ represents the approximated sensor value at time t and a, b, c, and d are parameters for the time series. Figure 7 shows an example of the approximated time series of the tactile sensor values. The parameters (a, b, c, d) encode tactile information of the object as follows: a represents the pressure level of a sensor, b is considered as time elapsing from the moment of contact to the end of the finger movement, c depends on the size of the object, and d is proportional to a. Since two parameters a and b are related to the hardness of the object, they are used as the feature vector of each sensor. Hence, a total of 162 feature vectors are obtained by grasping an object. Again, the bag-of-features model is applied to the data, so that any variation resulting from changes in the grasping point can be absorbed. The feature vectors are vector quantized using a codebook with 15 clusters and the histogram is constructed.

3.2.4 Word information

In this paper, nouns, adjectives, and adjective verb are considered as word information. These words are extracted by using a morphological analysis from partial sentences given by human users. Here, the word information is treated as a bag of words. In the experiments shown later, 145 words were used by the users to describe the objects. Hence, a 145-dimensional histogram is calculated.

4 Multimodal categorization

4.1 MLDA using Gibbs sampling

The problem of categorization is equivalent to the estimation of parameters of the graphical model in Fig. 2, using multimodal information observed by the robot. Gibbs sampling is used for the parameter estimation because no approximation is incorporated and it is easy to implement. In this paper, the LDA algorithm using Gibbs sampling [22] is extended to the multimodal version.

Now, let $\boldsymbol{w}^m$ be a set of captured multimodal information. In Gibbs sampling, the category z_{mij}, which is assigned to the ith datum of modality $m \in \{visual, auditory, tactile, words\}$ of the jth object, is sampled from the following conditional probability:

$$P\left(z_{mij} = k | \boldsymbol{z}^{-mij}, \boldsymbol{w}^m, \alpha, \pi^m\right) \propto (N_{kj}^{-mij} + \alpha) \cdot \frac{N_{mw^m k}^{-mij} + \pi^m}{N_{mk}^{-mij} + W^m \pi^m}, \tag{2}$$

where W^m denotes the dimension of modality m. $N_{mw^m kj}$ represents a frequency count of assigning w^m to the category k for the modality m of the jth object. Here, N_{kj} and N_{mk} can be calculated as follows:

$$N_{kj} = \sum_{m, w^m} N_{mw^m kj}, \tag{3}$$

$$N_{mk} = \sum_{w^m, j} N_{mw^m kj}. \tag{4}$$

N_{kj} represents number of times of assigning all modalities of the jth object to the category k, and N_{mk} represents the frequency of assigning modality m of all objects to the category k. The superscript with the minus sign in Eq. (2) denotes an exception, e.g., $\boldsymbol{z}^{-mij}$ represents assigned categories except for z_{mij}.

The category assigned to the ith datum of the modality m of the jth object is sampled according to Eq. (2). This process is repeated until N_* converges to a certain value. After the convergence, the final estimates of parameters $\beta_{w^m k}^m$ and θ_{kj} can be written as follows:

$$\beta_{w^m k}^m = \frac{N_{mw^m k} + \pi^m}{N_{mk} + W^m \pi^m}, \tag{5}$$

$$\theta_{kj} = \frac{N_{kj} + \alpha}{N_j + K\alpha}, \tag{6}$$

where K represents the number of categories.

4.2 Online MLDA

In standard batch Gibbs sampling MLDA, parameters are estimated by iterating the sampling according to Eq. (2) for all objects. The batch algorithm relies on the assumption that the system holds all multimodal data. Hence, a large amount of memory can be consumed as the number of training objects increases. Furthermore, the batch algorithm is inefficient since Gibbs sampling must be iterated for all object data every time a new datum arrives at the system. This may take a long time and be impractical, especially for an interactive learning scenario where a human user is involved. To solve this problem, we propose an online MLDA that sequentially updates parameters using new input data. After the update of parameters, the input multimodal data can be discarded in the online MLDA. The straightforward extension of MLDA to an online version is to take the idea of o-LDA, which uses current parameters as initial values for updating the model.

If we concentrate only on newly input information of an object, Eqs. (2), (3) and (4) can be written as follows:

$$P\left(z_{mi}=k|\boldsymbol{z}^{-mi},\boldsymbol{w}^{m},\alpha,\pi^{m}\right)\propto(N_{kj}^{-mi}+\alpha)\cdot\frac{N_{mw^{m}k}^{-mi}+\pi^{m}}{N_{mk}^{-mi}+W^{m}\pi^{m}}, \tag{7}$$

$$N_{k} = \sum_{m,w^{m}}N_{mw^{m}k}, \tag{8}$$

$$N_{mk} = \sum_{w^{m}}N_{mw^{m}k}, \tag{9}$$

where $N_{mw^{m}k}$ represents a frequency count of assigning w^{m} to the category k for the modality m of the target object.

In the same way, by repeating the results until convergence according to Eq. (7), the parameters $\beta_{w^{m}k}^{\hat{m}}$ and $\hat{\theta_{k}}$ can be determined by the following equations:

$$\beta_{w^{m}k}^{\hat{m}} = \frac{\hat{N}_{mw^{m}k}+\pi^{m}}{\hat{N}_{mk}+W^{m}\pi^{m}}, \tag{10}$$

$$\hat{\theta_{k}} = \frac{\hat{N}_{k}+\alpha}{\sum_{k}\hat{N}_{k}+K\alpha}. \tag{11}$$

Here, $\hat{N}_{*}$ represents a converged value of N_{*}.

As might be expected, the above idea has a problem that the order of input data seriously affects the performance of the trained model. In fact, o-LDA introduces a batch initialization phase to avoid this problem [20]. Here, we introduce the forgetting factor λ $(0<\lambda<1)$ as follows:

$$N_{mw^{m}k}^{(j+1)}=(1-\lambda)\hat{N}_{mw^{m}k}^{(j)}, \tag{12}$$

where $\hat{N}_{mw^{m}k}^{(j)}$ represents the converged value of $N_{mw^{m}k}$ for the jth object data. When a new object $j+1$ is input, $\hat{N}_{mw^{m}k}^{(j)}$ is used as the initial value of the Gibbs sampling by multiplying it by the factor $1-\lambda$. Here, the effects on the learning of a model caused by object order and/or the initial value can be reduced by the forgetting parameter. Gibbs sampling is carried out by applying Eq. (7) to the new input data iteratively until convergence. It is worth noting that $\lambda=0$ corresponds to o-LDA in [20].

Moreover, using the learned model, the category of the unseen object can be inferred. For given modal information of the novel object, $\boldsymbol{w}_{obs}^m$, its category can be determined as z that maximizes $P(z|\boldsymbol{w}_{obs}^m)$. Hence, category $\hat{z}$ of the novel object can be inferred as

$$\begin{aligned}\hat{z} &= \underset{z}{\operatorname{argmax}}\, P(z|\boldsymbol{w}_{obs}^m) \\ &= \underset{z}{\operatorname{argmax}} \int P(z|\theta)P(\theta|\boldsymbol{w}_{obs}^m)d\theta. \end{aligned} \tag{13}$$

It is also possible to recollect suitable words w^w for the unknown object. For this purpose, $P(w^w|\boldsymbol{w}_{obs}^m)$ is computed for given $\boldsymbol{w}_{obs}^m$ as

$$P(w^w|\boldsymbol{w}_{obs}^m) = \int \sum_z P(w^w|z)P(z|\theta)P(\theta|\boldsymbol{w}_{obs}^m)d\theta. \tag{14}$$

It should be noted that $P(z|\theta)$ and $P(\theta|\boldsymbol{w}_{obs}^m)$ in Eqs. (13) and (14) can be updated by recalculating θ for fixed β^m using Gibbs sampling.

4.3 Model selection based on a particle filter

The algorithm described in Section 4.2 enables incremental learning by sequentially updating the model using only the new input data. However, even if we use the forgetting parameters, dependency on order of the input data and/or the initial values still remains. Moreover, the forgetting factor λ should be set to the appropriate value. For instance, if the value of λ is too large, learning progress would be prohibited. To solve this problem, a model selection method using a particle filter is introduced in this paper.

Since words are given by the user as the ground truth, the model, which predicts correct words using the perceptual (visual, auditory, and haptic) information observed by the robot, can be considered as a model representing human-like concepts. That is, the model that gives a large value in Eq. (14) must fit better to the human sense. In the proposed method, the hyperparameter α, which determines the probability distribution of category z, as shown in Fig. 2, and the forgetting factor λ are varied as particles. Models with varying parameters are generated from ten to hundreds as particles. Then, the method for sequentially selecting the model that maximizes Eq. (14) is introduced. The model parameters with high accuracy in word prediction are selected and are used to replace the models with low accuracy as the initial values at the next learning step. Therefore, online learning progresses by varying the initial values and the forgetting factor at each step. When data for a new object are input, models with different parameters α and forgetting factors are created, and then, the model-updating process using the particle filter, which is based on the word prediction accuracy (Eq. (14)), is performed. Eventually, learning is achieved by performing sequential sampling according to Eq. (7). Since it becomes possible to select the appropriate initial values and forgetting parameters sequentially, dependency on initial values and learning order can be reduced. Algorithm 1 outlines the proposed online MLDA using the model selection based on word information for a single object. Every time the robot finds a novel object, the algorithm is applied to the new input data.

Algorithm 1 Online MLDA (for a single object)

Initialize λ and α
for all m, w^m, k **do**
  $N_{mw^mk} \leftarrow (1-\lambda)N_{mw^mk}$
end for
The following process is repeated until convergence
for all m, i (of new input data) **do**
  $u \leftarrow$ random value $[0, 1]$
  for $k \leftarrow 1$ to K **do**
    $P[k] \leftarrow P[k-1] + (N_k^{-mi} + \alpha)\frac{N_{mw^mk}^{-mi} + \pi^m}{N_{mk}^{-mi} + W^m\pi^m}$
  end for
  for $k \leftarrow 1$ to K **do**
    if $u < P[k]/P[K]$ **then**
      $z_{mi} = k$, break
    end if
  end for
end for
Select the model based on $P(w^w|\boldsymbol{w}_{obs}^v, \boldsymbol{w}_{obs}^a, \boldsymbol{w}_{obs}^h)$

5 Experiments

Five experiments are carried out to evaluate the proposed method. Figure 8 shows 50 objects used in the experiments. The objects framed by the red line are used as the objects for recognition in later experiments. During the observation of each object, volunteers gave a description for each object in turn. These words were used as linguistic information for the objects.

5.1 Human categorization

To evaluate the robot's classification, it is necessary to determine the classification ground truth. Therefore, experiments to classify the objects are performed by nine subjects. The subjects were asked to classify the objects shown in Fig. 8 into 8 to 14 classes according to their own criteria. In this paper, *Concordance*, which represents the degree of similarities among the subjects, is defined as follows:

$$Concordance = \frac{1}{J}\sum_{j}^{J}\delta(c_1(j), c_2(j)), \tag{15}$$

where J is the number of objects, $c_1(j)$ and $c_2(j)$ denote the category ID of the jth object categorized by two subjects, respectively, and $\delta(a, b)$ is a function that results in 1 for $a = b$ and 0 otherwise. The *Concordance* of all subjects, which has been calculated from Eq. (15), is shown in Fig. 9(a). In Fig. 9(a), the horizontal axis is the number of categories and the vertical axis is the average *Concordance* for all subjects; error bars represent the standard deviation.

The results reveal that the highest *Concordance* is achieved when the number of categories is 11. This means that most of the categorization made by the subjects is as shown in Fig. 8. In addition, when the number of categories was 11, the standard deviation is smaller than in other cases; i.e., there

Figure 8: The 50 objects used in the experiments. The red frame indicates objects used in recognition, which will be explained later.

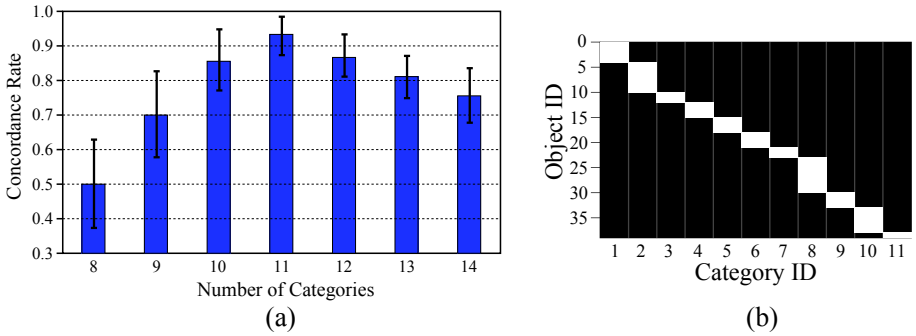

Figure 9: Human categorization: (a) concordance of nine subjects and (b) result of categorization (ground truth).

is not much difference between the classification of the subjects. If the number of categories is less than 10, the problem of combining two categories arises, such as the case when combining categories 2 and 3, categories 7 and 8, or categories 10 and 11 into the same category. Since the criteria for combining categories for each case is different and subject dependent, the *Concordance* will decrease as the number of categories decreases. However, if the number of categories is greater than or equal to 12, the *Concordance* is decreased, but by a smaller amount than when the number of categories is 10 or less. This is because the categories were formed from minor objects when 11 categories was taken as a standard.

Hence, the number of categories is chosen to be $K = 11$, and the most common categorization result (Fig. 8) is used as the ground truth to evaluate categorization by the robot. Moreover, Fig. 9(b) shows the correct classification result of each object ID. In this figure, the horizontal and vertical axes represent the category and object indices, respectively. The white bar in the figure indicates that the object is classified into the category.

5.2 Autonomous acquisition of multimodal information

The proposed multimodal information acquisition system has been implemented on the robot platform. All objects in Fig. 8 were placed on a table, and each description was decomposed into words using Japanese morphological analysis.

Figure 10: Examples of acquired multimodal information: (from top to bottom) color images, color mapped images, depth images, NIR intensity images, tactile array sensor output (grasped enough), tactile histograms, and auditory histograms (only 10 distinctive dimensions shown).

Table 1: Examples of words used for describing the objects.

tea	cup	can	red	pig	box	bear	elephant
frog	soap	hand	pink	lion	gray	hard	dressing
soft	food	blue	long	yarn	bell	snack	flooring
chick	glass	juice	chips	spray	paper	drink	shampoo
sound	metal	light	black	color	water	thick	biscuit
vinyl	brown	white	green	cookie	noodle	yellow	sports drink
square	purple	animal	fabric	monkey	plushie	cleaner	plastic bottle

The robot succeeded in autonomously acquiring multimodal information of all objects using the proposed system. Some examples of the acquired multimodal information are given in Fig. 10. It should be noted that only 10 distinctive dimensions out of 50 are shown for the auditory histograms in the figure. The word information used in the experiments consists of 145 words. Table 1 shows examples of the words. This multimodal information and these words are used in the following experiments.

5.3 Multimodal categorization

Multimodal categorizations have been carried out using batch VB [9], batch Gibbs sampling, and the proposed online Gibbs sampling. The robot categorized the multimodal data of 50 objects acquired in the previous section. The number of given words was varied to confirm the contribution of linguistic information to the categorization.

The object categorization results obtained by using the proposed online learning is shown in Fig. 11. In the figure, the x axis, y axis, and z axis are, respectively, the number of given words, the number of

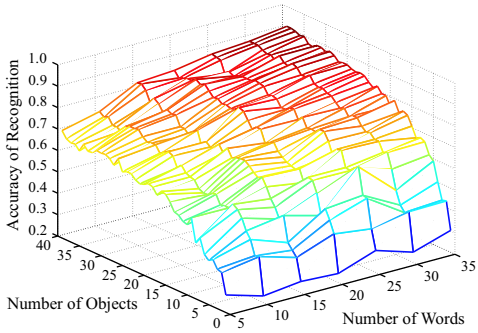

Figure 11: Accuracy of category recognition.

trained objects, and categorization accuracy defined as

$$Acc = \frac{(\text{number of objects categorized correctly})}{(\text{number of all objects})}. \tag{16}$$

Each accuracy value was calculated as an average over 100 trials. From the figure, one can confirm that the proposed online algorithm successfully learned objects incrementally. The object classification accuracy increases as the number of learning objects and number of words increase. Even if the word information is given only on five objects, the classification accuracy was about 70%. Finally, a high classification accuracy of 90.8% was achieved when word information is given on 35 objects. Examples of the categorization results are shown in Fig. 12. For comparison, the classification results by the previous method, i.e., VB batch and GS batch, are also shown. In these figures, as in Fig. 9(b), the horizontal axis shows the category ID and the vertical axis represents the index of the object. VB batch, GS batch, and GS online are shown from top to bottom. The number of objects with a given word information are shown increasing from left to right, i.e., 5, 20 (half of all objects), and 39 (all objects). From the figures it can be seen that Gibbs-sampling-based LDA yields better results than those of VB-based LDA. One can also see that better results are obtained by using word information in all methods. In particular, if the word information is sparse, plastic bottles (Object ID 5–10) and shampoo (Object ID 16–18) tend to be classified as the same category owing to the similar sound information. This tendency also occurs for yarn (Object ID 31–33) and plushie (Object ID 34–38) owing to their similar haptic information. When word information for 20 objects is provided, the categorization accuracy improves in all methods. Here, the proposed GS online learning is able to exceed the accuracy of GS batch learning. Moreover, almost all objects were categorized correctly when words were given for all objects in Gibbs sampling LDA. With respect to the results of the proposed online Gibbs sampling LDA, it can be seen that better categorization results are obtained by the proposed online LDA compared with the batch VB LDA. It is also revealed that comparable results are obtained by the online and batch LDAs.

To learn the geometrical structure of learned concepts, we visualize the models along with time. In the model of Fig. 2, parameter θ represents the probability ratio of a category for each object. The differences among θ values corresponding to the objects can be considered as the distances among concepts of these objects. Thus, the differences among θ values of the objects that belong to the same category are small and vice versa. In this paper, the Euclidean distance between θ values is used to

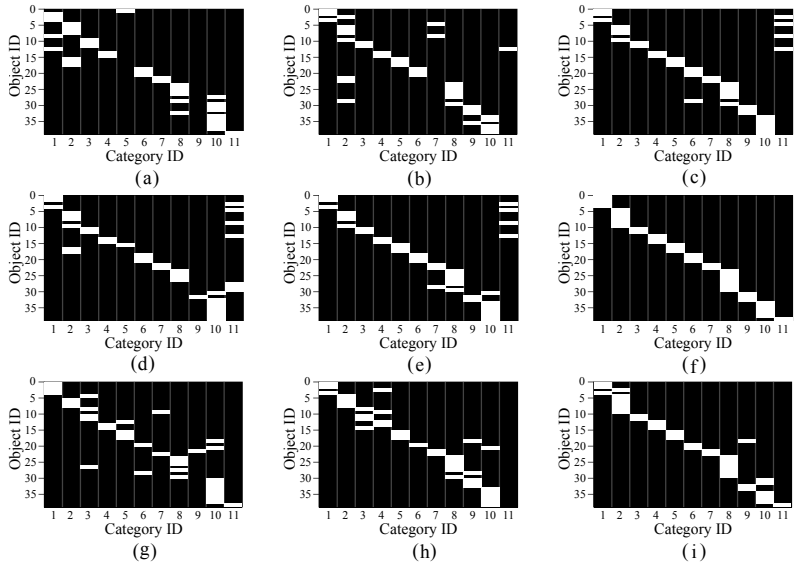

Figure 12: Result of categorization: (a) VB batch (with 5 words of information) categorization accuracy Acc is 0.64, (b) VB batch (with 20 words of information) categorization accuracy Acc is 0.76, (c) VB batch (with 39 words of information) categorization accuracy Acc is 0.84, (d) GS batch (with 5 words of information) categorization accuracy Acc is 0.76, (e) GS batch (with 20 words of information) categorization accuracy Acc is 0.82, (f) GS batch (with 39 words of information) categorization accuracy Acc is 1.00, (g) GS online (with 5 words of information) categorization accuracy Acc is 0.69, (h) GS online (with 20 words of information) categorization accuracy Acc is 0.85, and (i) GS online (with 39 words of information) categorization accuracy Acc is 0.92.

represent the structure of the model. Thus, the distance between objects at each learning step can be mapped onto two-dimensional space using multidimensional scaling (MDS). In Fig. 13, the model has been incrementally updated from left to right. Here, the color of each point in the figure represents a category, whereas the point denotes an object.

The results (Fig. 13) show that the objects that belong to the same category converge as the learning of the models progresses. At the beginning of learning, all the objects are distributed near the center because all of them are classified into each category with equal probability. By updating the models, the distance between the categories is increased. By the end of learning, objects that belong to the same category lie within a close range. In addition, pairs of objects such as "plastic bottle" and "glass bottle," "chips" and "cookie," "yarn" and "plushie," i.e., objects that have similar modalities and/or word information, converged with each other. From the results, we conclude that the proposed online algorithm enables incremental learning of the relationships among categories.

5.4 Forgetting factor and model selection based on word information

In this section, we tested online MLDA for various forgetting factors λ and model selection based on word information. We also made comparisons among batch-VB, batch Gibbs sampling, and online Gibbs sampling. The results when word information is given to 15 objects are shown in Fig. 14. In the figure,

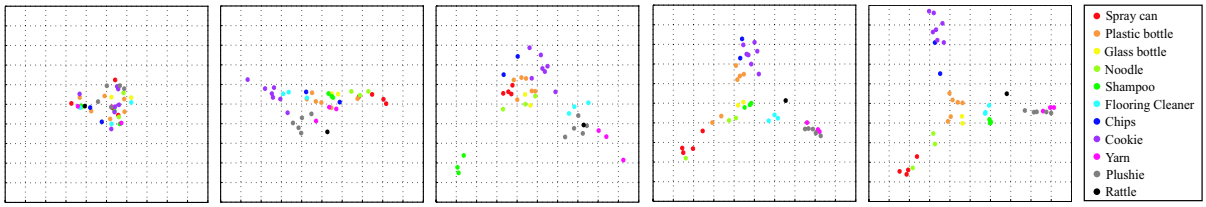

Figure 13: Learning model plotted by multidimensional scaling.

the horizontal and vertical axes represent the number of trained objects and the category recognition rate, respectively. Each object data set including word information was used for training the model one by one in an online manner. Every time new object data were input to train the model, category recognition was carried out by using all 50 objects for evaluating the category recognition rate. The forgetting factor was varied in increments of 0.1 from 0.0 to 1.0, and each recognition rate was calculated as an average over 100 trials.

From the figure, we see that the recognition rate is about 30% when only one object is input to the system in all cases. In online learning, when the forgetting factor was set to the fixed value of 0.0 and 0.1, the accuracy of categorization was 66.8% and 73.0%, respectively. Here, the best accuracy achieved when $\lambda = 0.1$ was set as a fixed value of the forgetting factor, and then, the rate dwindles as λ increases. Because $\lambda = 1.0$ is equivalent to forgetting all previously learned objects, online learning does not work at all in this particular case. As mentioned earlier, $\lambda = 0$ corresponds to o-LDA [20]. Here, we consider that learning starts from scratch, and thus having no batch initialization phase results in moderate results of o-LDA. In fact, Fig. 14 shows that $\lambda = 0.1$ achieved better results than o-LDA. These results indicate that the proposed online algorithm with appropriate forgetting factor λ works well. In addition, by using the proposed model selection method, classification accuracy was improved up to 83.1%. Although the best result is obtained by the batch-based Gibbs sampling, the proposed method exceeds the GS batch learning when the number of learned objects ranged from about 10 to 20. This result shows that concept formation close to that of a human can be achieved by learning based on sequentially varying the forgetting factor using model selection based on word information. Thus, the proposed algorithm can be considered as effective for object concept formation.

5.5 Inferring words for unseen objects

We evaluated word inference performance for unseen objects. Recognition and word prediction was carried out for unknown objects with red frames in Fig. 8. Note that the test samples contain one object from each category. At first, the training samples were used for training the proposed online MLDA. Then, word-generation probability $P(w^w|\bar{\boldsymbol{w}}^v, \bar{\boldsymbol{w}}^a, \bar{\boldsymbol{w}}^h)$ was calculated for all words using the test samples, i.e., unknown objects.

To determine the accuracy of the predicted word, threshold values are introduced. The words that had been estimated with a probability above the thresholds were evaluated as being suitable or not for describing the object. To determine the thresholds, the relationship among the number of predicted words, recall, and precision are calculated. The highest probability, $\mathrm{argmax}_{w^w} P(w^w|\bar{\boldsymbol{w}}^v, \bar{\boldsymbol{w}}^a, \bar{\boldsymbol{w}}^h)$, is

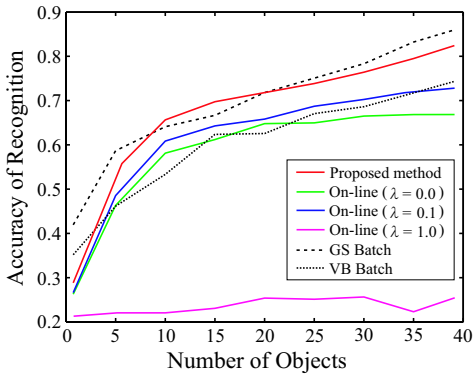

Figure 14: Forgetting factor and model selection versus accuracy of categorization.

used to calculate the following threshold:

$$Threshold = A \underset{w^w}{\operatorname{argmax}} P(w^w | \bar{\boldsymbol{w}}^v, \bar{\boldsymbol{w}}^a, \bar{\boldsymbol{w}}^h), \tag{17}$$

where A represents a coefficient that determines the threshold. Recall and precision are calculated from the words whose probabilities are above the threshold as follows:

$$Recall = \frac{(\text{number of correctly estimated words})}{(\text{number of grounded words})}, \tag{18}$$

$$Precision = \frac{(\text{number of correctly estimated words})}{(\text{number of estimated words that are larger than a threshold})}. \tag{19}$$

In this experiment, the coefficient of the threshold A is varied in increments of 0.001 from 0 to 1. The results are illustrated in Fig. 15. In the figure, the horizontal axis represents the coefficient of the threshold A. The vertical axes on the left and right correspond to the recall–precision rate and the number of estimated words, respectively. From the figure, one can see that recall and the number of predicted words dwindle, whereas precision improves as the coefficient increases. When the coefficient is set to 0.3, i.e., 1/3 the maximum predicted probability, recall and precision are approximately 60%. In this case, the result of 45 predicted words is obtained. However, in the scenario in which interaction between the user and the robot is assumed, it is desirable for the percentage of correct answers of the predicted words, i.e., the precision, to be high. Therefore, 50% of the maximum predicted probability ($A = 0.5$) is used as the coefficient in the later experiments. In this case, the number of estimated words is about 30, indicating that about three words were estimated from each object. Figure 16 shows some examples of word inference. In this figure, the horizontal and vertical axes represent words and probability, respectively. The red dashed line shows the threshold for each case. The words listed in the figure are some examples, and the predicted words are represented as a 145-dimensional probability vector. Bars with different colors represent the words with high predicted probability over the threshold. The red bar indicates an incorrectly inferred word, whereas the green bar indicates a correct word. Although some irrelevant words are inferred, one can see that the online algorithm inferred words reasonably well for each unseen object. At least suitable words for representing the category of

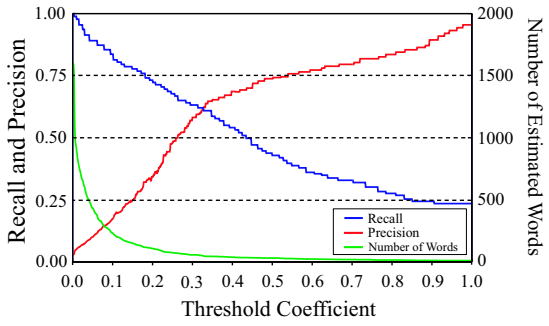

Figure 15: Number of estimated words and recall precision.

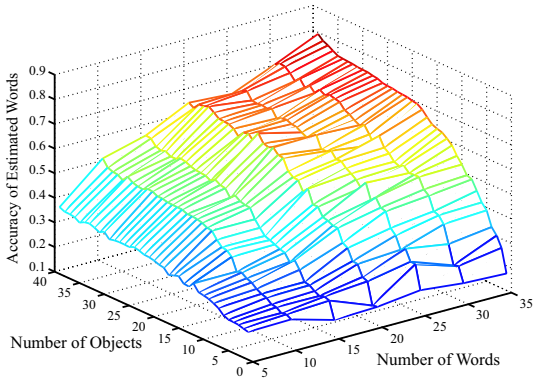

Figure 17: Accuracy of estimated words.

the object, e.g., "spray can," "plushie," "snack," etc., can be inferred. Common nouns representing the object category such as "plushie," "plastic bottle," etc. can be suitably recollected for unseen objects, even when the word is included only once in the training samples. In comparing objects that can generate sound with those that cannot, a major difference is seen in the probability of occurrence of the words associated with "sound." Also, in any object, it can be seen that words that are common to multiple categories such as "hard" and "soft" were predicted with high probability compared to the other ones. When the noodle was recognized, the word that does not represent noodle such as "hard" was predicted with a higher probability than the threshold. This is because the noodle has similar visual features with the spray can, which leads to misclassification during the recognition process. In fact, the words "noodle" and "spray" are predicted with the same probabilities. However, even if incorrect recognition is performed in this manner, reliable word information such as "sound" can be estimated with higher probability compared to other words.

In contrast, words that are not shared within a category but spread over several categories were predicted with relatively low probability $P(w^w|\bar{\boldsymbol{w}}^v, \bar{\boldsymbol{w}}^a, \bar{\boldsymbol{w}}^h)$. For instance, a color such as "red" is not shared in a single object category; therefore, it is hard to understand the meaning of "red" through the category. This problem can be easily resolved by having a category of red objects. This is an issue having to do with the granularity of categories. Selective attention is a key for modeling this granularity of categories, as discussed in [27]. If the same threshold value is used, the accuracy of predicted words according to the number of learning objects and the number of given words is shown in Fig. 17. In the training phase, the numbers of objects and given words were varied, and the word prediction accuracy was evaluated in each case. Each accuracy value was calculated as an average over 100 trials. In the figure, the x axis, y axis, and z axis are the number of given words, the number of training objects, and word prediction accuracy, respectively. From the figure, one can see that the word prediction accuracy improves as the numbers of words and objects increase. Only 10 objects are required to have word information to obtain about 50% word prediction accuracy for unseen objects. It is natural that the highest prediction accuracy (79.4%) is obtained when all objects and 35 words are used for the training.

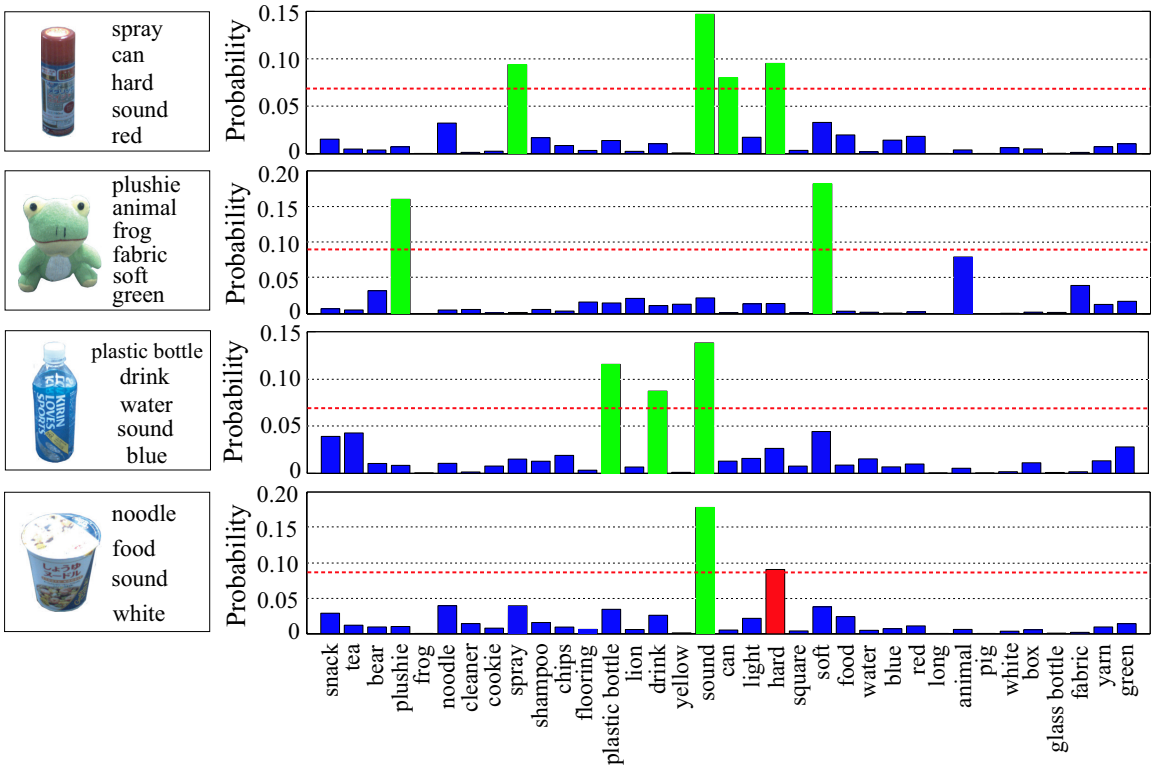

Figure 16: Some examples of word inference. The volunteers used the words in the box for describing the object. Green and red bars represent correctly inferred words and incorrectly inferred words, respectively. Only the words that had been estimated with a probability above the thresholds are evaluated.

These results show that the proposed online algorithm enables the robot to learn the meaning of words incrementally and describe unseen objects using suitable words.

6 Conclusion

This paper discussed an online object concept formation method by autonomous robots. To develop autonomous learning robots, we first proposed a fully autonomous acquisition method of multimodal information. The latter part of this paper was devoted to online MLDA using Gibbs sampling. We demonstrated that the proposed particle-filter-based approach enhances the performance of online learning. These frameworks make it possible for the robot to learn object concepts naturally in everyday operation in conjunction with a small amount of linguistic information from human users. Performance improvement of the online MLDA and estimation of the number of categories are left for future work. We are planning to apply a nonparametric Bayesian approach such as [28] to the latter problem.

REFERENCES

[1] Fergus, R., Perona, P., and Zisserman, A., "Object Class Recognition by Unsupervised Scale-invariant Learning," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Vol. 2, pp. 264–271 (2003).

[2] Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A. and Freeman, W.T., "Discovering Object Categories in Image Collections," in Proc. Int. Conf. on Computer Vision, Beijing, pp. 370–377 (2005).

[3] Wang, C., Blei, D., and Fei-Fei, L., "Simultaneous Image Classification and Annotation," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 1903–1910 (2009).

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

[5] Sinapov, J. and Stoytchev, A., "Object Category Recognition by a Humanoid Robot Using Behavior-Grounded Relational Learning," in Proc. IEEE Int. Conf. on Robotics and Automation, pp. 184–190 (2011).

[6] Natale, L., Metta, G. and Sandini, G., "Learning Haptic Representation of Objects," in Proc. IEEE Int. Conf. on Intelligent Manipulation and Grasping, (2004).

[7] Schneider, A., Sturm, J., Stachniss, C., Reisert, M., Burkhardt, H. and Burgard, W. "Object Identification with Tactile Sensors Using Bag-of-Features," in Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 243–248 (2009).

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

[9] Nakamura, T., Nagai, T. and Iwahashi, N., "Grounding of Word Meanings in Multimodal Concepts Using LDA," in Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 3943–3948 (2009).

[10] Rosch, E., "Principles of Categorization," Concepls: Core Readings, pp. 189–206 (1998).

[11] Welke, K., Issac, J., Schiebener, D., Asfour, T. and Dillmann, R., "Autonomous Acquisition of Visual Multi-view Object Representations for Object Recognition on a Humanoid Robot," in Proc. IEEE Int. Conf. on Robotics and Automation, pp. 2012–2019 (2010).

[12] Kojima, M., Okada, K. and Inaba, M., "Visual Memory Acquisition Behavior in Humanoid Through Picking-up Motion," in Proc. of the 25th Annual Conference on Robotics Society of Japan, 3H22 (2007) (in Japanese).

[13] Stasse, O., Larlus, D., Lagarde, B., Escande, A., Saidi, F., Kheddar, A., Yokoi, K. and Jurie, F., "Towards Autonomous Object Reconstruction for Visual Search by the Humanoid Robot HRP-2," in Proc. IEEE/RAS Int. Conf. on Humanoid Robots, pp. 151–158 (2007).

[14] Kojima, M., Okada, K., Inamura, T. and Inaba, M., "Humanoid Robot That Observes Handed Rotation Symmetry Object and Generates Appearance Model," in Proc. JSME Conf. on Robotics and Mechatronics, 2A1-D27 (2006) (in Japanese).

[15] Yamazaki, K., Tomono, M., Tsubouchi, T. and Yuta, S., "Object Shape Reconstruction and Pose Estimation by a Camera Mounted on a Mobile Robot," in Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 4019–4025 (2004).

[16] Attamimi, M., Mizutani, A., Nakamura, T., Sugiura, K., Nagai, T., Iwahashi, N., Okada, H., and Omori, T., "Learning Novel Objects Using Out-of-Vocabulary Word Segmentation and Object Extraction for Home Assistant Robots," in Proc. IEEE Int. Conf. on Robotics and Automation, pp. 745–750 (2010).

[17] Attamimi, M., Mizutani, A., Nakamura, T., Nagai, T., Funakoshi, K. and Nakano, M., "Real-time 3D Visual Sensor for Robust Object Recognition," in Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 4560–4565 (2010).

[18] Blei, D.M., Ng, A.Y., and Jordan, M.I., "Latent Dirichlet Allocation," Journal of Machine Learning Research, pp. 993–1022 (2003).

[19] Hoffman, M., Blei, D. and Bach, F., "Online Learning for Latent Dirichlet Allocation," in Proc. Conf. on Neural Information Processing Systems, pp. 856–864 (2010).

[20] Banerjee, A. and Basu, S., "Topic Models over Text Streams: A Study of Batch and Online Unsupervised Learning," in Proc. SIAM Int. Conf. on Data Mining, pp. 431–436 (2007).

[21] Canini, K.R., Shi, L. and Griffiths, T.L., "Online Inference of Topics with Latent Dirichlet Allocation," in Proc. Int. Conf. Artificial Intelligence and Statistics, Vol. 5, pp. 65–72 (2009).

[22] Griffiths, T. and Steyvers, M., "Finding Scientific Topics," in Proc. National Academy of Sciences, Vol. 101, Suppl, No. 1, pp. 5228–5235 (2004).

[23] Rusu, R.B., Bradski, G., Thibaux, R. and Hsu, J., "Fast 3D Recognition and Pose Using the Viewpoint Feature Histogram," in Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 2155–2162 (2010).

[24] Maldonado, A., Klank, U. and Beetz, M., "Robotic Grasping of Unmodeled Objects Using Time-of-Flight Range Data and Finger Torque Information," in Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp. 2586–2591 (2010)

[25] Ito, K., Nakamura, T. and Nagai, T., "Joint Path Planning of Dual Arms Using Configuration Space RRT," in Proc. 28th Annual Conf. on Robotics Society of Japan, 1M2-2 (2010) (in Japanese).

[26] Vedaldi, A. and Fulkerson, B., "VLFeat—An Open and Portable Library of Computer Vision Algorithms," in Proc. of Association for Computing Machinery Multimedia, pp. 1469–1472 (2010).

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

[28] Teh, Y.W., Jordan, M.I., Beal, M.J. and Blei, D.M., "Hierarchical Dirichlet Processes," Journal of the American Statistical Association, Vol. 101, pp. 1566–1581 (2006).