frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 22 May 2018
doi: 10.3389/fnbot.2018.00022

Multimodal Hierarchical Dirichlet Process-Based Active Perception by a Robot

Tadahiro Taniguchi[1]*, *Ryo Yoshino*[1] and *Toshiaki Takano*[2]

[1] *Emergent Systems Laboratory, College of Information Science and Engineering, Ritsumeikan University, Ksatsu, Japan,* [2] *Adaptive Systems Laboratory, Department of Computer Science, Shizuoka Institute of Science and Technology, Fukuroi, Japan*

In this paper, we propose an active perception method for recognizing object categories based on the multimodal hierarchical Dirichlet process (MHDP). The MHDP enables a robot to form object categories using multimodal information, e.g., visual, auditory, and haptic information, which can be observed by performing actions on an object. However, performing many actions on a target object requires a long time. In a real-time scenario, i.e., when the time is limited, the robot has to determine the set of actions that is most effective for recognizing a target object. We propose an active perception for MHDP method that uses the information gain (IG) maximization criterion and lazy greedy algorithm. We show that the IG maximization criterion is optimal in the sense that the criterion is equivalent to a minimization of the expected Kullback–Leibler divergence between a final recognition state and the recognition state after the next set of actions. However, a straightforward calculation of IG is practically impossible. Therefore, we derive a Monte Carlo approximation method for IG by making use of a property of the MHDP. We also show that the IG has submodular and non-decreasing properties as a set function because of the structure of the graphical model of the MHDP. Therefore, the IG maximization problem is reduced to a submodular maximization problem. This means that greedy and lazy greedy algorithms are effective and have a theoretical justification for their performance. We conducted an experiment using an upper-torso humanoid robot and a second one using synthetic data. The experimental results show that the method enables the robot to select a set of actions that allow it to recognize target objects quickly and accurately. The numerical experiment using the synthetic data shows that the proposed method can work appropriately even when the number of actions is large and a set of target objects involves objects categorized into multiple classes. The results support our theoretical outcomes.

Keywords: active perception, cognitive robotics, topic model, multimodal machine learning, submodular maximization

OPEN ACCESS

Edited by:
Minoru Asada,
Osaka University, Japan

Reviewed by:
Shingo Murata,
National Institute of Informatics, Japan
J. Michael Herrmann,
University of Edinburgh,
United Kingdom
Jivko Sinapov,
Tufts University, United States

****Correspondence:***
Tadahiro Taniguchi
taniguchi@ci.ritsumei.ac.jp

Received: *24 August 2017*
Accepted: *23 April 2018*
Published: *22 May 2018*

Citation:
Taniguchi T, Yoshino R and Takano T (2018) Multimodal Hierarchical Dirichlet Process-Based Active Perception by a Robot. Front. Neurorobot. 12:22. doi: 10.3389/fnbot.2018.00022

1. INTRODUCTION

Active perception is a fundamental component of our cognitive skills. Human infants autonomously and spontaneously perform actions on an object to determine its nature. The sensory information that we can obtain usually depends on the actions performed on the target object. For example, when people find a gift box placed in front of them, they cannot perceive its weight without holding the box, and they cannot determine its sound without hitting or shaking it. In other words, we can obtain sensory information about an object by selecting and executing actions to manipulate it. Adequate action selection is important for recognizing objects quickly and accurately. This example about a human also holds for a robot. An autonomous robot that moves and helps people in a living environment should also select adequate actions to recognize target objects. For example, when a person asks an autonomous robot to bring an empty plastic bottle, the robot has to examine many objects by applying several actions (**Figure 1**). This type of information is important, because our object categories are formed on the basis of multimodal information, i.e., not only visual information is used, but also auditory, haptic, and other information. Therefore, a computational model of the active perception should be consistently based on a computational model for multimodal object categorization and recognition.

In spite of the wide range of studies about active perception (e.g., Borotschnig et al., 2000; Dutta Roy et al., 2004; Eidenberger and Scharinger, 2010; Krainin et al., 2011; Ferreira et al., 2013) and multimodal categorization for robots (e.g., Nakamura et al., 2007, 2011a; Sinapov and Stoytchev, 2011; Celikkanat et al., 2014; Sinapov et al., 2014), active perception methods for a robot, i.e., action selection methods for perception for unsupervised multimodal categorization, have not been sufficiently explored (see section 2).

This paper considers the active perception problem for unsupervised multimodal object categorization under the condition that a robot has already obtained several action primitives that are used to examine target objects. In the context of this study, we need to study active perception on an unsupervised multimodal categorization method having generality as much as possible because it is believed that unsupervised multimodal categorization is important for future language learning by robots, and the findings obtained in this study should be able to be applied to other unsupervised multimodal categorization models. It was suggested that a child forms a category based on his/her sensorimotor experience before learning a word for the category in a Bayesian manner, and learning the word is a matter of attaching a new label to this preexisting category (Kemp et al., 2010). The multimodal hierarchical Dirichlet process (MHDP) is a mathematically very general and sophisticated nonparametric Bayesian multimodal categorization method. Therefore, we adopt MHDP proposed by Nakamura et al. (2011b) as a representative computational model for unsupervised multimodal object categorization.

We develop an active perception method based on the MHDP in this paper. The MHDP is a sophisticated, fully Bayesian, probabilistic model for multimodal object categorization (Nakamura et al., 2011b) that is developed by enabling hierarchical Dirichlet process (HDP) (Teh et al., 2006) to have multimodal emission distributions corresponding to multiple sensor information[1]. Nakamura et al. (2011b) showed that the MHDP enables a robot to form object categories using multimodal information, i.e., visual, auditory, and haptic information, in an unsupervised manner. The MHDP can estimate the number of object categories as well because of the nature of Bayesian nonparametrics.

This paper describes a new MHDP-based active perception method for multimodal object recognition based on object categories formed by a robot itself. We found that an active perception method that has a good theoretical nature, i.e., the performance of the greedy algorithm is theoretically guaranteed (see section 4), can be derived for MHDP. Our formulation is based on a hierarchical Bayesian model. If a cognitive system of a robot is modeled by using hierarchical Bayesian model, a recognition state are usually represented by posterior distribution over latent variables, e.g., object categories. The purpose of an active perception is to infer appropriate posterior distribution with a small number of actions. In our approach, we propose an action selection method that can reduce the distance between inferred posterior distributions and true posterior distributions.

In this study, we define the active perception problem in the context of unsupervised multimodal object categorization as following. Which set of actions should a robot take to recognize a target object as accurately as possible under the constraint that the number of actions is restricted[2]? Our MHDP-based active perception method uses an IG maximization criterion, Monte Carlo approximation, and the lazy greedy algorithm. In this paper, we show that the MHDP provides the following three advantages for deriving an efficient active perception method.

1. The *IG maximization criterion* is *optimal* in the sense that a selected set of actions minimizes the expected Kullback–Leibler (KL) divergence between the final posterior distribution estimated using the information regarding all modalities and the posterior distribution of the category estimated using the selected set of actions (see section 4.1).
2. The IG has a *submodular* and non-decreasing property as a set function. Therefore, for performance, the greedy and lazy greedy algorithms are guaranteed to be near-optimal strategies (see section 4.2).
3. A *Monte Carlo approximation* method for the IG can be derived by exploiting MHDP's properties (see section 4.3).

Although the above properties follow from the theoretical characteristics of the MHDP, this has never been pointed out in previous studies.

The main contributions of this paper are that we

- develop an MHDP-based active perception method, and
- show its effectiveness through experiments using an upper-torso humanoid robot and synthetic data.

The proposed active perception method can be used for general purposes, i.e., not only for robots but also for other target

[1] HDP is a nonparametric Bayesian extension of latent Dirichlet allocation (LDA) (Blei et al., 2003), which has been widely used for document-word clustering. The nonparametric Bayesian extension allows HDP to estimate the number of topics, i.e., clusters, as well.

[2] We can consider an extension of this problem by introducing different cost to each action, i.e., different action requires different time and energy. However, for simplicity, this paper focuses on the problem in which cost for each action is the same.

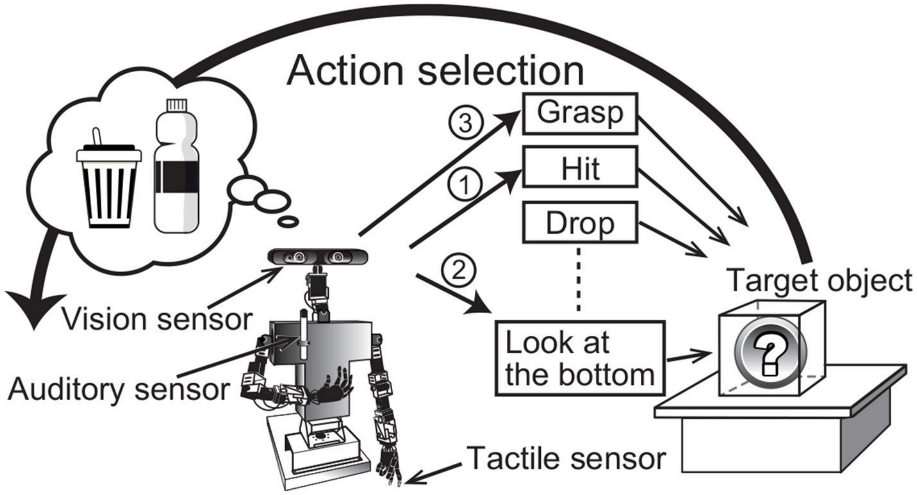

FIGURE 1 | Overview of active perception for multimodal object category recognition. The numbers attached to the arrows show a sample of the order of action selection by the robot.

domains to which the MHDP can be applied. In addition, The proposed method can be easily extended for other multimodal categorization methods with similar graphical models, e.g., multimodal latent Dirichlet allocation (MLDA) (Nakamura et al., 2009). However, in this paper, we focus on the MHDP and the robot active perception scenario, and explain our method on the basis of this task.

The remainder of this paper is organized as follows. Section 2 describes the background and work related to our study. Section 3 briefly introduces the MHDP, proposed by Nakamura et al. (2011b), which enables a robot to obtain object categories by fusing multimodal sensor information in an unsupervised manner. Section 4 describes our proposed action selection method. Section 5 discusses the effectiveness of the action selection method through experiments using an upper-torso humanoid robot. Section 6 describes a supplemental experiment using synthetic data. Section 7 concludes this paper.

2. BACKGROUND AND RELATED WORK

2.1. Multimodal Categorization

The human capability for object categorization is a fundamental topic in cognitive science (Barsalou, 1999). In the field of robotics, adaptive formation of object categories that considers a robot's embodiment, i.e., its sensory-motor system, is gathering attention as a way to solve the symbol grounding problem (Harnad, 1990; Taniguchi et al., 2016).

Recently, various computational models and machine learning methods for multimodal object categorization have been proposed in artificial intelligence, cognitive robotics, and related research fields (Roy and Pentland, 2002; Natale et al., 2004; Nakamura et al., 2007, 2009, 2011a,b, 2014; Iwahashi et al., 2010; Sinapov and Stoytchev, 2011; Araki et al., 2012; Griffith et al., 2012; Ando et al., 2013; Celikkanat et al., 2014; Sinapov et al., 2014). For example, Sinapov and Stoytchev (2011) proposed a graph-based multimodal categorization method that allows a robot to recognize a new object by its similarity to a set of familiar objects. They also built a robotic system that categorizes 100 objects from multimodal information in a supervised manner (Sinapov et al., 2014). Celikkanat et al. (2014) modeled the context in terms of a set of concepts that allow many-to-many relationships between objects and contexts using LDA.

Our focus of this paper is not a supervised learning-based, but an unsupervised learning-based multimodal categorization method and an active perception method for categories formed by the method. Of these, a series of statistical unsupervised multimodal categorization methods for autonomous robots have been proposed by extending LDA, i.e., a topic model (Nakamura et al., 2007, 2009, 2011a,b, 2014; Araki et al., 2012; Ando et al., 2013). All these methods are Bayesian generative models, and the MHDP is a representative method of this series (Nakamura et al., 2011b). The MHDP is an extension of the HDP, which was proposed by Teh et al. (2006), and the HDP is a nonparametric Bayesian extension of LDA (Blei et al., 2003). Concretely, the generative model of the MHDP has multiple types of emissions that correspond to various sensor data obtained through various modality inputs. In the HDP, observation data are usually represented as a bag-of-words (BoW). In contrast, the observation data in the MHDP use bag-of-features (BoF) representations for multimodal information. BoF is a histogram-based feature representation that is generated by quantizing observed feature vectors. Latent variables that are regarded as indicators of *topics* in the HDP correspond to *object categories* in the MHDP. Nakamura et al. (2011b) showed that the MHDP enables a robot to categorize a large number of objects in a home environment into categories that are similar to human categorization results.

To obtain multimodal information, a robot has to perform actions and interact with a target object in various ways, e.g., grasping, shaking, or rotating the object. If the number of actions and types of sensor information increase, multimodal categorization and recognition can require a longer time. When the recognition time is limited and/or if quick recognition is required, it becomes important for a robot to select a small number of actions that are effective for accurate recognition. Action selection for recognition is often called active perception. However, an active perception method for the MHDP has not been proposed. This paper aims to provide an active perception method for the MHDP.

2.2. Active Perception

Generally, active perception is one of the most important cognitive capabilities of humans. From an engineering viewpoint, active perception has many specific tasks, e.g., localization, mapping, navigation, object recognition, object segmentation, and self–other differentiation.

In machine learning, *active learning* is defined as a task in which a method interactively queries an information source to obtain the desired outputs at new data points to learn efficiently Settles (2012). Active learning algorithms select an unobserved input datum and ask a user (labeler) to provide a training signal (label) in order to reduce uncertainty as quickly as possible (Cohn et al., 1996; Muslea et al., 2006; Settles, 2012). These algorithms usually assume a supervised learning problem. This problem is related to the problem in this paper, but is fundamentally different.

Historically, active vision, i.e., active visual perception, has been studied as an important engineering problem in computer vision. Dutta Roy et al. (2004) presented a comprehensive survey of active three-dimensional object recognition. For example, Borotschnig et al. (2000) proposed an active vision method in a parametric eigenspace to improve the visual classification results. Denzler and Brown (2002) proposed an information theoretic action selection method to gather information that conveys the true state of a system through an active camera. They used the mutual information (MI) as a criterion for action selection. Krainin et al. (2011) developed an active perception method in which a mobile robot manipulates an object to build a three-dimensional surface model of it. Their method uses the IG criterion to determine when and how the robot should grasp the object.

Modeling and/or recognizing a single object as well as modeling a scene and/or segmenting objects are also important tasks in the context of robotics. Eidenberger and Scharinger (2010) proposed an active perception planning method for scene modeling in a realistic environment. van Hoof et al. (2012) proposed an active scene exploration method that enables an autonomous robot to efficiently segment a scene into its constituent objects by interacting with the objects in an unstructured environment. They used IG as a criterion for action selection. InfoMax control for acoustic exploration was proposed by Rebguns et al. (2011).

Localization, mapping, and navigation are also targets of active perception. Velez et al. (2012) presented an online planning algorithm that enables a mobile robot to generate plans that maximize the expected performance of object detection. Burgard et al. (1997) proposed an active perception method for localization. Action selection is performed by maximizing the weighted sum of the expected entropy and expected costs. To reduce the computational cost, they only consider a subset of the next locations. Roy and Thrun (1999) proposed a coastal navigation method for a robot to generate trajectories for its goal by minimizing the positional uncertainty at the goal. Stachniss et al. (2005) proposed an information-gain-based exploration method for mapping and localization. Correa and Soto (2009) proposed an active perception method for a mobile robot with a visual sensor mounted on a pan-tilt mechanism to reduce localization uncertainty. They used the IG criterion, which was estimated using a particle filter.

In addition, various studies on active perception by a robot have been conducted (Natale et al., 2004; Ji and Carin, 2006; Schneider et al., 2009; Tuci et al., 2010; Saegusa et al., 2011; Fishel and Loeb, 2012; Pape et al., 2012; Sushkov and Sammut, 2012; Gouko et al., 2013; Hogman et al., 2013; Ivaldi et al., 2014; Zhang et al., 2017). In spite of a large number of contributions about active perception, few theories of active perception for multimodal object category recognition have been proposed. In particular, an MHDP-based active perception method has not yet been proposed, although the MHDP-based categorization method and its series have obtained many successful results and extensions.

2.3. Active Perception for Multimodal Categorization

Sinapov et al. (2014) investigated multimodal categorization and active perception by making a robot perform 10 different behaviors; obtain visual, auditory, and haptic information; explore 100 different objects, and classify them into 20 object categories. In addition, they proposed an active behavior selection method based on confusion matrices. They reported that the method was able to reduce the exploration time by half by dynamically selecting the next exploratory behavior. However, their multimodal categorization is performed in a supervised manner, and the theory of active perception is still heuristic. The method does not have theoretical guarantees of performance.

IG-based active perception is popular, as shown above, but the theoretical justification for using IG in each task is often missing in many robotics papers. Moreover, in many cases in robotics studies, IG cannot be evaluated directly, reliably, or accurately. When one takes an IG criterion-based approach, how to estimate the IG is an important problem. In this study, we focus on MHDP-based active perception and develop an efficient near-optimal method based on firm theoretical justification.

3. MULTIMODAL HIERARCHICAL DIRICHLET PROCESS FOR STATISTICAL MULTIMODAL CATEGORIZATION

We assume that a robot forms object categories using the MHDP from multimodal sensory data. In this section, we briefly introduce the MHDP on which our proposed active perception method is based (Nakamura et al., 2011b). The MHDP assumes that an observation node in its graphical model corresponds to an action and its corresponding modality. Nakamura et al. (2011b) employed three observation nodes in their graphical model, i.e., haptic, visual, and auditory information nodes. Three actions, i.e., grasping, looking around, and shaking, correspond to these modalities, respectively. However, the MHDP can be easily extended to a model with additional types of sensory inputs. It is without doubt that autonomous robots will also gain more types of action for perception. For modeling more general cases, an MHDP with M actions is described in this paper. A graphical model of the MHDP is illustrated in **Figure 2**. In this section, we describe the MHDP briefly. For more details, please refer to Nakamura et al. (2011b).

The index $m \in \mathbf{M}$ ($\#(\mathbf{M}) = M$) in **Figure 2** represents the type of information that corresponds to an action for perception, e.g., hitting an object to obtain its sound, grasping an object to test its shape and hardness, or looking at all of an object by rotating it. We assume that a robot has action primitives and it can execute one of the actions by selecting the index of the action primitives. The observation $x^m_{jn} \in X^m$ is the m-th modality's n-th feature for the j-th target object. X^m represents a set of observation of m-th modality. The observation x^m_{jn} is assumed to be drawn from a categorical distribution whose parameter is θ^m_k, where k is an index of a latent topic. Each index k is drawn from a categorical distribution whose parameter is β that is drawn from a Dirichlet distribution parametrized by γ. Parameter θ^m_k is assumed to be drawn from the Dirichlet prior distribution whose parameter is α^m_0. The MHDP assumes that a robot obtains each modality's sensory information as a BoF representation. Each latent variable t^m_{jn} is drawn from a topic proportion, i.e., a parameter of a multinomial distribution, of the j-th object π_j whose prior is a Dirichlet distribution parametrized by λ.

Similarly to the generative process of the original HDP (Teh et al., 2006), the generative process of the MHDP can be described as a Chinese restaurant franchise, which is the name of a special type of probabilistic process in Bayesian nonparametrics (Teh et al., 2005). The learning and recognition algorithms are both derived using Gibbs sampling. In its learning process, the MHDP estimates a latent variable t^m_{jn} for each feature of the j-th object and a topic index k_{jt} for each latent variable t. The combination of latent variable and topic index corresponds to a topic in LDA (Blei et al., 2003). Using the estimated latent variables, the categorical distribution parameter θ^m_k and topic proportion of the j-th object π_j are drawn from the posterior distribution.

The selection procedure for latent variable t^m_{jn} is as follows. The prior probability that x^m_{jn} selects t is

$$P(t^m_{jn} = t|\lambda) = \begin{cases} \frac{\sum_m w^m N^m_{jt}}{\lambda + \sum_m w^m N^m_j - 1}, & (t = 1, \cdots, T_j), \\ \frac{\lambda}{\lambda + \sum_m w^m N^m_j - 1}, & (t = T_j + 1), \end{cases}$$

where w^m is a weight for the m-th modality, To balance the influence of different modalities, w^m are set as hyperparameters. The weight w^m increases the influence of the modality m on multimodal category formation. N^m_{jt} is the number of m-th modality observations that are allocated to t in the j-th object, and λ is a hyperparameter. In the Chinese restaurant process, if the number of observed features $N_{jt} = \sum_m w^m N^m_{jt}$ that are allocated to t increases, the probability at which a new observation is allocated to the latent variable t increases. Using the prior distribution, the posterior probability that observation x^m_{jn} is allocated to the latent variable t becomes

$$\begin{aligned} P(t^m_{jn} = t|X^m, \lambda) &= \frac{P(x^m_{jn}|X^m_{k=k_{jt}})P(t^m_{jn} = t|\lambda)}{P(x^m_{jn}|X^m \setminus \{x^m_{jn}\}, \lambda)} \\ &\propto \begin{cases} P(x^m_{jn}|X^m_{k=k_{jt}})\frac{\sum_m w^m N^m_{jt}}{\lambda + \sum_m w^m N^m_j - 1}, & (t = 1, \cdots, T_j), \\ P(x^m_{jn}|X^m_{k=k_{jt}})\frac{\lambda}{\lambda + \sum_m w^m N^m_j - 1}, & (t = T_j + 1), \end{cases} \end{aligned}$$

where N^m_j is the number of the m-th modality's observations about the j-th object. The set of observations that correspond to the m-th modality and have the k-th topic in any object are represented by X^m_k.

In the Gibbs sampling procedure, a latent variable for each observation is drawn from the posterior probability distribution. If $t = T_j + 1$, a new observation is allocated to a new latent variable. The dish selection procedure is as follows. The prior probability that the k-th topic is allocated on the t-th latent variable becomes

$$P(k_{jt} = k|\gamma) = \begin{cases} \frac{M_k}{\gamma + M - 1}, & (k = 1, \cdots, K), \\ \frac{\gamma}{\gamma + M - 1}, & (k = K + 1), \end{cases}$$

where K is the number of topic types, and M_k is the number of latent variables on which the k-th topic is placed. Therefore, the posterior probability that the k-th topic is allocated on the t-th latent variable becomes

$$\begin{aligned} P(k_{jt} = k|X, \gamma) &= P(X_{jt}|X_k)P(k_{jt} = k|\gamma) \\ &= \begin{cases} P(X_{jt}|X_k)\frac{M_k}{\gamma + M - 1}, & (k = 1, \cdots, K), \\ P(X_{jt}|X_k)\frac{\gamma}{\gamma + M - 1}, & (k = K + 1) \end{cases} \end{aligned}$$

where $X = \cup_m X^m$, $X_k = \cup_m X^m_k$, and X_{jt} is the set of the j-th object's observations allocated to the t-th latent variable. A topic index for the latent variable t for the j-th object is drawn using the posterior probability, where γ is a hyperparameter. If $k = K + 1$, a new topic is placed on the latent variable.

By sampling t^m_{jn} and k_{jt}, the Gibbs sampler performs probabilistic object clustering:

$$t^m_{jn} \sim P(t^m_{jn}|X^{-mjn}, \lambda), \tag{1}$$

$$k_{jt} \sim P(k_{jt}|X^{-jt}, \gamma), \tag{2}$$

where $X^{-mjn} = X \setminus \{x^m_{jn}\}$, and $X^{-jt} = X \setminus X_{jt}$. By sampling t^m_{jn} for each observation in every object using (1) and sampling k_{jt} for each latent variable t in every object using (2), all of the latent variables in the MHDP can be inferred.

If t^m_{jn} and k_{jt} are given, the probability that the j-th object is included in the k-th category becomes

$$P(k|X_j) = \frac{\Sigma^{T_j}_{t=1} \delta_k(k_{jt}) \sum_m w^m N^m_{jt}}{\sum_m w^m N^m_j}, \tag{3}$$

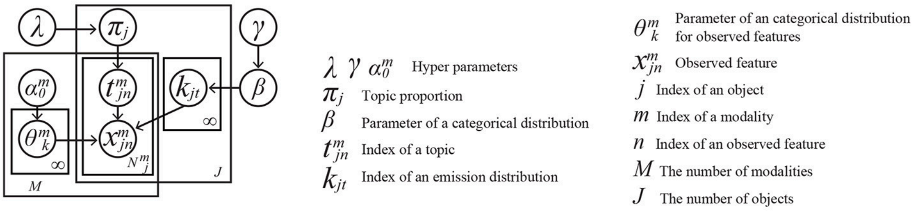
λ
π_j
γ
α_0^m
t_{jn}^m
k_{jt}
β
θ_k^m
x_{jn}^m
∞
∞
N_j^m
J
M
$\lambda\ \gamma\ \alpha_0^m$ Hyper parameters
π_j Topic proportion
β Parameter of a categorical distribution
t_{jn}^m Index of a topic
k_{jt} Index of an emission distribution
θ_k^m Parameter of an categorical distribution for observed features
x_{jn}^m Observed feature
j Index of an object
m Index of a modality
n Index of an observed feature
M The number of modalities
J The number of objects

FIGURE 2 | Graphical representation of an MHDP with *M* modalities corresponding to actions for perception.

where $X_j = \cup_m X_j^m$, w^m is the weight for the m-th modality and $\delta_a(x)$ is a delta function.

When a robot attempts to recognize a new object after the learning phase, the probability that feature x_{jn}^m is generated from the k-th topic becomes

$$P(x_{jn}^m|X_k^m) = \frac{w^m N_{kx_{jn}^m}^m + \alpha_0^m}{w^m N_k^m + d^m \alpha_0^m},$$

where d^m denotes the dimension of the m-th modality input, and $N_{kx_{jn}^m}^m$ represents the number of features x_{jn}^m that is corresponding to the index k. Topic k_t allocated to t for a new object is sampled from

$$k_t \sim P(k_{jt} = k|X, \gamma) \propto P(X_{jt}|X_k)\frac{\gamma}{\gamma + M - 1}.$$

These sampling procedures play an important role in the Monte Carlo approximation of our proposed method (see section 4.2.).

For a more detailed explanation of the MHDP, please refer to Nakamura et al. (2011b). Basically, a robot can autonomously learn object categories and recognize new objects using the multimodal categorization procedure described above. The performance and effectiveness of the method was evaluated in the paper.

4. ACTIVE PERCEPTION METHOD

4.1. Basic Formulation

A robot should have already conducted several actions and obtained information from several modalities when it attempts to select next action set for recognizing a target object. For example, visual information can usually be obtained by looking at the front face of the j-th object from a distance before interacting with the object physically. We assume that a robot has already obtained information corresponding to a subset of modalities $\mathbf{m_o}_j \subset \mathbf{M}$, where the subscript **o** means"originally" obtained modality information. When a robot faces a new object and has not obtained any information, $\mathbf{m_o}_j = \emptyset$.

The purpose of object recognition in multimodal categorization is different from conventional supervised learning-based pattern recognition problems. In supervised learning, the recognition result is evaluated by checking whether the output is the same as the truth label. However, in unsupervised learning, there are basically no truth labels. Therefore, the performance of active perception should be measured in a different manner.

The action set the robot selects is described as $\mathbf{A} = \{a_1, a_2, \ldots, a_{N_\mathbf{A}}\} \in \mathbf{2}^{\mathbf{M} \setminus \mathbf{m_o}_j}$, where $\mathbf{2}^{\mathbf{M} \setminus \mathbf{m_o}_j}$ is a family of subsets of $\mathbf{M} \setminus \mathbf{m_o}_j$, i.e., $\mathbf{A} \subset \mathbf{M} \setminus \mathbf{m_o}_j$, $a_i \in \mathbf{M} \setminus \mathbf{m_o}_j$ and N_A represents the number of available actions. We consider an effective action set for active perception to be one that largely reduces the distance between the final recognition state after the information from all modalities **M** is obtained and the recognition state after the robot executes the selected action set **A**. The recognition state is represented by the posterior distribution $P(\mathbf{z}_j|X_j^{\mathbf{m_o}_j \cup \mathbf{A}})$. Here, $\mathbf{z}_j = \{\{k_{jt}\}_{1 \le t \le T_j}, \{t_{jn}^m\}_{m \in \mathbf{M}, 1 \le n \le N_j^m}\}$ is a latent variable representing the j-th object's topic information, where $X_j^\mathbf{A} = \cup_{m \in \mathbf{A}} X_j^m, X_j^m = \{x_{j1}^m, \ldots, x_{jn}^m, \ldots, x_{jN_j^m}^m\}$. Probability $P(\mathbf{z}_j|X_j^{\mathbf{m_o}_j \cup \mathbf{A}})$ represents the posterior distribution related to the object category after taking actions $\mathbf{m_o}_j$ and **A**.

The final recognition state, i.e., posterior distribution over latent variables after obtaining the information from all modalities **M**, becomes $P(\mathbf{z}_j|X_j^\mathbf{M})$. The purpose of active perception is to select a set of actions that can estimate the posterior distribution most accurately. When L actions can be executed, if we employ KL divergence as the metric of the difference between the two probability distributions,

$$\underset{\mathbf{A} \in \mathbf{F}_L^{\mathbf{m_o}_j}}{\text{minimize}}\, \text{KL}\left(P(\mathbf{z}_j|X_j^\mathbf{M}), P(\mathbf{z}_j|X_j^{\mathbf{m_o}_j \cup \mathbf{A}})\right) \tag{4}$$

is a reasonable evaluation criterion for realizing effective active perception, where $\mathbf{F}_L^{\mathbf{m_o}_j} = \{\mathbf{A}|\mathbf{A} \subset \mathbf{M} \setminus \mathbf{m_o}_j, N_\mathbf{A} \le L\}$ is a feasible set of actions.

However, neither the true $X_j^\mathbf{M}$ nor $X_j^{\mathbf{m_o}_j \cup \mathbf{A}}$ can be observed before taking **A** on the j-th target object, and hence cannot be used at the moment of action selection. Therefore, a rational alternative for the evaluation criterion is the expected value of the KL divergence at the moment of action selection:

$$\underset{\mathbf{A} \in \mathbf{F}_L^{\mathbf{m_o}_j}}{\text{minimize}}\, \mathbb{E}_{X_j^{\mathbf{M} \setminus \mathbf{m_o}_j}|X_j^{\mathbf{m_o}_j}}[\text{KL}\left(P(\mathbf{z}_j|X_j^\mathbf{M}), P(\mathbf{z}_j|X_j^{\mathbf{m_o}_j \cup \mathbf{A}})\right)]. \tag{5}$$

Here, we propose to use the IG maximization criterion to select the next action set for active perception:

$$\mathbf{A}_j^* = \underset{\mathbf{A} \in \mathbf{F}_L^{\mathbf{m_o}_j}}{\text{argmax}}\, \text{IG}(\mathbf{z}_j; X_j^\mathbf{A}|X_j^{\mathbf{m_o}_j}) \tag{6}$$

$$= \underset{\mathbf{A} \in \mathbf{F}_L^{\mathbf{m_o}_j}}{\text{argmin}}\, \mathbb{E}_{X_j^\mathbf{A}|X_j^{\mathbf{m_o}_j}}[\text{KL}\left(P(\mathbf{z}_j|X_j^{\mathbf{m_o}_j \cup \mathbf{A}}), P(\mathbf{z}_j|X_j^{\mathbf{m_o}_j})\right)], \tag{7}$$

where $\mathrm{IG}(X;\, Y|Z)$ is the IG of Y for X, which is calculated on the basis of the probability distribution commonly conditioned by Z as follows:

$$\mathrm{IG}(X;\, Y|Z) = \mathrm{KL}\left(P(X, Y|Z), P(X|Z)P(Y|Z)\right).$$

By definition, the expected KL divergence is the same as $\mathrm{IG}(X;\, Y)$. The definition of IG and its relation to KL divergence are as follows.

$$\begin{aligned}\mathrm{IG}(X;\, Y) &= H(X) - H(X|Y)\\ &= \mathrm{KL}\left(P(X, Y), P(X)P(Y)\right)\\ &= \mathbb{E}_Y[\mathrm{KL}\left(P(X|Y), P(X)\right)].\end{aligned}$$

The optimality of the proposed criterion (6) is supported by Theorem 1.

Theorem 1. *The set of next actions* $\mathbf{A} \in \mathbf{F}_L^{\mathbf{m_o}_j}$ *that maximizes the* $\mathrm{IG}(\mathbf{z}_j;\, X_j^{\mathbf{A}}|X_j^{\mathbf{m_o}_j})$ *minimizes the expected KL divergence between the posterior distribution over* $\mathbf{z}_j$ *after all modality information has been observed and after* $\mathbf{A}$ *has been executed.*

$$\begin{aligned}&\underset{\mathbf{A}\in\mathbf{F}_L^{\mathbf{m_o}_j}}{\mathrm{argmin}}\, \mathbb{E}_{X_j^{\mathbf{M}\setminus\mathbf{m_o}_j}|X_j^{\mathbf{m_o}_j}}[\mathrm{KL}\left(P(\mathbf{z}_j|X_j^{\mathbf{M}}), P(\mathbf{z}_j|X_j^{\mathbf{m_o}_j\cup\mathbf{A}})\right)]\\ &\quad= \underset{\mathbf{A}\in\mathbf{F}_L^{\mathbf{m_o}_j}}{\mathrm{argmax}}\, \mathrm{IG}(\mathbf{z}_j;\, X_j^{\mathbf{A}}|X_j^{\mathbf{m_o}_j})\end{aligned}$$

Proof: See Appendix A.

This theorem is essentially the result of well-known characteristics of IG (see MacKay, 2003; Russo and Van Roy, 2016 for example). This means that maximizing IG is the optimal policy for active perception in an MHDP-based multimodal object category recognition task. As a special case, when only a single action is permitted, the following corollary is satisfied.

Corollary 1.1. *The next action* $m \in \mathbf{M} \setminus \mathbf{m_o}_j$ *that maximizes* $\mathrm{IG}(\mathbf{z}_j;\, X_j^m|X_j^{\mathbf{m_o}_j})$ *minimizes the expected KL divergence between the posterior distribution over* $\mathbf{z}_j$ *after all modality information has been observed and after the action has been executed.*

$$\begin{aligned}&\underset{m\in\mathbf{M}\setminus\mathbf{m_o}_j}{\mathrm{argmin}}\, \mathbb{E}_{X_j^{\mathbf{M}\setminus\mathbf{m_o}_j}|X_j^{\mathbf{m_o}_j}}[\mathrm{KL}\left(P(\mathbf{z}_j|X_j^{\mathbf{M}}), P(\mathbf{z}_j|X_j^{\{m\}\cup\mathbf{m_o}_j})\right)]\\ &\quad= \underset{m\in\mathbf{M}\setminus\mathbf{m_o}_j}{\mathrm{argmax}}\, \mathrm{IG}(\mathbf{z}_j;\, X_j^m|X_j^{\mathbf{m_o}_j}). \qquad (8)\end{aligned}$$

Proof: By substituting $\{m\}$ into $\mathbf{A}$ in Theorem 1, we can obtain the corollary.

Using IG, the active perception strategy for the next single action is simply described as follows:

$$m_j^* = \underset{m\in\mathbf{M}\setminus\mathbf{m_o}_j}{\mathrm{argmax}}\, \mathrm{IG}(\mathbf{z}_j;\, X_j^m|X_j^{\mathbf{m_o}_j}). \qquad (9)$$

This means that the robot should select the action m_j^* that can obtain the $X_j^{m_j^*}$ that maximizes the IG for the recognition result $\mathbf{z}_j$ under the condition that the robot has already observed $X_j^{\mathbf{m_o}_j}$.

However, we still have two problems, as follows.

1. The argmax operation in (6) is a combinatorial optimization problem and incurs heavy computational cost when $\#(\mathbf{M}\setminus\mathbf{m_o}_j)$ and L become large.
2. The calculation of $\mathrm{IG}(\mathbf{z}_j;\, X_j^{\mathbf{A}}|X_j^{\mathbf{m_o}_j})$ cannot be performed in a straightforward manner.

Based on some properties of the MHDP, we can obtain reasonable solutions for these two problems.

4.2. Sequential Decision Making as a Submodular Maximization

If a robot wants to select L actions $\mathbf{A}_j = \{a_1, a_2, \ldots, a_L\}$ ($a_i \in \mathbf{M} \setminus \mathbf{m_o}_j$), it has to solve (6), i.e., a combinatorial optimization problem. The number of combinations of L actions is ${}_{\#(\mathbf{M}\setminus\mathbf{m_o}_j)}C_L$, which increases dramatically when the number of possible actions $\#(\mathbf{M} \setminus \mathbf{m_o}_j)$ and L increase. For example, Sinapov et al. (2014) gave a robot 10 different behaviors in their experiment on robotic multimodal categorization. Future autonomous robots will have more available actions for interacting with a target object and be able to obtain additional types of modality information through these interactions. Hence, it is important to develop an efficient solution for the combinatorial optimization problem.

Here, the MHDP has advantages for solving this problem.

Theorem 2. *The evaluation criterion for multimodal active perception* $\mathrm{IG}(\mathbf{z}_j;\, X_j^{\mathbf{A}}|X_j^{\mathbf{m_o}_j})$ *is a submodular and non-decreasing function with regard to* $\mathbf{A}$.

Proof: As shown in the graphical model of the MHDP in **Figure 2**, the observations for each modality X_j^m are conditionally independent under the condition that a set of latent variables $\mathbf{z}_j = \{\{k_{jt}\}_{1\le t\le T_j}, \{t_{jn}^m\}_{m\in\mathbf{M}, 1\le n\le N_j^m}\}$is given. This satisfies the conditions of the theorem by Krause and Guestrin (2005). Therefore, $\mathrm{IG}(\mathbf{z}_j;\, X_j^m|X_j^{\mathbf{m_o}_j})$ is a submodular and non-decreasing function with regard to X_j^m.

Submodularity is a property similar to the convexity of a real-valued function in a vector space. If a set function $F: V \rightarrow R$ satisfies

$$F(A \cup x) - F(A) \ge F(A' \cup x) - F(A'),$$

where V is a finite set $\forall A \subset A' \subseteq V$ and $x \notin A$, the set function F has submodularity and is called a submodular function.

Function IG is not always a submodular function. However, Krause et al. proved that $\mathrm{IG}(U; A)$ is submodular and non-decreasing with regard to $A \subseteq S$ if all of the elements of S are conditionally independent under the condition that U is given. With this theorem, Krause and Guestrin (2005) solved the sensor allocation problem efficiently. Theorem 2 means that the problem (6) is reduced to a *submodular maximization problem*.

It is known that the greedy algorithm is an efficient strategy for the submodular maximization problem. Nemhauser et al. (1978) proved that the greedy algorithm can select a subset that is at most a constant factor $(1 - 1/\mathrm{e})$ worse than the optimal set, if the evaluation function $F(A)$ is submodular, non-decreasing, and $F(\emptyset) = 0$, where $F(\cdot)$ is a set function, and A

is a set. If the evaluation function is a submodular set function, a greedy algorithm is practically sufficient for selecting subsets in many cases. In sum, a greedy algorithm gives a near-optimal solution. However, the greedy algorithm is still inefficient because it requires an evaluation of all choices at each step of a sequential decision making process.

Minoux (1978) proposed lazy greedy algorithm to make the greedy algorithm more efficient for the submodular evaluation function. The lazy greedy algorithm can reduce the number of evaluations by using the characteristics of a submodular function.

4.3. Monte Carlo Approximation of IG

Equations (6) and (9) provide a robot with an appropriate criterion for selecting an action to efficiently recognize a target object. However, at first glance, it looks difficult to calculate the IG. First, the calculation of the expectation procedure $\mathbb{E}_{X_j^{\mathbf{A}}|X_j^{\mathbf{m_o}_j}}[\cdot]$ requires a sum operation over all possible $X_j^{\mathbf{A}}$. The number of possible $X_j^{\mathbf{A}}$ exponentially increases when the number of elements in the BoF increases. Second, the calculation of $P(\mathbf{z}_j|X_j^{\mathbf{A}\cup\mathbf{m_o}_j})$ for each possible observation $X_j^{\mathbf{A}}$ requires the same computational cost as recognition in the multimodal categorization itself. Therefore, the straightforward calculation for solving (9) is computationally impossible in a practical sense.

However, by exploiting a characteristic property of the MHDP, a Monte Carlo approximation can be derived. First, we describe IG as the expectation of a logarithm term.

$$\begin{aligned}\mathrm{IG}(\mathbf{z}_j; X_j^m|X_j^{\mathbf{m_o}_j}) &= \sum_{\mathbf{z}_j, X_j^m} P(\mathbf{z}_j, X_j^m|X_j^{\mathbf{m_o}_j}) \log \frac{P(\mathbf{z}_j, X_j^m|X_j^{\mathbf{m_o}_j})}{P(\mathbf{z}_j|X_j^{\mathbf{m_o}_j})P(X_j^m|X_j^{\mathbf{m_o}_j})}\\ &= \mathbb{E}_{\mathbf{z}_j, X_j^m|X_j^{\mathbf{m_o}_j}}\left[\log \frac{P(\mathbf{z}_j, X_j^m|X_j^{\mathbf{m_o}_j})}{P(\mathbf{z}_j|X_j^{\mathbf{m_o}_j})P(X_j^m|X_j^{\mathbf{m_o}_j})}\right]. \quad (10)\end{aligned}$$

An analytic evaluation of (10) is also practically impossible. Therefore, we adopt a Monte Carlo method. Equation (10) suggests that an efficient Monte Carlo approximation can be performed as shown below if we can sample

$$(z_j^{[k]}, X_j^{m[k]}) \sim P(\mathbf{z}_j, X_j^m|X_j^{\mathbf{m_o}_j}), \quad (k \in \{1, \ldots, K\}).$$

Fortunately, the MHDP provides a sampling procedure for $z_j^{[k]} \sim P(\mathbf{z}_j|X_j^{\mathbf{m_o}_j})$ and $X_j^{m[k]} \sim P(X_j^m|z_j^{[k]})$ in its original paper (Nakamura et al., 2011b). In the context of multimodal categorization by a robot, $X_j^{m[k]} \sim P(X_j^m|z_j^{[k]})$ is a prediction of an unobserved modality's sensation using observed modalities' sensations, i.e., cross-modal inference. The sampling process of $(z_j^{[k]}, X_j^{m[k]})$ can be regarded as a mental simulation by a robot that predicts the unobserved modality's sensation leading to a categorization result based on the predicted sensation and observed information.

$$\begin{aligned}(10) &\approx \frac{1}{K}\sum_k \log \frac{P(\mathbf{z}_j^{[k]}, X_j^{m[k]}|X_j^{\mathbf{m_o}_j})}{P(\mathbf{z}_j^{[k]}|X_j^{\mathbf{m_o}_j})P(X_j^{m[k]}|X_j^{\mathbf{m_o}_j})}\\ &= \frac{1}{K}\sum_k \log \frac{P(X_j^{m[k]}|\mathbf{z}_j^{[k]}, X_j^{\mathbf{m_o}_j})}{\underbrace{P(X_j^{m[k]}|X_j^{\mathbf{m_o}_j})}_{*}}. \quad (11)\end{aligned}$$

In (11), $P(X_j^{m[k]}|\mathbf{z}_j^{[k]}, X_j^{\mathbf{m_o}_j})$ in the numerator can be easily calculated because all the parent nodes of $X_j^{m[k]}$ are given in the graphical model shown in **Figure 2**. However, $P(X_j^{m[k]}|X_j^{\mathbf{m_o}_j})$ in the denominator cannot be evaluated in a straightforward way. Again, a Monte Carlo method can be adopted, as follows:

$$\begin{aligned}(*) = P(X_j^{m[k]}|X_j^{\mathbf{m_o}_j}) &= \sum_{\mathbf{z}_j} P(X_j^{m[k]}|\mathbf{z}_j, X_j^{\mathbf{m_o}_j})P(\mathbf{z}_j|X_j^{\mathbf{m_o}_j})\\ &= \mathbb{E}_{\mathbf{z}_j|X_j^{\mathbf{m_o}_j}}[P(X_j^{m[k]}|\mathbf{z}_j, X_j^{\mathbf{m_o}_j})]\\ &\approx \frac{1}{K'}\sum_{k'} P(X_j^{m[k]}|\mathbf{z}_j^{[k']}, X_j^{\mathbf{m_o}_j}) \quad (12)\end{aligned}$$

where K' is the number of samples for the second Monte Carlo approximation. Fortunately, in this Monte Carlo approximation (12), we can reuse the samples drawn in the previous Monte Carlo approximation efficiently, i.e., $K' = K$. By substituting (12) for (11), we finally obtain the approximate IG for the criterion of active perception, i.e., our proposed method, as follows:

$$\mathrm{IG}(\mathbf{z}_j; X_j^m|X_j^{\mathbf{m_o}_j}) \approx \frac{1}{K}\sum_k \log \frac{P(X_j^{m[k]}|\mathbf{z}_j^{[k]}, X_j^{\mathbf{m_o}_j})}{\frac{1}{K}\sum_{k'} P(X_j^{m[k]}|\mathbf{z}_j^{[k']}, X_j^{\mathbf{m_o}_j})}.$$

Note that the computational cost for evaluating IG becomes $O(K^2)$. In summary, a robot can approximately estimate the IG for unobserved modality information by generating virtual observations based on observed data and evaluating their likelihood.

4.4. MHDP-Based Active Perception Methods

We propose the use of the *greedy* and *lazy greedy algorithms* for selecting L actions to recognize a target object on the basis of the submodular property of IG. The final greedy and lazy greedy algorithms for MHDP-based active perception, i.e., our proposed methods, are shown in Algorithms 1 and 2, respectively.

The main contribution of the lazy greedy algorithm is to reduce the computational cost of active perception. The majority of the computational cost originates from the number of times a robot evaluates IG_m for determining action sequences. When a robot has to choose L actions, the brute-force algorithm that directly evaluates all alternatives $\mathbf{A} \in \mathbf{F}_L^{\mathbf{m_o}_j}$ using (6) requires ${}_{\#(\mathbf{M}\setminus\mathbf{m_o}_j)}C_L$ evaluations of $\mathrm{IG}(\mathbf{z}_j; X_j^{\mathbf{A}}|X_j^{\mathbf{m_o}_j})$. In contrast, the greedy algorithm requires $\{\#(\mathbf{M} \setminus \mathbf{m_o}_j) + (\#(\mathbf{M} \setminus \mathbf{m_o}_j) - 1) + \ldots +$

Algorithm 1 Greedy algorithm.

```
Require: MHDP is trained using a training data set.
  The $j$-th object is found.
  $\mathbf{m}_{\mathbf{o}j}$ is initialized, and $X_j^{\mathbf{m}_{\mathbf{o}j}}$ is observed.
  for $l = 1$ to $L$ do
    for all $m \in \mathbf{M} \setminus \mathbf{m}_{\mathbf{o}j}$ do
      for $k = 1$ to $K$ do
        Draw
          $(z_j^{[k]}, X_j^{m[k]}) \sim P(\mathbf{z}_j, X_j^m | X_j^{\mathbf{m}_{\mathbf{o}j}})$
      end for
      $\mathrm{IG}_m \leftarrow \frac{1}{K} \sum_k \log \frac{P(X_j^{m[k]} | \mathbf{z}_j^{[k]}, X_j^{\mathbf{m}_{\mathbf{o}j}})}{\frac{1}{K} \sum_{k'} P(X_j^{m[k]} | \mathbf{z}_j^{[k']}, X_j^{\mathbf{m}_{\mathbf{o}j}})}$
    end for
    $m^* \leftarrow \underset{m}{\operatorname{argmax}}\, \mathrm{IG}_m$
    Execute the $m^*$-th action to the $j$-th target object and obtain $X_j^{m^*}$.
    $\mathbf{m}_{\mathbf{o}j} \leftarrow \mathbf{m}_{\mathbf{o}j} \cup \{m^*\}$
  end for
```

Algorithm 2 Lazy greedy algorithm.

```
Require: The MHDP is trained using a training data set.
  The $j$-th object is found.
  $\mathbf{m}_{\mathbf{o}j}$ is initialized, and $X_j^{\mathbf{m}_{\mathbf{o}j}}$ is observed.
  for all $m \in \mathbf{M} \setminus \mathbf{m}_{\mathbf{o}j}$ do
    for $k = 1$ to $K$ do
      Draw
        $(z_j^{[k]}, X_j^{m[k]}) \sim P(\mathbf{z}_j, X_j^m | X_j^{\mathbf{m}_{\mathbf{o}j}})$
    end for
    $\mathrm{IG}_m \leftarrow \frac{1}{K} \sum_k \log \frac{P(X_j^{m[k]} | \mathbf{z}_j^{[k]}, X_j^{\mathbf{m}_{\mathbf{o}j}})}{\frac{1}{K} \sum_{k'} P(X_j^{m[k]} | \mathbf{z}_j^{[k']}, X_j^{\mathbf{m}_{\mathbf{o}j}})}$
  end for
  $m^* \leftarrow \underset{m}{\operatorname{argmax}}\, \mathrm{IG}_m$
  Execute the $m^*$-th action to the $j$-th target object and obtain $X_j^{m^*}$.
  $\mathbf{m}_{\mathbf{o}j} \leftarrow \mathbf{m}_{\mathbf{o}j} \cup \{m^*\}$
  Prepare a stack $S$ for the modality indices and initialize it.
  for all $m \in \mathbf{M} \setminus \mathbf{m}_{\mathbf{o}j}$ do
    $push(S, (m, \mathrm{IG}_m))$
  end for
  for $l = 1$ to $L - 1$ do
    repeat
      $S \leftarrow descending\_sort(S)$ // w.r.t. $\mathrm{IG}_m$
      $(m^1, \mathrm{IG}_{m^1}) \leftarrow pop(S)$ , $(m^2, \mathrm{IG}_{m^2}) \leftarrow pop(S)$
      // Re-evaluate $\mathrm{IG}_{m^1}$ as follows.
      for $k = 1$ to $K$ do
        Draw
          $(z_j^{[k]}, X_j^{m^1[k]}) \sim P(\mathbf{z}_j, X_j^{m^1} | X_j^{\mathbf{m}_{\mathbf{o}j}})$
      end for
      $\mathrm{IG}_{m^1} \leftarrow \frac{1}{K} \sum_k \log \frac{P(X_j^{m^1[k]} | \mathbf{z}_j^{[k]}, X_j^{\mathbf{m}_{\mathbf{o}j}})}{\frac{1}{K} \sum_{k'} P(X_j^{m^1[k]} | \mathbf{z}_j^{[k']}, X_j^{\mathbf{m}_{\mathbf{o}j}})}$
      $push(S, (m^2, \mathrm{IG}_{m^2}))$, $push(S, (m^1, \mathrm{IG}_{m^1}))$
    until $\mathrm{IG}_{m^1} \geq \mathrm{IG}_{m^2}$
    $m^* \leftarrow m^1$
    $pop(S)$
    Execute the $m^*$-th action to the $j$-th target object and obtain $X_j^{m^*}$.
    $\mathbf{m}_{\mathbf{o}j} \leftarrow \mathbf{m}_{\mathbf{o}j} \cup \{m^*\}$
  end for
```

$(\#(\mathbf{M} \setminus \mathbf{m}_{\mathbf{o}j}) - L + 1)\}$ evaluations of $\mathrm{IG}(\mathbf{z}_j; X_j^m | X_j^{\mathbf{m}_{\mathbf{o}j}})$, i.e., $O(ML)$. The lazy greedy algorithm incurs the same computational cost as the greedy algorithm only in the worst case. However, practically, the number of re-evaluations in the lazy greedy algorithm is quite small. Therefore, the computational cost of the lazy greedy algorithm increases almost in proportion to L, i.e., almost linearly. The memory requirement of the proposed method is also quite small. Both the greedy and lazy greedy algorithms only require memory for IG_m for each modality and K samples for the Monte Carlo approximation. These requirements are negligibly small compared with the MHDP itself.

Note that the IG_m is not the exact IG, but an approximation. Therefore, the differences between IG and IG_m may harm the performance of greedy and lazy greedy algorithms to a certain extent. However, the algorithms are expected to work practically. We evaluated the algorithms through experiments.

5. EXPERIMENT 1: HUMANOID ROBOT

5.1. Conditions

An experiment using an upper-torso humanoid robot was conducted to verify the proposed active perception method in the real-world environment. In this experiment, RIC-Torso, developed by the RT Corporation, was used (see **Figure 3**). RIC-Torso is an upper-torso humanoid robot that has two robot hands. We prepared an experimental environment that is similar to the one in the original MHDP paper (Nakamura et al., 2011b). The robot has four available actions and four corresponding modality information. The set of modalities was $\mathbf{M} = \{m^v, m^{as}, m^{ah}, m^h\}$, which represent visual information, auditory information obtained by shaking an object, one by hitting an object and haptic information, respectively.

5.1.1. Visual Information (m^v)

Visual information was obtained from the Xtion PRO LIVE set on the head of the robot. The camera was regarded as the eyes of the robot. The robot captured 74 images of a target object while it rotated on a turntable (see **Figure 3**). The size of each image was re-sized to 320 × 240. Scale-invariant feature transform (SIFT) feature vectors were extracted from each captured image (Lowe, 2004). A certain number of 128-dimensional feature vectors were obtained from each image. Note that the SIFT feature did not consider hue information. All of the obtained feature vectors were transformed into BoF representations using k-means clustering with $k = 25$. The number of clusters k was determined empirically, considering prior works (Nakamura et al., 2011b; Araki et al., 2012). The k-means clustering was performed using data from all objects in a training set, and the centroids of the clusters were determined. BoF representations were used as observation data for the visual modality of the MHDP. The index for this modality was defined as m^v.

5.1.2. Auditory Information (m^{as} and m^{ah})

Auditory information was obtained from a multipowered shotgun microphone NTG-2 by RODE Microphone. The microphone was regarded as the ear of the robot. In this experiment, two types of auditory information were acquired. One was generated by hitting the object, and the other was generated by shaking it. The two sounds were regarded as different auditory information and hence different modality observations in the MHDP model. The two actions, i.e., hitting and shaking, were manually programmed for the robot. Each action was implemented as a fixed trajectory. When the robot began to execute an action, it also started recording the objects's sound (see **Figure 3**). The sound was recorded until two seconds after the robot finished the action. The recorded auditory data were temporally divided into frames, and each frame was transformed into 13-dimensional Mel-frequency cepstral coefficients (MFCCs). The MFCC feature vectors were transformed into BoF representations using k-means clustering with $k = 25$ in the same way as the visual information. The indices of these modalities were defined as m^{as} and m^{ah}, respectively, for "shake" and "hit."

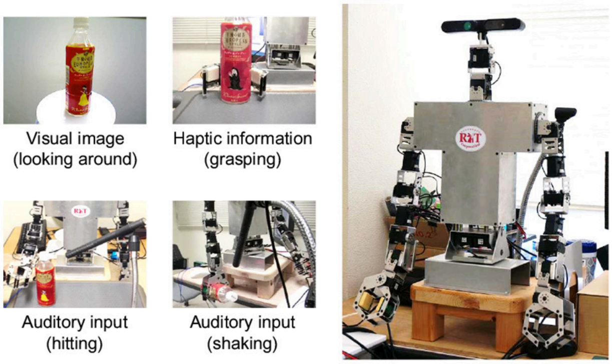

FIGURE 3 | A humanoid robot used in the experiment.

5.1.3. Haptic Information (m^h)

Haptic information was obtained by grasping a target object using the robot's hand. When the robot attempted to obtain haptic information from an object placed in front of it, it moved its hand to the object and gradually closed its hand until a certain amount of counterforce was detected (see **Figure 3**). The joint angle of the hand was measured when the hand touched the target object and when the hand stopped. The two variables and difference between the two angles were used as a three-dimensional feature vector. When obtaining haptic information, the robot grasped the target object 10 times and obtained 10 feature vectors. The feature vectors were transformed into BoF representations using k-means clustering with $k = 5$ in the same way as for the other information types. The index of the haptic modality was defined as m^h.

5.1.4. Multimodal Information as BoF Representations

In summary, a robot could obtain multimodal information from four modalities for perception. The dimensions of the BoFs were set to 25, 25, 25, and 5 for m^v, m^{as}, m^{ah}, and m^h, respectively. The dimension of each BoF corresponds to the number of clusters for k-means clustering. The numbers of clusters, i.e., the sizes of the dictionaries, were empirically determined on the basis of a preliminary experiment on multimodal categorization. All of the training datasets were used to train the dictionaries. The histograms of the feature vectors, i.e., the BoFs, were resampled to make their counts $N_j^{m^v} = 100, N_j^{m^{as}} = 80, N_j^{m^{ah}} = 130$, and $N_j^{m^h} = 30$. The weight of each modality w^m was set to 1. The formation of multimodal object categories itself is out of the scope of this paper. Therefore, the constants were empirically determined so that the robot could form object categories that are similar to human participants. The number of samples K in the Monte Carlo approximation for estimating IG was set to $K = 5,000$. The constant K was determined empirically. The effect of K will be examined in the experiment as well (see **Figure 11**).

5.1.5. Target Objects

For the target objects, 17 types of commodities were prepared for the experiment shown in **Figure 4**. An object was provided for obtaining a training data, i.e., data for object categorization, and another object was provided for obtaining test data, i.e., data for active perception, for each type of objects. Each index on the right-hand side of the figure indicates the index of each object. The hardness of the balls, the striking sounds of the cups, and the sounds made while shaking the bottles were different depending on the object categories. Therefore, ground-truth categorization could not be achieved using visual information alone.

5.2. Procedure

The experimental procedure was as follows. First, the robot formed object categories through multimodal categorization in an unsupervised manner. An experimenter placed each object

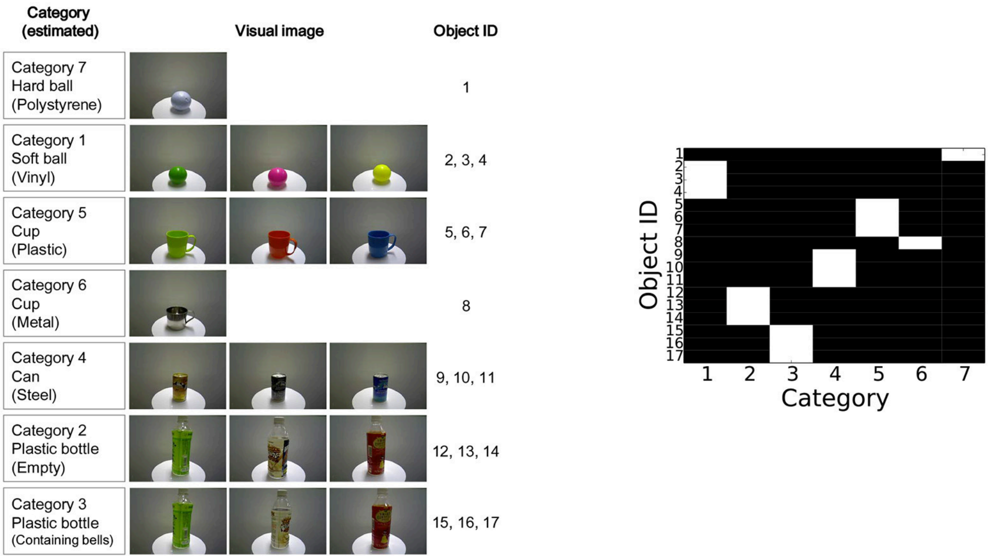

FIGURE 4 | (Left) target objects used in the experiment and **(Right)** categorization results obtained in the experiment.

in front of the robot one by one. In this training phase, two objects for each type of objects were provided. The robot looked at the object to obtain visual features, grasped it to obtain haptic features, shook it to obtain auditory shaking features, and hit it to obtain the auditory striking features. After obtaining the multimodal information of the objects as a training data set, the MHDP was trained using a Gibbs sampler. The results of multimodal categorization are shown in **Figure 4**. The category that has the highest posterior probability for each object is shown in white. These results show that the robot can form multimodal object categories using MHDP, as described in Nakamura et al. (2011b). After the robot had formed object categories, we fixed the latent variables for the training data set[3].

Second, an experimental procedure for active perception was conducted. An experimenter placed an object in front of the robot. The robot observed the object using its camera, obtained visual information, and set $\mathbf{m}_{\mathbf{o}j} = \{m^v\}$. An object was provided for each type of objects shown in **Figure 4** to the robot one by one. Therefore, 17 objects were used for evaluating each active perception strategy. The sequential action selection and object recognition were performed once per an object. At each step of the sequential action selection, Gibbs sampler for MHDP was performed and it updated its latent variables, i.e., recognition state, of the MHDP. The robot then determined its next set of actions for recognizing the target object using its active perception strategy shown in Algorithms 1 and 2.

5.3. Results

5.3.1. Selecting the Next Action

First, we describe results for the first single action selection after obtaining visual information. In this experiment, the robot had three choices for its next action, i.e., m^{as}, m^{ah}, and m^h. To evaluate the results of active perception, we used $\mathrm{KL}\left(P(k|X_j^{\mathbf{M}}), P(k|X_j^{\mathbf{A}\cup\mathbf{m}_{\mathbf{o}j}})\right)$, i.e., the distance between the posterior distribution over the object categories k in the final recognition state and that in the next recognition state as an evaluation criterion on behalf of $\mathrm{KL}\left(P(\mathbf{z}_j|X_j^{\mathbf{M}}), P(\mathbf{z}_j|X_j^{\mathbf{A}\cup\mathbf{m}_{\mathbf{o}j}})\right)$, which is the original evaluation criterion in (4). The computational cost for numerical evaluation of $\mathrm{KL}\left(P(\mathbf{z}_j|X_j^{\mathbf{M}}), P(\mathbf{z}_j|X_j^{\mathbf{A}\cup\mathbf{m}_{\mathbf{o}j}})\right)$ using a Monte Carlo method is too high because $\mathbf{z}_j = \{\{k_{jt}\}_{1\leq t\leq T_j}, \{t^m_{jn}\}_{m\in\mathbf{M},1\leq n\leq N^m_j}\}$ has so many variables and a posterior distributions over $\mathbf{z}_j$ is very complex.

Figure 5 (Top) shows samples of the KL divergence between the posterior probabilities of the category after obtaining the information from all modalities and after obtaining only visual information.

With regard to some objects, e.g., objects 6 and 7, the figure shows samples of that visual information seems to be sufficient for the robot to recognize the objects as compared the other objects[4]. However, with regard to many objects, visual information alone could not lead the recognition state to the final state. However, it could be reached using the information of all modalities. **Figure 5** (Middle) shows samples of IG_m calculated using the visual information for each action. **Figure 5** (Bottom) shows the KL divergence between the final recognition state and the posterior probability estimated after obtaining visual information and the information of each selected action. We observe that an action with a higher value of IG_m tended to further reduce the KL divergence, as Theorem 1

[3] The collected datasets for this experiment can be found in GitHub: https://github.com/tanichu/data-active-perception-hmdp

[4] Note that currently we don't have a good criteria of KL divergence to determine whether performing further actions are necessary or not.

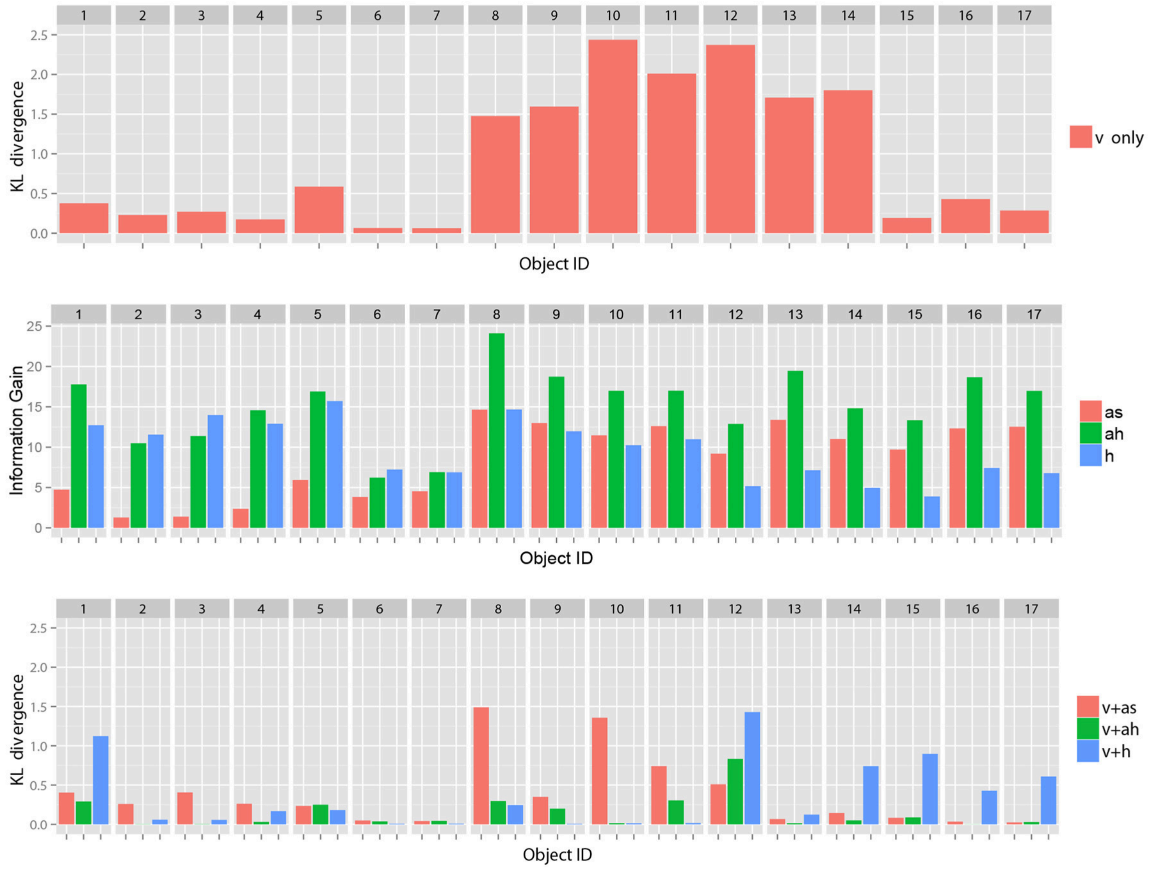

FIGURE 5 | (Top) Samples of KL divergence between the final recognition state and the posterior probability estimated after obtaining only visual information, **(Middle)** samples of estimated IG_m for each object based on visual information (v), and **(Bottom)** samples of KL divergence between the final recognition state and the posterior probability estimated after obtaining only visual information and each selected action where as, ah, h represent represent auditory information obtained by shaking an object, one by hitting an object and haptic information, respectively. Our theory of multimodal active perception suggests that the action with the highest information gain (shown in the middle) tends to lead its initial recognition state (whose KL divergence from the final recognition state is shown at the top) to a recognition state whose KL divergence from the final recognition state (shown at the bottom) is the smallest. These figures suggest the probabilistic relationships were satisfied as a whole.

suggests. **Figure 6** shows the average KL divergence for the final recognition state after executing an action selected by the IG_m criterion. Actions IG .min, IG .mid, and IG .max denote actions that have the minimum, middle, and maximum values of IG_m, respectively. These results show that IG .max clearly reduced the uncertainty of the target objects.

The precision of category recognition after an action execution is summarized in **Table 1**. Basically, a category recognition result is obtained as the posterior distribution (3) in the MHDP. The category with the highest posterior probability is considered to be the recognition result for illustrative purposes in **Table 1**. Obtaining information by executing IG .max almost always increased recognition performance.

Examples of changes in the posterior distribution are shown in **Figure 7** (Left, Right) for objects 8 ("metal cup") and 12 ("empty plastic bottle"), respectively. The robot could not clearly recognize the category of object 8 after obtaining visual information. Action IG_m in **Figure 5** shows that m^{ah} was IG .max for the 8th object. **Figure 7** (Left) shows that m^{ah} reduced the uncertainty and allowed the robot to correctly recognize the object, as evidenced by category 6, a metal cup. This means that the robot noticed that the target object was a metal cup by hitting it and listening to its metallic sound. The metal cup did not make a sound when the robot shook it. Therefore, the IG for m^{as} was small. As **Figure 7** (Right) shows, the robot first recognized the 12th object as a plastic bottle containing bells with high probability and as an empty plastic bottle with a low probability. **Figure 5** shows that the IG_m criterion suggested m^{ah} as the first alternative and m^{as} as the second alternative. **Figure 7** (Right) shows that m^{as} and m^{ah} could determine that the target object was an empty plastic bottle, but m^{h} could not.

As humans, we would expect to differentiate an empty bottle from a bottle containing bells by shaking or hitting the bottle, and differentiate a metal cup from a plastic cup by hitting it. The proposed active perception method constructively reproduced this behavior in a robotic system

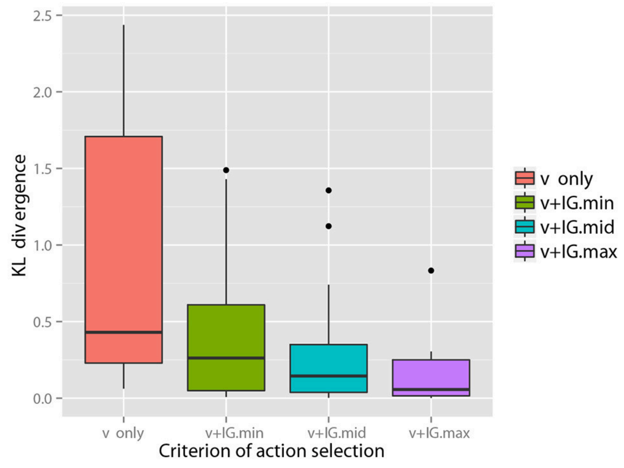

FIGURE 6 | Reduction in the KL divergence by executing an action selected on the basis of the IG_m maximization criterion. The KL divergences between the recognition state after executing the second action and the final recognition state are calculated for all objects and shown with box plot. This shows that an action with more information brings the recognition of its state closer to the final recognition state.

TABLE 1 | Number of successfully recognized objects.

v only	v+IG.min	v+IG.mid	v+IG.max	Full information
8/17	11/17	15/17	16/17	17/17

using an unsupervised multimodal machine learning approach.

5.3.2. Selecting the Next Set of Multiple Actions

We evaluated the greedy and lazy greedy algorithms for active perception sequential decision making. The KL divergence from the final state for all target objects is averaged at each step and shown in **Figure 8**. For each condition, the KL divergence gradually decreased and reached almost zero. However, the rate of decrease notably differed. As the theory of submodular optimization suggests, the greedy algorithm was shown to be a better solution on average and slightly worse than the best case (Nemhauser et al., 1978). The best and worst cases were selected after all types of sequential actions had been performed. The "average" is the average of the KL divergence obtained by all possible types of sequential actions. The results for the lazy greedy algorithm were almost same as those of the greedy algorithm, as Minoux (1978) suggested.

The sequential behaviors of IG_m were observed to determine if their behaviors were consistent with our theories. For example, the changes in IG_m at each step as the robot sequentially selected its action to perform on object 10 using the greedy algorithm is shown in **Figure 9**. Theorem 2 shows that the IG is a submodular function. This predicts that IG_m decreases monotonically when a new action is executed in active perception. When the robot obtained only visual information (v only in **Figure 9**), all values of IG_m were still large. After m^{ah} was executed on the basis of the greedy algorithm, $\mathrm{IG}_{m^{ah}}$ became zero. At the same time, $\mathrm{IG}_{m^{as}}$ and $\mathrm{IG}_{m^{h}}$ decreased. In the same way, all values of IG_m gradually decreased monotonically.

Figure 10 shows the time series of the posterior probability of the category for object 10 during sequential active perception. Using only visual information, the robot misclassified the target object as a plastic bottle containing bells (category 3). The action sequence in reverse order did not allow the robot to recognize the object as a steel can at its first step and change its recognition state to an empty plastic bottle (category 4). After the second action, i.e., grasping (m^h), the robot recognized the object as a steel can. In contrast, the greedy algorithm could determine that the target object was in category 4, i.e., steel can, with its first action.

The effect of the number of samples K for the Monte Carlo approximation was observed. **Figure 11** shows the relation between K and the standard deviation of the estimated IG_m for the 15th object for each action after obtaining a visual image. This figure shows that estimation error gradually decreases when K increases. Roughly speaking, $K \geq 1,000$ seems to be required for an appropriate estimate of IG_m in our experimental setting. Evaluation of IG_m required less than 1 second, which is far shorter than the time required for action execution by a robot. This means that our method can be used in a real-time manner.

These empirical results show that the proposed method for active perception allowed a robot to select appropriate actions sequentially to recognize an object in the real-world environment and in a real-time manner. It was shown that the theoretical results were supported, even in the real-world environment.

6. EXPERIMENT 2: SYNTHETIC DATA

In experiment 1, the numbers of classes, actions, and modalities as well as the size of dataset were limited. In addition, it was difficult to control the robotic experimental settings so as to check some interesting theoretical properties of our proposed method. Therefore, we performed a supplemental experiment, Experiment 2, using synthetic data comprising 21 object types, 63 objects, and 20 actions, i.e., modalities.

First, we checked the validity of our active perception method when the number of types of actions increases. Second, we checked how the method worked when two classes were assigned to the same object. Although the MHDP can categorize an object into two or more categories in a probabilistic manner, each object was classified into a single category in the previous experiment.

6.1. Conditions

A synthetic dataset was generated using the generative model that the MHDP assumes (see **Figure 2**). We prepared 21 virtual object classes, and three objects were generated from each object class, i.e., we obtained 63 objects in total. Among the object classes, 14 object classes are "pure," and seven object classes are "mixed." For each pure object class, a multinomial distribution was drawn from the Dirichlet distribution corresponding to each modality. We set the number of modalities $M = 20$. The hyperparameters of the Dirichlet distributions of the modalities were set to $\alpha_0^m =$

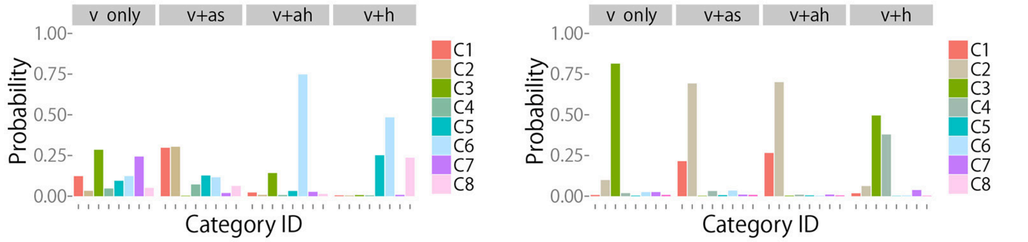

FIGURE 7 | (Left) Posterior probability of the category for object 8 after executing each action. These results show that the action with the highest information gain, i.e., *ah*, allowed the robot to efficiently estimate that the true object category was "metal cup". **(Right)** Posterior probability of the category for object 12 after executing each action. These results show that the actions with the highest and second highest information gain, i.e., *ah* and *as*, allowed the robot to efficiently estimate that the true object category was "empty plastic bottle".

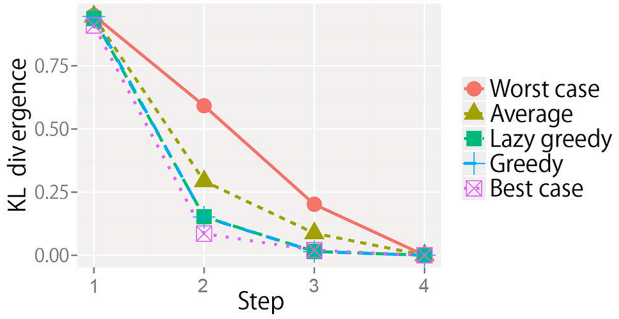

FIGURE 8 | KL divergence from the final state at each step for each sequential action selection procedure. Note that the line of the lazy greedy algorithm is overlapped by that of the greedy algorithm.

0.4(m − 1) for $m > 1$. For $m = 1$, we set $\alpha_0^1 = 10$. For each mixed object class, a multinomial distribution for each modality was prepared by mixing the distributions of the two pure object classes. Specifically, the multinomial distribution for the i-th mixed object was obtained by averaging those of the $(2i - 1)$-th and the $2i$-th object classes. The observations for each modality of each object were drawn from the multinomial distributions corresponding to the object's class. The count of the BoFs for each modality was set to 20. Finally, 42 pure virtual objects and 21 mixed virtual objects were generated.

The experiment was performed almost in the same way as experiment 1. First, multimodal categorization was performed for the 63 virtual objects, and 14 categories were successfully formed in an unsupervised manner. The posterior distributions over the object categories are shown in **Figure 12**. Generally speaking, mixed objects were categorized into two or more classes. After categorization, a virtual robot was asked to recognize all of the target objects using the proposed active perception method.

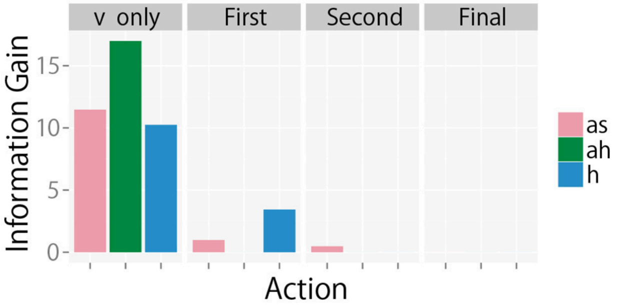

FIGURE 9 | IG$_m$ at each step for object 10 when the greedy algorithm is used.

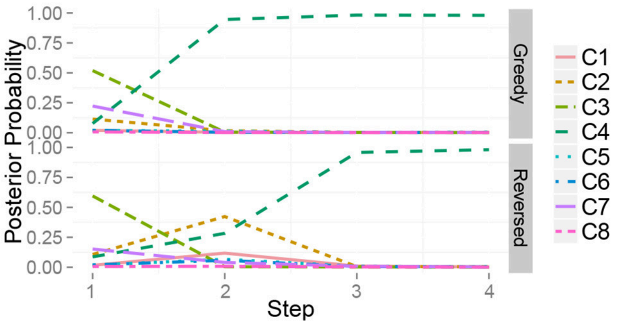

FIGURE 10 | Time series of the posterior probability of the category for object 10 during sequential action selection based on **(top)** the greedy algorithm, i.e., $m^{ah} \rightarrow m^{h} \rightarrow m^{as}$, and **(bottom)** its reverse order , i.e., $m^{as} \rightarrow m^{h} \rightarrow m^{ah}$.

6.2. Results

We compared the greedy, lazy greedy, and random algorithms for the active perception sequential decision making process. The random algorithm is a baseline method that determines the next action randomly from the remaining actions that have not been taken. In other words, the random algorithm is the case in which a robot does not employ any active perception algorithms.

The KL divergence from the final state for all target objects is averaged at each step and shown in **Figure 13**. For each condition, the KL divergence gradually decreased and reached almost zero. However, the rate of decrease was different. The greedy and lazy greedy algorithms were clearly shown to be better solutions on average than the random algorithm. In contrast with experiment 1, the best and worst cases could not practically be calculated because of the prohibitive computational cost.

Interestingly, the lazy greedy algorithm has almost the same performance as the greedy algorithm, as the theory suggests, although the laziness reduced the computational cost in reality.

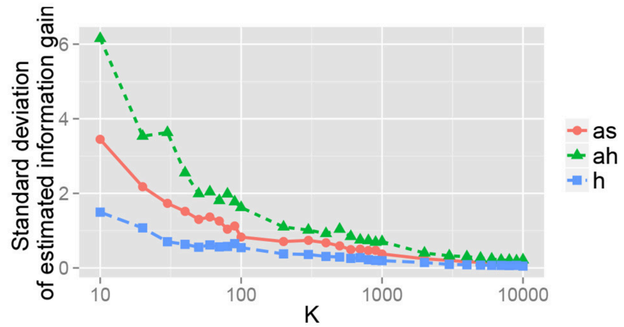

FIGURE 11 | Standard deviation of the estimated information gain IG_m for the 15th object. For each K, 100 values of the estimated information gain IG_m were obtained, and their standard deviation is shown.

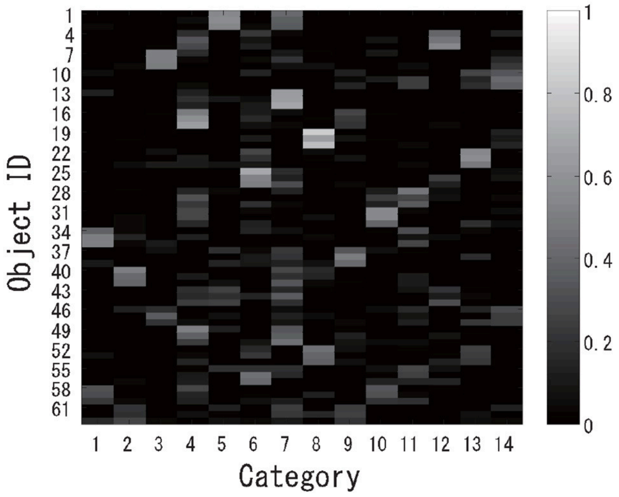

FIGURE 12 | Categorization results for the posterior probability distributions for each object.

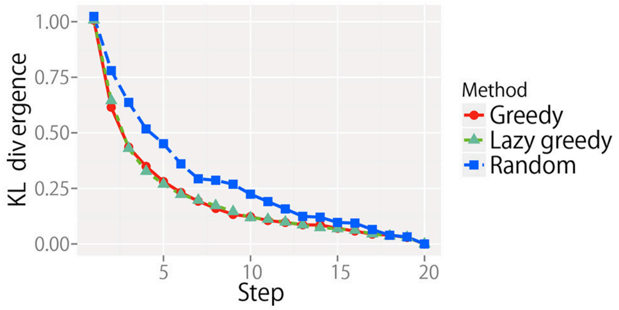

FIGURE 13 | KL divergence from the final state at each step for each sequential action selection procedure.

The number of times the robot evaluated IG_m to determine the action sequences for all executable counts of actions $L = 1, 2, \ldots, M$ is summarized for each method. The number of times the lazy greedy algorithm was required for each target object was 71.7 ($SD = 5.2$) on average, and that of the greedy algorithm was 190. Theoretically, the greedy and lazy greedy algorithms require $O(M^2)$ evaluations. Practically, the number of re-evaluations needed by the lazy greedy algorithm is quite small. In contrast, the brute-force algorithm requires $O(2^M)$ evaluations, i.e., far more evaluations of IG are required.

Next, a case in which two classes were assigned to the same object was investigated. The target dataset contained "mixed" objects. The results also imply that our method works well even when two classes are assigned to the same object. This is because our theory is completely derived on the basis of the probabilistic generative model, i.e., the MHDP. We show a typical result. **Figure 14** shows the time series of the posterior probability of the category for object 51, i.e., one of the mixed objects, during sequential active perception. This shows that the greedy and lazy greedy algorithms quickly categorized the target object into two categories "correctly." Our formulation assumes the categorization result to be a posterior distribution. Therefore, this type of probabilistic case can be treated naturally.

7. CONCLUSION AND DISCUSSION

In this paper, we described an MHDP-based active perception method for robotic multimodal object category recognition. We formulated a new active perception method on the basis of the MHDP (Nakamura et al., 2011b) .

First, we proposed an action selection method based on the IG criterion and showed that IG is an optimal criterion for active perception from the viewpoint of reducing the expected

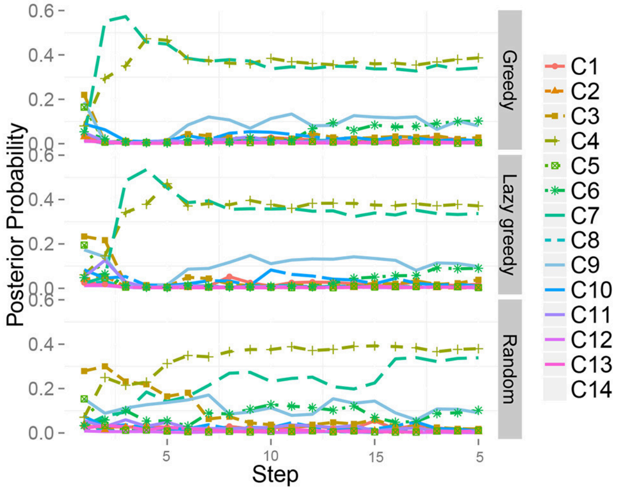

FIGURE 14 | Time series of the posterior probability of the category for object 51 during sequential action selection based on **(Top)** the greedy algorithm, **(Middle)** the lazy greedy algorithm, and **(Bottom)** the random selection procedure.

KL divergence between the final and current recognition states. Second, we proved that the IG has a submodular property and reduced the sequential active perception problem to a submodular maximization problem. Third, we derived a Monte Carlo approximation method for evaluating IG efficiently and made the action selection method executable. Given the theoretical results, we proposed to use the greedy and lazy greedy algorithms for selecting a set of actions for active perception. It is important to note that all of the three theoretical contributions mentioned above were naturally derived from the characteristics of the MHDP. These contributions are clearly a result of the theoretical soundness of the MHDP. In this sense, our theorems reveal a new advantage of the MHDP that other several heuristic multimodal object categorization methods do not have.

To evaluate the proposed methods empirically, we conducted experiments using an upper-torso humanoid robot and a synthetic dataset. Our results showed that the method enables the robot to actively select actions and recognize target objects quickly and accurately.

One of the most interesting points of this paper is that not only object categories but also an action selection for object recognition can be formed in an unsupervised manner. From the viewpoint of cognitive developmental robotics, providing an unsupervised learning model for bridging the development between perceptual and action systems is meaningful for shedding a new light on the computational understanding of cognitive development (Asada et al., 2009; Cangelosi and Schlesinger, 2015). It is believed that the coupling of action and perception is important for an embodied cognitive system (Pfeifer and Scheier, 2001).

The advantage of this paper compared with the related works in robotics is that our action selection method for multimodal category recognition has a clear theoretical basis and is tightly connected to the computational model for multimodal object categorization, i.e., MHDP. The theoretical basis gives the method preferable characteristics, i.e., theoretical guarantee.

However, note that the theoretical guarantee is satisfied only when IG is correctly estimated. We assumed that outcome of each action is deterministic and fully observable when we apply the theory of submodular optimization to active perception in multimodal categorization. However, observations X^m and IG are measured somehow probabilistically because of real-world uncertainty and Monte Carlo approximation. For example, IG is approximately estimated at each step of the greedy and lazy greedy algorithms. Theoretically, given this approximation in evaluating the objective being maximized, the $(1 - 1/e)$ bound no longer holds. Streeter et al. proposed to introduce an additional penalty based on a function approximation (Streeter and Golovin, 2009). Golovin et al. extended submodularity to adaptive submodularity to consider stochastic property (Golovin and Krause, 2011). Though we discussed the proposed method from the viewpoint of submodular optimization, this algorithm can be regarded as a version of the sequential information maximization, more specifically (Chen et al., 2015). Extending our idea by referring the adaptive submodularity and/or the sequential information maximization, and update our method is our future challenge.

We assumed that each action requires same cost, and tried to reduce the number of actions in active perception, i.e., to maximize the performance of perception with the fixed number of actions. However, practically, each action, e.g., shake, hit and look at, requires different duration and different energy. Therefore, practical cost is not always the number of actions, but total cost of actions. Zhang et al. (2017) tried to deal with this problem in the context of multimodal object identification. This problem leads us a knapsack problem-like formulation. This type of submodular optimization has been studied by many researchers (Streeter and Golovin, 2009; Zhou et al., 2013). Our method will be able to be extended in the similar way.

In addition to active perception, active "learning/exploration" for multimodal categorization is also an important research topic. It takes a longer time for a robot to gather multimodal information to form multimodal object categories from a massive number of daily objects than it does to recognize a new object. If a robot can notice that "the object is obviously a sample of learned category," the robot need not obtain knowledge about object categories from such an object. In contrast, if a target object appears to be completely new to the robot, the robot should carefully interact with the object to obtain multimodal information from the object. Such a scenario will be achieved by developing an active "learning/exploration" method for multimodal categorization. It is likely that such a method will be able to be obtained by extending our proposed active perception method.

Considering more complex categorization scenario is our future challenge. For example, Schenck et al. (2014) is dealing with the more complex categorization scenario, i.e., 36 plastic containers with identical shape and 3 colors, 4 types of contents, and 3 different amounts of those contents. In this paper, we used MHDP which assumes an object is classified into a single object category and infers the posterior distribution over categories. When we consider human cognition, we can find that object categories have more complex characteristics. For example, object categories have a hierarchical structure, an object is categorized into several classes, and they have different modality-dependency based on the types of categories. Unsupervised machine learning methods for such complex categorization problem have proposed by several researchers based on hierarchical Bayesian models (Griffiths and Ghahramani, 2006; Ando et al., 2013; Nakamura et al., 2015). Theoretically, the main assumption we used was that the MHDP is a hierarchical Bayesian model and action selection is corresponding to obtaining an observation which is a probabilistic variable on the leaf node of its graphical model. Therefore, by applying the same idea to the more complex categorization methods, we will be able to extend our theory to more complex categorization problems. This is on of our future works.

Another challenge lies in feature representation for multimodal categorization. The MHDP assumed that observations are given as bag-of-features representations. However, there are many kinds of feature representations for visual, auditory and haptic information. In particular, the feature extraction capability of deep neural networks is gathering attention, recently. Theoretically, our main theorems do not depend on the type of emission distributions, i.e., bag-of-features representations. It is likely that the same approach can be used even when a multimodal categorization method uses different feature representations, e.g., the features in the last hidden layer of a pre-trained deep neural network. This extension is also a part of our future challenges.

In addition, the MHDP model treated in this paper assumed that an action for perception is related to only one modality, e.g., grasping only corresponds to m^h. However, in reality, when we interact with an object with a specific action, e.g., grasping, shaking, or hitting, we obtain rich information related to various modalities. For example, when we shake a box to obtain auditory information, we also unwittingly obtain haptic information and information about its weight. The tight linkage between the modality information and an action is a type of approximation taken in this research. An extension of our model and the MHDP to a model that can treat actions that are related to various modalities is also a task for our future work.

AUTHOR CONTRIBUTIONS

The main theory was developed by TaT. The experiments were conceived by RY. The data were analyzed by RY and ToT with help of TaT. The manuscript was written by TaT.

FUNDING

This research was partially supported by Tateishi Science and Technology Foundation, and JST, CREST. This was also partially supported by a Grant-in-Aid for Scientific Research on Innovative Areas (16H06569) and a Grant-in-Aid for Young Scientists (B) (24700233) funded by the Ministry of Education, Culture, Sports, Science, and Technology.

ACKNOWLEDGMENTS

The authors would like to thank undergraduate student Takuya Takeshita and graduate student Hajime Fukuda of Ritsumeikan University, who helped us develop the experimental instruments for obtaining our preliminary results.

REFERENCES

Ando, Y., Nakamura, T., Araki, T., and Nagai, T. (2013). "Formation of hierarchical object concept using hierarchical latent dirichlet allocation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems* (Tokyo), 2272–2279.

Araki, T., Nakamura, T., Nagai, T., Nagasaka, S., Taniguchi, T., and Iwahashi, N. (2012). "Online learning of concepts and words using multimodal LDA and hierarchical Pitman-Yor Language Model," in *IEEE/RSJ International Conference on Intelligent Robots and Systems* (Algarve), 1623–1630.

Asada, M., Hosoda, K., Kuniyoshi, Y., Ishiguro, H., Inui, T., Yoshikawa, Y., et al. (2009). Cognitive Developmental Robotics: A Survey. *IEEE Trans. Auton. Mental Develop.* 1, 12–34. doi: 10.1109/TAMD.2009.2021702

Barsalou, L. W. (1999). Perceptual symbol systems. *Behav. Brain Sci.* 22, 1–16.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022.

Borotschnig, H., Paletta, L., Prantl, M., and Pinz, A. (2000). Appearance-based active object recognition. *Image Vision Comput.* 18, 715–727. doi: 10.1016/S0262-8856(99)00075-X

Burgard, W., Fox, D., and Thrun, S. (1997). "Active obile robot localization," in *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI)* (Nagoya), 1346–1352.

Cangelosi, A., and Schlesinger, M. (2015). *Developmental Robotics*. Cambridge, MA: The MIT press.

Celikkanat, H., Orhan, G., Pugeault, N., Guerin, F., Erol, S., and Kalkan, S. (2014). "Learning and Using Context on a Humanoid Robot Using Latent Dirichlet Allocation," in *Joint IEEE International Conferences on Development and Learning and Epigenetic Robotics (ICDL-Epirob)* (Genoa), 201–207.

Chen, Y., Hassani, S. H., Karbasi, A., and Krause, A. (2015). "Sequential information maximization: When is greedy near-optimal?" in *Conference on Learning Theory* (Paris), 338–363.

Cohn, D. A., Ghahramani, Z., and Jordan, M. I. (1996). Active learning with statistical models. *J. Artif. Intell. Res.* 4, 129–145.

Correa, J., and Soto, A. (2009). Active Visual Perception for Mobile Robot Localization. *J. Intell. Robot. Sys.* 58, 339–354. doi: 10.1007/s10846-009-9348-4

Denzler, J., and Brown, C. M. (2002). Information Theoretic Sensor Data Selection for Active Object Recognition and State Estimation. *IEEE Trans. Patt. Anal. Mach. Intell.* 24, 1–13. doi: 10.1109/34.982896

Dutta Roy, S., Chaudhury, S., and Banerjee, S. (2004). Active recognition through next view planning: a survey. *Patt. Recogn.* 37, 429–446. doi: 10.1016/j.patcog.2003.01.002

Eidenberger, R., and Scharinger, J. (2010). "Active perception and scene modeling by planning with probabilistic 6D object poses," in *IEEE/RSJ International Conference on Intelligent Robots and Systems* (Taipei), 1036–1043.

Ferreira, J., Lobo, J., Bessiere, P., Castelo-Branco, M., and Dias, J. (2013). A Bayesian framework for active artificial perception. *IEEE Trans. Cyber.* 43, 699–711. doi: 10.1109/TSMCB.2012.2214477

Fishel, J. A. and Loeb, G. E. (2012). Bayesian exploration for intelligent identification of textures. *Front. Neurorobot.* 6, 1–20. doi: 10.3389/fnbot.2012.00004

Golovin, D., and Krause, A. (2011). Adaptive submodularity: theory and applications in active learning and stochastic optimization. *J. Artif. Intell. Res.* 42, 427–486. doi: 10.1613/jair.3278

Gouko, M., Kobayashi, Y., and Kim, C. H. (2013). "Online exploratory behavior acquisition of mobile robot based on reinforcement learning," in *26th International Conference on Industrial Engineering and Other Applications of Applied Intelligence Systems, IEA/AIE 2013* (Amsterdam), 272–281.

Griffith, S., Sinapov, J., Sukhoy, V., and Stoytchev, A. (2012). A behavior-grounded approach to forming object categories: Separating containers from noncontainers. *IEEE Trans. Auton. Mental Develop.* 4, 54–69. doi: 10.1109/TAMD.2011.2157504

Griffiths, T. L., and Ghahramani, Z. (2006). "Infinite latent feature models and the indian buffet process," in *Advances in Neural Information Processing Systems 2006* (Vancouver, BC), 475–482.

Harnad, S. (1990). The symbol grounding problem. *Phys. D* 42, 335–346.

Hogman, V., Bjorkman, M., and Kragic, D. (2013). "Interactive object classification using sensorimotor contingencies," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Tokyo), 2799–2805.

Ivaldi, S., Nguyen, S. M., Lyubova, N., Droniou, A., Padois, V., Filliat, D., et al. (2014). Object learning through active exploration. *IEEE Trans. Auton. Mental Develop.* 6, 56–72. doi: 10.1109/TAMD.2013.2280614

Iwahashi, N., Sugiura, K., Taguchi, R., Nagai, T., and Taniguchi, T. (2010). "Robots that learn to communicate: a developmental approach to personally and physically situated human-robot conversations," in *Dialog with Robots Papers from the AAAI Fall Symposium* (Palo Alto, CA), 38–43.

Ji, S., and Carin, L. (2006). Cost-Sensitive Feature Acquisition and Classification. *Patt. Recogn.* 40, 1474–1485. doi: 10.1016/j.patcog.2006.11.008

Kemp, C., Chang, K. M., and Lombardi, L. (2010). Category and feature identification. *Acta Psychol.* 133, 216–233. doi: 10.1016/j.actpsy.2009.11.012

Krainin, M., Curless, B., and Fox, D. (2011). "Autonomous generation of complete 3D object models using next best view manipulation planning," in *IEEE International Conference on Robotics and Automation* (Shanghai), 5031–5037.

Krause, A., and Guestrin, C. E. (2005). "Near-optimal nonmyopic alue of information in graphical models," in *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence* (Edinburgh).

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* 60, 91–110. doi: 10.1023/B:VISI.0000029664.99615.94

MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge, UK: Cambridge University Press.

Minoux, M. (1978). "Accelerated greedy algorithms for maximizing submodular set functions," in *Optimization Techniques*, ed J. Stoer (Berlin: Springer), 234–243.

Muslea, I., Minton, S., and Knoblock, C. A. (2006). Active learning with multiple views. *J. Art. Intell. Res.* 27, 203–233. doi: 10.1613/jair.2005

Nakamura, T., Ando, Y., Nagai, T., and Kaneko, M. (2015). "Concept formation by robots using an infinite mixture of models," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Hamburg), 4593–4599.

Nakamura, T., Nagai, T., Funakoshi, K., Nagasaka, S., Taniguchi, T., and Iwahashi, N. (2014). "Mutual learning of an object oncept and language model based on MLDA and NPYLM," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'14)* (Chicago, IL), 600–607.

Nakamura, T., Nagai, T., and Iwahashi, N. (2007). "Multimodal object categorization by a robot," in *IEEE/RSJ International Conference on Intelligent Robots and Systems* (San Diego, CA), 2415–2420.

Nakamura, T., Nagai, T., and Iwahashi, N. (2009). "Grounding of word meanings in multimodal concepts using LDA," in *IEEE/RSJ International Conference on Intelligent Robots and Systems* (St. Louis, MO), 3943–3948.

Nakamura, T., Nagai, T., and Iwahashi, N. (2011a). "Bag of multimodal LDA models for concept formation," in *IEEE International Conference on Robotics and Automation* (Shanghai), 6233–6238.

Nakamura, T., Nagai, T., and Iwahashi, N. (2011b). "Multimodal categorization by hierarchical dirichlet process," in *IEEE/RSJ International Conference on Intelligent Robots and Systems* (San Francisco, CA), 1520–1525.

Natale, L., Metta, G., and Sandini, G. (2004). "Learning haptic representation of objects," in *International Conference of Intelligent Manipulation and Grasping* (Genoa).

Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L. (1978). An analysis of approximations for maximizing submodular set functions-I. *Math. Program.* 14, 265–294.

Pape, L., Oddo, C. M., Controzzi, M., Cipriani, C., Förster, A., Carrozza, M. C., et al. (2012). Learning tactile skills through curious exploration. *Front. Neurorobot.* 6:6. doi: 10.3389/fnbot.2012.00006

Pfeifer, R., and Scheier, C. (2001). *Understanding Intelligence*. A Bradford Book. Cambridge, MA: MIT Press.

Rebguns, A., Ford, D., and Fasel, I. (2011). "InfoMax control for acoustic exploration of objects by a mobile robot," in *AAAI11 Workshop on Lifelong Learning* (San Francisco, CA), 22–28.

Roy, D. K., and Pentland, A. P. (2002). Learning words from sights and sounds: a computational model. *Cogn. Sci.* 26, 113–146. doi: 10.1207/s15516709cog2601_4

Roy, N., and Thrun, S. (1999). "Coastal navigation with mobile robots," in *Advances in Neural Processing Systems 12*. Cambridge, MA: The MIT Press.

Russo, D., and Van Roy, B. (2016). An information-theoretic analysis of thompson sampling. *J. Mach. Learn. Res.* 17, 2442–2471. Available online at: http://jmlr.org/papers/v17/14-087.html

Saegusa, R., Natale, L., Metta, G., and Sandini, G. (2011). "Cognitive Robotics - Active Perception of the Self and Others," in *The 4th International Conference on Human System Interactions (HSI)* (Yokohama), 419–426.

Schenck, C., Sinapov, J., Johnston, D., and Stoytchev, A. (2014). Which object fits best? solving matrix completion tasks with a humanoid robot. *IEEE Trans. Auton. Mental Develop.* 6, 226–240. doi: 10.1109/TAMD.2014.2325822

Schneider, A., Sturm, J., Stachniss, C., Reisert, M., Burkhardt, H., and Burgard, W. (2009). "Object identification with tactile sensors using bag-of-features," in *IEEE/RSJ International Conference on Intelligent Robots and Systems* (St. Louis, MO), 243–248.

Settles, B. (2012). Active learning. *Synth. Lect. Artif. Intell. Mach. Learn.* 6, 1–114. doi: 10.2200/S00429ED1V01Y201207AIM018

Sinapov, J., Schenck, C., Staley, K., Sukhoy, V., and Stoytchev, A. (2014). Grounding semantic categories in behavioral interactions: experiments with 100 objects. *Robot. Auton. Sys.* 62, 632–645. doi: 10.1016/j.robot.2012.10.007

Sinapov, J., and Stoytchev, A. (2011). "Object category recognition by a humanoid robot using behavior-Grounded Relational Learning," in *IEEE International Conference on Robotics and Automation (ICRA)* (Shanghai), 184–190.

Stachniss, C., Grisetti, G., and Burgard, W. (2005). Information gain-based exploration using rao-blackwellized particle filters. in *Robotics Science and Systems (RSS)* (Cambridge, MA).

Streeter, M., and Golovin, D. (2009). "An online algorithm for maximizing submodular functions," in *Advances in Neural Information Processing Systems* (Vancouver, BC), 1577–1584.

Sushkov, O. O., and Sammut, C. (2012). "Active robot learning of object properties," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Algarve: IEEE), 2621–2628.

Taniguchi, T., Nagai, T., Nakamura, T., Iwahashi, N., Ogata, T., and Asoh, H. (2016). Symbol emergence in robotics: a survey. *Adv. Robot.* 30, 706–728. doi: 10.1080/01691864.2016.1164622

Teh, Y., Jordan, M., Beal, M., and Blei, D. (2006). Hierarchical Dirichlet processes. *J. Am. Stat. Assoc.* 101, 1566–1581. doi: 10.1198/016214506000000302

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2005). "Sharing clusters among related groups: Hierarchical dirichlet processes," in *Advances in Neural Information Processing Systems* (Vancouver, BC), 1385–1392.

Tuci, E., Massera, G., and Nolfi, S. (2010). Active categorical perception of object shapes in a simulated anthropomorphic robotic arm. *IEEE Trans. Evol. Comput.* 14, 885–899. doi: 10.1109/TEVC.2010.2046174

van Hoof, H., Kroemer, O., Ben Amor, H., and Peters, J. (2012). "Maximally informative interaction learning for scene exploration," in *IEEE/RSJ International Conference on Intelligent Robots and Systems* (Algarve), 5152–5158.

Velez, J., Hemann, G., Huang, A. S., Posner, I., and Roy, N. (2012). Modelling observation correlations for active exploration and robust object detection. *J. Artif. Intell. Res.* 44, 423–453. doi: 10.1613/jair.3516

Zhang, S., Sinapov, J., Wei, S., and Stone, P. (2017). "Robot behavioral exploration and multimodal perception using pomdps," in *AAAI 2017 Spring Symposium on Interactive Multisensory Object Perception for Embodied Agents* (Palo Alto, CA).

Zhou, J., Ross, S., Yue, Y., Dey, D., and Bagnell, J. A. (2013). "Knapsack constrained contextual submodular list prediction with application to multi-document summarization," *ICML 2013 Workshop on Inferning: Interactions between Inference and Learning* (Atlanta).

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX A: PROOF OF THE OPTIMALITY OF THE PROPOSED ACTIVE PERCEPTION STRATEGY

In this appendix, we show that the proposed active perception strategy, which maximizes the expected KL divergence between the current state and the posterior distribution of $\mathbf{z}_j$ after a selected set of actions, minimizes the expected KL divergence between the next and final states.

$$\begin{aligned}
\mathbf{A}_j^* &= \operatorname*{argmin}_{\mathbf{A}\in\mathbf{F}_L^{\mathbf{m}_{\mathbf{o}j}}} \mathbb{E}_{X_j^{\mathbf{M}\backslash\mathbf{m}_{\mathbf{o}j}}|X_j^{\mathbf{m}_{\mathbf{o}j}}} \left[\mathrm{KL}\left(P(\mathbf{z}_j|X_j^{\mathbf{M}}), P(\mathbf{z}_j|X_j^{\mathbf{A}\cup\mathbf{m}_{\mathbf{o}j}})\right)\right] \\
&= \operatorname*{argmin}_{\mathbf{A}\in\mathbf{F}_L^{\mathbf{m}_{\mathbf{o}j}}} \sum_{X_j^{\mathbf{M}\backslash\mathbf{m}_{\mathbf{o}j}}} \sum_{\mathbf{z}_j} \big[P(X_j^{\mathbf{M}\backslash\mathbf{m}_{\mathbf{o}j}}|X_j^{\mathbf{m}_{\mathbf{o}j}})P(\mathbf{z}_j|X_j^{\mathbf{M}}) \\
&\quad \log \frac{P(\mathbf{z}_j|X_j^{\mathbf{M}})}{P(\mathbf{z}_j|X_j^{\mathbf{m}_{\mathbf{o}j}}, X_j^{\mathbf{A}})}\big]
\end{aligned} \tag{A1}$$

The numerator inside of the log function does not depend on $\mathbf{A}$. Therefore, the term related to the numerator can be deleted. In addition, by negating the remaining term, we obtain

$$\begin{aligned}
(\mathrm{A1}) &= \operatorname*{argmax}_{\mathbf{A}\in\mathbf{F}_L^{\mathbf{m}_{\mathbf{o}j}}} \sum_{X_j^{\mathbf{M}\backslash\mathbf{m}_{\mathbf{o}j}}} \sum_{\mathbf{z}_j} [P(X_j^{\mathbf{M}\backslash\mathbf{m}_{\mathbf{o}j}}|X_j^{\mathbf{m}_{\mathbf{o}j}})P(\mathbf{z}_j|X_j^{\mathbf{M}}) \\
&\quad \log P(\mathbf{z}_j|X_j^{\mathbf{m}_{\mathbf{o}j}}, X_j^{\mathbf{A}})] \\
&= \operatorname*{argmax}_{\mathbf{A}\in\mathbf{F}_L^{\mathbf{m}_{\mathbf{o}j}}} \sum_{X_j^{\mathbf{M}\backslash\mathbf{m}_{\mathbf{o}j}}} \sum_{\mathbf{z}_j} [P(\mathbf{z}_j, X_j^{\mathbf{M}\backslash\mathbf{m}_{\mathbf{o}j}}|X_j^{\mathbf{m}_{\mathbf{o}j}}) \\
&\quad \log P(\mathbf{z}_j|X_j^{\mathbf{m}_{\mathbf{o}j}}, X_j^{\mathbf{A}})].
\end{aligned} \tag{A2}$$

By marginalizing $X_j^{\mathbf{M}\backslash(\mathbf{m}_{\mathbf{o}j}\cup\mathbf{A})}$ from (A2), we obtain

$$\begin{aligned}
(\mathrm{A2}) &= \operatorname*{argmax}_{\mathbf{A}\in\mathbf{F}_L^{\mathbf{m}_{\mathbf{o}j}}} \sum_{X_j^{\mathbf{A}}} \sum_{\mathbf{z}_j} P(\mathbf{z}_j, X_j^{\mathbf{A}}|X_j^{\mathbf{m}_{\mathbf{o}j}}) \log P(\mathbf{z}_j|X_j^{\mathbf{m}_{\mathbf{o}j}}, X_j^{\mathbf{A}}) \\
&= \operatorname*{argmax}_{\mathbf{A}\in\mathbf{F}_L^{\mathbf{m}_{\mathbf{o}j}}} [\big(\sum_{X_j^{\mathbf{A}}} \sum_{\mathbf{z}_j} P(\mathbf{z}_j, X_j^{\mathbf{A}}|X_j^{\mathbf{m}_{\mathbf{o}j}}) \log P(\mathbf{z}_j|X_j^{\mathbf{m}_{\mathbf{o}}}, X_j^{\mathbf{A}})\big) \\
&\quad \times \underbrace{\big(- \sum_{\mathbf{z}_j} P(\mathbf{z}_j|X_j^{\mathbf{m}_{\mathbf{o}j}}) \log P(\mathbf{z}_j|X_j^{\mathbf{m}_{\mathbf{o}j}})\big)}_{\textit{constant w.r.t. } \mathbf{A}}] \\
&= \operatorname*{argmax}_{\mathbf{A}\in\mathbf{F}_L^{\mathbf{m}_{\mathbf{o}j}}} \big[\sum_{X_j^{\mathbf{A}}} \sum_{\mathbf{z}_j} [P(X_j^{\mathbf{A}}|X_j^{\mathbf{m}_{\mathbf{o}j}})P(\mathbf{z}_j|X_j^{\mathbf{m}_{\mathbf{o}j}}, X_j^{\mathbf{A}}) \\
&\quad \log P(\mathbf{z}_j|X_j^{\mathbf{m}_{\mathbf{o}j}}, X_j^{\mathbf{A}})] - \sum_{X_j^{\mathbf{A}}} \sum_{\mathbf{z}_j} [P(X_j^{\mathbf{A}}|X_j^{\mathbf{m}_{\mathbf{o}j}})P(\mathbf{z}_j|X_j^{\mathbf{m}_{\mathbf{o}j}}, X_j^{\mathbf{A}}) \\
&\quad \log P(\mathbf{z}_j|X_j^{\mathbf{m}_{\mathbf{o}j}})]\big] \\
&= \operatorname*{argmax}_{\mathbf{A}\in\mathbf{F}_L^{\mathbf{m}_{\mathbf{o}j}}} \sum_{X_j^{\mathbf{A}}} \sum_{\mathbf{z}_j} [P(X_j^{\mathbf{A}}|X_j^{\mathbf{m}_{\mathbf{o}j}})\, \mathrm{KL}\big(P(\mathbf{z}_j|X_j^{\mathbf{m}_{\mathbf{o}j}}, X_j^{\mathbf{A}}), \\
&\quad P(\mathbf{z}_j|X_j^{\mathbf{m}_{\mathbf{o}j}})\big)] \\
&= \operatorname*{argmax}_{\mathbf{A}\in\mathbf{F}_L^{\mathbf{m}_{\mathbf{o}j}}} \mathbb{E}_{X_j^{\mathbf{A}}|X_j^{\mathbf{m}_{\mathbf{o}j}}} [\mathrm{KL}\big(P(\mathbf{z}_j|X_j^{\mathbf{A}\cup\mathbf{m}_{\mathbf{o}j}}), P(\mathbf{z}_j|X_j^{\mathbf{m}_{\mathbf{o}j}})\big)].
\end{aligned}$$