Autonomous Acquisition of Multimodal Information for Online Object Concept Formation by Robots

Takaya Araki¹, Tomoaki Nakamura¹ Takayuki Nagai¹, Kotaro Funakoshi², Mikio Nakano², Naoto Iwahashi³
¹Department of Electronic Engineering, The University of Electro-Communications
1-5-1 Chofugaoka Chofu-shi, Tokyo 182-8585, Japan
²Honda Research Institute Japan Co., Ltd., 8-1 Honcho, Wako-shi, Saitama 351-0188, Japan
³NICT 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 — This paper proposes a robot that acquires multimodal information, i.e. auditory, visual, and haptic information, fully autonomous way using its embodiment. We also propose an online algorithm of multimodal categorization based on the acquired multimodal information and words, which are partially given by human users. The proposed 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. In order to obtain multimodal information, the robot detects an object on a flat surface. Then the robot grasps and shakes it for gaining haptic and auditory information. For obtaining visual information, the robot uses a hand held small observation table, so that the robot can control the viewpoints for observing the object. As for the multimodal concept formation, the multimodal LDA using Gibbs sampling is extended to the online version in this paper. The proposed algorithms are implemented on a real robot and tested using real everyday objects in order to show validity of the proposed system.

I. INTRODUCTION

In recent years, there has been a great deal of research on intelligent robots that coexist with people on a daily basis. However, many problems still remain unsolved. One of the important problems is management of objects by the robot. Since vast numbers of objects exist in the real world and the objects, which 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 the robot in the real environment. At this time, multimodal perceptual information that the robot obtains are of importance for the learning. Moreover, it is desirable for the robot to observe the object and obtain multimodal information fully autonomous fashion without any human’s help so that the robot develops object concepts by itself.

We have proposed an object concept formation framework based on the multimodal categorization by robots using statistical models such as pLSA and LDA [1][2]. We showed that the multimodal categorization makes it possible for the robot to categorize objects in a same way as humans do. This means that suited object concepts can be formed through the 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 a very important factor for human-robot coexistence.

In order to achieve this kind of learning in real environments autonomously, the robot has to obtain multimodal data such as visual, auditory, and haptic information by itself. Hence the 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 [6]–[9]. For example, in [7], the authors propose a method for obtaining multi-view object representations by handing over the target object to the robot. [8] proposes a robot that can grasp and obtain visual information of objects autonomously. In [10], the robotic systems can autonomously acquire 3D information of an object by going around it. We have proposed a method that a human user shows a target object to the robot for learning novel objects [11]. It is worth noting that all of these works do not consider the acquisition of multimodal information.

The goal of this paper is to develop a robot that learns object concepts online by autonomously gaining multimodal information on a daily basis in any way as illustrated in Fig.1. On the other hand, linguistic labels are also important for the object concept formation. In [2], we have shown that the robot can learn meanings of words by connecting multimodal concepts, which are formed by the multimodal categorization, and corresponding words. Here we take a step further to consider the learning process of whole object concepts including word meaning 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 the linguistic information for the multimodal categorization. Of course, the formed object concepts can be used for inferring suitable words to unseen objects. It should be noted that the word information must be given by the human user. However, it is not plausible that the human user always accompanies for giving the

978-1-61284-455-8/11/$26.00 ©2011 IEEE 1540
linguistic information. As we mentioned earlier, the learning process should be autonomous as much as possible. The robot expects to have linguistic information only when the human user is available. Therefore the robot is required to have an ability of forming object concepts using perceptual information and partially given (incomplete) words.

In this paper, a robot that acquires multimodal information fully autonomous way using the mounted sensors is proposed. The robot can acquire visual information form the 3D visual sensor [12], auditory information by shaking the object, and haptic information by grasping it. We also propose an online algorithm of multimodal categorization based on the autonomously acquired multimodal information and words, which are partially given by the human user. Multimodal LDA using Gibbs sampling is extended to the online version in this paper so that the robot can discard the data after using them for learning. This is because the perceptual information consumes a large amount of memory and the batch-type learning is inefficient.

In [13]–[15], some online algorithms for LDA have been proposed. An online variational Bayes (VB) algorithm for LDA is shown to be converged to a local optimum of the VB objective function in [13]. We extend Gibbs sampling-based LDA instead of VB-based LDA because Gibbs sampling yields better results in multimodal categorization [3]. Moreover, Gibbs sampling-based LDA has an easy to implement property. An online LDA using Gibbs sampler called o-LDA has been proposed in [15]; the batch initialization phase is required for o-LDA. In [14], the incremental Gibbs sampler has been proposed to improve performance of o-LDA. The basic idea is to rejuvenate old assignments, which means old data cannot be discarded.

II. 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 a grammar, which allow the robot to decompose recognized sentences into words. Here nouns and adjectives are considered. All perceptual information is vector quantized and represented as Bag-of-Features model.

The object concepts are represented by the multimodal LDA (MLDA) as illustrated in Fig. 2. In the figure, \( u^v, w^a, w^h, \) and \( w^w \) represent visual, auditory, haptic, and words information and are assumed to be drawn from each multinomial distribution parameterized by \( \beta^v, \beta^a, \beta^h, \) and \( \beta^w \), respectively. \( \pi^v, \pi^a, \pi^h, \) and \( \pi^w \) denote parameters of Dirichlet prior distributions for \( \beta^* \). \( z \) represents the category and is assumed to be drawn from multinomial distribution parameterized by \( \alpha \), which depends on Dirichlet prior distribution parameterized by \( \alpha \).

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 of words meaning is also possible through the MLDA model. The important point of the proposed system is that the robot can learn object concepts incrementally and autonomously in conjunction with partially given words information. In the next section, autonomous acquisition of multimodal information is described at first.

III. MULTIMODAL INFORMATION ACQUISITION

In Fig. 3, the robot platform used in this paper is shown. The robot equips a 3D visual sensor [12], 6-DOF arms, and 4-DOF hands with a tactile array sensor consisting of 162 elements and a microphone. These sensors are used for obtaining multimodal information: color images from multiple viewpoints, 3D information from the TOF camera, NIR reflectance intensities, pressure information from the
tactile sensors, and sound that is made by shaking the object.

In order to capture such information autonomously, there are three problems to be solved: (1) novel objects detection, (2) grasping of novel objects, and (3) observation method for perceptual information. The first one is simply resolved by the plane-based object detection as in [12] and [16]. After the 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. 3D information 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 [17]. The most important issue in 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 by its own fingers. To cope with this problem a small handheld 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 planned by dual-arm RRT-connect [18] considering the shape of the object, which is grasped by the hand. Then the robot can observe the object from various viewpoints by controlling the position of the observation table. The target object on the observation table can be easily segmented out by applying the plane-based object detection using 3D information again. Finally, color, texture, 3D, and NIR intensities of the object can be obtained. The scenario of the acquisition process for visual information is shown in Fig. 4. Figure 5 shows an example of visual information that is obtained by the actual robot.

The haptic information is acquired by performing the grasping actions in Fig. 6 five times. During the grasping action fingers are controlled to close at constant velocity and stop when the torque reaches to a predefined threshold value. The lower part of Fig. 6 shows actual sensor values of the tactile array sensor.

As for auditory information, the microphone that is mounted on the robot hand is used to capture the sound, which is produced when the robot shakes it. The problem here is the motor noise of the robot arm, since the arm moves relatively fast so that the object makes sound enough audible. Here we utilize the spectral subtraction technique to solve this problem. Specifically, the average spectrum of the motor noise is calculated in advance and it is subtracted from the input sound spectrum. Figure 7 shows a sequence of actions for obtaining the auditory information.

IV. ONLINE MULTIMODAL CATEGORIZATION

A. Processing of perceptual information

1) Visual information: The target object is segmented out in each image frame and then 36-dimensional PCA-SIFT descriptors [19] are computed. In the later experiment, ten image frames are captured for each object. From 300 to 400 feature vectors are extracted at each image that
results in about 3000–4000 features for each object. Each feature vector is vector quantized using a codebook with 500 clusters. The codebook is generated by k-means algorithm in advance. Finally, a histogram is built for the bag of features representation.

2) Auditory information: The sound is recorded while the robot grasps and shakes an object. Then the sound data is divided into frames and transformed into 13-dimensional MFCCs (Mel-Frequency Cepstrum Coefficient) as the feature vector. Finally, the feature vectors are vector quantized using the codebook with 50 clusters, and then a histogram is taken.

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

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

where $p(t)$ represents the approximated sensor value at time $t$ and $a$, $b$, $c$, and $d$ are parameters for the time series. Figure 8 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 a moment of contact to the end of the finger movement, $c$ depends on size of the object, and $d$ is in proportion to $a$. Since two parameters $a$ and $b$ are related to 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 the variation due to grasping point change can be absorbed. The feature vectors are vector quantized using the codebook with 15 clusters and the histogram is taken.

B. Online MLDA using Gibbs Sampling

The problem of the 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 involved in the parameter estimation because no approximation is incorporated and it is easy to implement. Now let $w^m$ be a set of captured multimodal information. In Gibbs sampling, the category $z_{mi}$, which is assigned to $i$-th data of modality $m \in \{\text{visual, auditory, tactile}\}$ of the target object, is sampled from the following conditional probability:

$$P(z_{mi} = k | z^{-mi}, w^m, \alpha, \pi^m) \propto (N_k^{-mi} + \alpha) \cdot \frac{N^{-mi}w^m_k + \pi^m}{N^{-mi}w^m_k + W^m\pi^m},$$  

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

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

and

$$N_{mk} = \sum_{w^m} N_{mw^m k}.$$  

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

The category assigned to $i$-th data of the modality $m$ is sampled according to Eq. (2). This process is repeated until $N_k$ converges to a certain value. After the convergence, the final estimates of parameters $\hat{\beta}_{w^m k}$ and $\hat{\theta}_k$ can be written as follows:

$$\hat{\beta}_{w^m k} = \frac{\hat{N}_{w^m k} + \pi^m}{\hat{N}_{mk} + W^m\pi^m},$$  

and

$$\hat{\theta}_k = \frac{\hat{N}_k + \alpha}{\sum_k \hat{N}_k + K\alpha},$$

where $K$ represents number of categories and $\hat{N}_k$ represents a converged value of $N_k$.

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 data input to the system. This may take a long time and be impractical especially for the interactive learning scenario where the human user is involved. In order 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. 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 the batch initialization phase to avoid this problem [15]. On the other hand, we introduce the forgetting factor $\lambda (0 < \lambda < 1)$ as follows:

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

where $\hat{N}_{w^m k}^{(j)}$ represents the converged value of $N_{w^m k}$ for the $j$-th object data. When a new object $j + 1$ is input, $\hat{N}_{w^m k}^{(j)}$ is used as the initial value of the Gibbs sampling by multiplying the forgetting factor $1 - \lambda$. Gibbs sampling is carried out by applying Eq. (2) to the new input data iteratively until convergence. Algorithm 1 outlines the proposed online MLDA for a single object. Every time the robot finds a novel object, the algorithm is applied to the new input data. It is worth noting that $\lambda = 0$ corresponds to o-LDA in [15].

C. Inference

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

\[
\hat{z} = \arg\max_z P(z|w_{obs}^m)
\]

\[
= \arg\max_z \int P(z|\theta)P(\theta|w_{obs}^m)d\theta. \tag{8}
\]

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

\[
P(w^w|w_{obs}^m) = \int \sum_z P(w^w|z)P(z|\theta)P(\theta|w_{obs}^m)d\theta. \tag{9}
\]

It should be noted that \( P(z|\theta) \) and \( P(\theta|w_{obs}^m) \) in Eqs. (8) and (9) can be updated by recalculating \( \theta \) for fixed \( \beta^m \) using Gibbs sampling.

V. EXPERIMENTS

Four experiments are carried out to evaluate the proposed method. Figure 9 shows 48 objects with 8 categories (ground truth) used in the experiments. We asked some volunteers to give a brief description of each object. Then each description was decomposed into words using Japanese morphological analysis, and only nouns and adjectives were extracted. These words were used as linguistic information of the objects. In the experiments the number of categories \( K \) is fixed as \( K = 8 \).

A. Autonomous acquisition of multimodal information

The proposed multimodal information acquisition system has been implemented on the robot platform. All objects in Fig. 9 were put on a table and the robot was succeeded in acquiring multimodal information of all objects autonomously. Some examples of the acquired multimodal information are given in Fig. 10. It should be noted that only distinctive 6 dimensions out of 50 are shown for the auditory histograms in the figure. During the observation of each object, volunteers gave a description for the object in turn. All nouns and adjectives that appeared in their sentences are listed in tab. I. These multimodal information and words are used in the following experiments.

B. Multimodal categorization

Multimodal categorizations have been carried out using batch VB [2], batch Gibbs sampling, and the proposed online Gibbs sampling. The robot categorized the multimodal data of 48 objects acquired in the previous section. The number of given words was varied to confirm the contribution of linguistic information for the categorization. The results are shown in Fig. 11. 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 same category because of their similarities in appearance when no words are used for the categorization. Meanwhile, by providing words information for 24 objects, the categorization accuracy improves in any methods. Moreover, almost all objects were categorized correctly when words were given for all objects in Gibbs sampling LDA.

\[
\text{Acc} = \frac{\text{Number of objects categorized correctly}}{\text{Number of all objects}}. \tag{10}
\]

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 given by using word information in all methods. Especially, plastic bottles (Object ID 13–20) and spray cans (Object ID 21–25) are classified into the same category because of their similarities in appearance when no words are used for the categorization. Meanwhile, by providing words information for 24 objects, the categorization accuracy improves in any methods.
Fig. 10. Examples of acquired multimodal information: (from top to bottom) color mapped images, depth images, NIR intensity images, tactile array sensor output (grasped enough), tactile histograms, and auditory histograms (only distinctive 6 dimensions are shown).

Fig. 11. Results of categorization: (a) VB-batch (w/o words info.) categorization accuracy $Acc$ is 0.65, (b) VB-batch (w/ 50% words info.) categorization accuracy is 0.77, (c) VB-batch (w/ all words info.) categorization accuracy is 0.83, (d) GS-batch (w/o words info.) categorization accuracy is 0.67, (e) GS-batch (w/ 50% words info.) categorization accuracy is 0.85, (f) GS-batch (w/ all words info.) categorization accuracy is 1.00, (g) GS-online (w/o words info.) categorization accuracy is 0.63, (h) GS-online (w/ 50% words info.) categorization accuracy is 0.88, (i) GS-online (w/ all words info.) categorization accuracy is 0.96, and (j) ground truth.

With respect to results of the proposed online Gibbs sampling LDA, it can be seen that better categorization results are given 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.

C. Forgetting factor of online MLDA

Here we have tested online MLDA for various forgetting factors $\lambda$. All 48 objects were ordered randomly and each object data including words information was used for training the model one by one in the online manner. Every time a new object data was input to train the model, category recognition was conducted using all of 48 objects for evaluating the category recognition rate. The forgetting factor was varied in increments of 0.05 form 0.00 to 1.00. Each recognition rate was calculated as an average over 100 times trials. We also made comparisons among batch-VB, batch Gibbs sampling, and online Gibbs sampling. The results are illustrated in Fig. 12. In the figure, the horizontal and vertical axes indicate number of trained objects and the category recognition rate, respectively.

From the figure one can confirm that the proposed online algorithm successfully learned objects incrementally. In all cases, the recognition rate is about 40% when only one object is input to the system. When the system learns all of 48 objects, the rate goes up to 76.9%, 84.9%, and 83.4% for $\lambda = 0.00$, $\lambda = 0.05$, and $\lambda = 0.10$, respectively. $\lambda = 0.05$ shows the best result among the proposed online algorithms, and then the rate dwindles as $\lambda$ increases. Because $\lambda = 1.00$ is equivalent to forget all previously learned objects, online
learning does not work at all in this particular case. Although the best result is given by the batch-based Gibbs sampling, the online algorithms with $\lambda = 0.05$ and 0.10 provide better results than that of batch-VB.

As we mentioned earlier, $\lambda = 0$ corresponds to o-LDA [15]. Here we consider the learning starts from scratch, thus no batch initialization phase results in moderate results of o-LDA. In fact, Fig. 12 shows that $\lambda = 0.05$ and 0.1 achieve better results than o-LDA. These results indicate that the proposed online algorithm with appropriate forgetting factor $\lambda$ works well. We use $\lambda = 0.05$ in all experiments in this paper.

In the online algorithm with $\lambda = 0.05$, category recognition failure occurred primarily for “plastic bottle” and “spray can” due to their similarities in perceptual properties.

D. Inferring words for unseen objects

We made an evaluation on words inference performance for unseen objects. Forty-eight objects are divided into 40 training and 8 test samples as shown in Fig. 9. 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 $P(w|\vec{w}^v, \vec{w}^a, \vec{w}^h)$ was calculated for all words using the test samples. The words with three highest probabilities were evaluated, that is, whether each word was suitable for describing the object or not. In the training phase, the numbers of objects and given words were varied, and words prediction accuracy was evaluated in each case.

The result is shown in Fig. 13. Each accuracy value was calculated as an average over 100 times trials. In the figure, $x$-axis, $y$-axis, and $z$-axis are number of given words, number of training objects, and words prediction accuracy, respectively. From the figure, one can see that the words prediction accuracy improves as the numbers of words and objects increase. Only five objects are required to have words information in order to obtain about 50% words prediction accuracy for unseen objects. It is natural that the highest prediction accuracy (73.3%) is obtained when all of 40 objects and 40 words are used for the training. Common nouns representing the object category such as “cup”, “bottle”, and so on can be suitably recollected for unseen objects even when the word is included only once in the training samples. The words that are shared among some categories, e.g. “hard” and “sound”, also have the similar property to common nouns. These results show that the proposed online algorithm enables the robot to learn words meaning incrementally and to describe unseen objects using suitable words.

On the other hand, the words that are not shared within a category but spread over some categories were predicted with relatively low probability $P(w|\vec{w}^v, \vec{w}^a, \vec{w}^h)$. For instance, 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 solved by having a category of red objects. This is an issue on granularity of categories. Selective attention is the key to model this granularity of categories as discussed in [3].

Another problem to be considered is features of the perceptual information. It is impossible to learn the concepts regarding color essentially because no color information is included in the visual features. In fact, almost all of false recollections are due to the problem that the local visual feature cannot represent color and global characteristics such as “square” and “long”. Figure 14 shows some examples of words inference. Although some irrelevant words are inferred, one can see that the online algorithm reasonably inferred words for each unseen object. At least suitable words for representing the categories of objects, e.g. plushie, plastic bottle, spray can, cookie, etc., can be inferred. On the other hand, some packages of cookies and kitchen wraps have similar appearance that leads to false recognition of their categories. These are responsible for the inference of inadequate words.

VI. CONCLUSION

This paper discussed online object concept formation by autonomous robots. In order 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. 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 the future works. We are planning to apply no parametric Bayesian approach such as [20] to the latter problem.

REFERENCES