Check for updates

Hierarchical Spatial Concept Formation Based on Multimodal Information for Human Support Robots

Yoshinobu Hagiwara*, Masakazu Inoue, Hiroyoshi Kobayashi and Tadahiro Taniguchi

Emergent Systems Laboratory, College of Information Science and Engineering, Ritsumeikan University, Shiga, Japan

In this paper, we propose a hierarchical spatial concept formation method based on the Bayesian generative model with multimodal information e.g., vision, position and word information. Since humans have the ability to select an appropriate level of abstraction according to the situation and describe their position linguistically, e.g., "I am in my home" and "I am in front of the table," a hierarchical structure of spatial concepts is necessary in order for human support robots to communicate smoothly with users. The proposed method enables a robot to form hierarchical spatial concepts by categorizing multimodal information using hierarchical multimodal latent Dirichlet allocation (hMLDA). Object recognition results using convolutional neural network (CNN), hierarchical k-means clustering result of self-position estimated by Monte Carlo localization (MCL), and a set of location names are used, respectively, as features in vision, position, and word information. Experiments in forming hierarchical spatial concepts and evaluating how the proposed method can predict unobserved location names and position categories are performed using a robot in the real world. Results verify that, relative to comparable baseline methods, the proposed method enables a robot to predict location names and position categories closer to predictions made by humans. As an application example of the proposed method in a home environment, a demonstration in which a human support robot moves to an instructed place based on human speech instructions is achieved based on the formed hierarchical spatial concept.

Keywords: spatial concept, hierarchy, human-robot interaction, multimodal categorization, human support robot, unsupervised learning

1. INTRODUCTION

Space categorization is an important function for human support robots. It is believed that humans predict unknown information flexibly by forming categories of space through their multimodal experiences. We define categories of spaces formed by self-organization from experience as spatial concepts. Furthermore, prediction based on the connection between concepts and words is thought to lead to a semantic understanding of words. It means that spatial concept formation is an important function of human intelligence, and having this ability is important for human support robots.

Spatial concepts form a hierarchical structure. The use of this hierarchical structure enables humans to predict unknown information using concepts in an appropriate layer. For example,

humans can linguistically represent their own positions at an appropriate level of abstraction according to the context of communication, such as "I'm in my home" at the global level, "I'm in the living room" at the intermediate level, and "I'm in front of the TV" at the local level. In this case, the living room has the home in the higher layer and front of the TV in the lower layer. By learning such a hierarchical structure, even if the unknown place does not have features such as front of the TV, its characteristics can be predicted if it has features of the living room. It is expected that the robot acquires spatial concepts in a higher layer by learning the commonality of features in spatial concepts at the lower layer.

Furthermore, the hierarchical structure of spatial concepts plays an important role when a robot moves based on linguistic instructions from a user. As shown in **Figure 1**, even if multiple tables are present in a room, robots can recognize them individually by using a spatial concept at a higher layer, such as "the front of the table in the living space." Indeed, in RoboCup@Home, an international competition in which intelligent robots coexist with humans in home environments, location names are defined as two layers in the tasks of a General Purpose Service Robot[1] as shown in **Table 1**. This table indicates that having sense of space relations is important for a robot coexisting with humans, e.g., that the living space has a center table. By having such hierarchical spatial concepts, it becomes possible to describe and move within a space based on linguistic communication with a user.

We assume that a computational model, which considers the hierarchical structure of spatial concepts, enables robots to acquire not only the spatial concepts, but also the hierarchical structure hiding among the spatial concepts through a bottom-up approach and form spatial concepts similar to those perceived by humans. The goal of this study was to develop a robot that can predict unobserved location names and positions from observed information using formed hierarchical spatial concepts. The main contributions of this paper are as follows.

- We propose a method of forming hierarchical spatial concepts using a Bayesian generative model based on multimodal information, e.g., vision, position, and word information.
- We show that spatial concepts formed by the proposed method enable a robot to predict location names and positions similar to prediction made by humans.
- We demonstrate application examples in which a robot moves within and describes a space based on linguistic communication with a user through the hierarchical spatial concept formed by the proposed method.

The rest of this paper is structured as follows. Section 2 describes related works. Section 3 presents an overview and the computational model of hierarchical spatial concept formation. Section 4 presents experimental results evaluating the effectiveness of the proposed method in space categorization. Section 5 describes application examples of using hierarchical spatial concepts in a home environment. Finally, section 6 presents conclusions.

---

[1]GPSR Command Generator: https://github.com/kyordhel/GPSRCmdGen

2. RELATED WORKS

In order for a robot to move within a space, a metric map consisting of occupancy grids that encode whether or not an area is navigable is generally used. The simultaneous localization and mapping (SLAM) (Durrant-Whyte and Bailey, 2006) is a famous localization method for mobile robots. However, the tasks that are coordinated with a user cannot be performed using only a metric map, since semantic information is required for interaction with a user. Nielsen et al. (2004) proposed a method of expanding a metric map into a semantic map by attaching a single-frame snapshot in order to share spatial information between a user and a robot. As a bridge between a metric map and human-robot interaction, research on semantic maps that provide semantic attributes (such as object recognition results) to metric maps has been performed (Pronobis et al., 2006; Ranganathan and Dellaert, 2007). Studies have also been reported on giving semantic object annotations to 3D point cloud data (Rusu et al., 2008, 2009). Moreover, in terms of studies based on multiple cues, Espinace et al. (2013) proposed a method of characterizing places according to low-level visual features associated to objects. Although these approaches could categorize spaces based on semantic information, they did not deal with linguistic information about the names that represent spaces.

In the field of navigation tasks with human-robot interaction, methods of classifying corridors and rooms using a predefined ontology based on shape and image features have been proposed (Zender et al., 2008; Pronobis and Jensfelt, 2012). In studies on semantic space categorization, Kostavelis and Gasteratos (2013) proposed a method of generating a 3D metric map that is semantically categorized by recognizing a place using bag of features and support vector machines. Granda et al. (2010) performed spatial labeling and region segmentation by applying a Gaussian model to the SLAM module of a robot operating system (ROS). Mozos and Burgard (2006) proposed a method of classifying metric maps into semantic classes by using adaboost as a supervised learning method. Galindo et al. (2008) utilized semantic maps and predefined hierarchical spatial information for robot task planning. Although these approaches were able to ground several predefined names to spaces, the learning of location names through human-robot communication in a bottom-up manner has not been achieved.

Many studies have been conducted on spatial concept formation based on multimodal information observed in individual environments (Hagiwara et al., 2016; Heath et al., 2016; Rangel et al., 2017). Spatial concepts are formed in a bottom-up manner based on multimodal observed information, and allow predictions of different modalities. This makes it possible to estimate the linguistic information representing a space from position and image information in a probabilistic way. Gu et al. (2016) proposed a method of learning relative space categories from ambiguous instructions. Taniguchi et al. (2014, 2016) proposed computational models for a mobile robot to acquire spatial concepts based on information from recognized speech and estimated self-location. Here, the spatial concept was defined as the distributions of names and positions at each place.
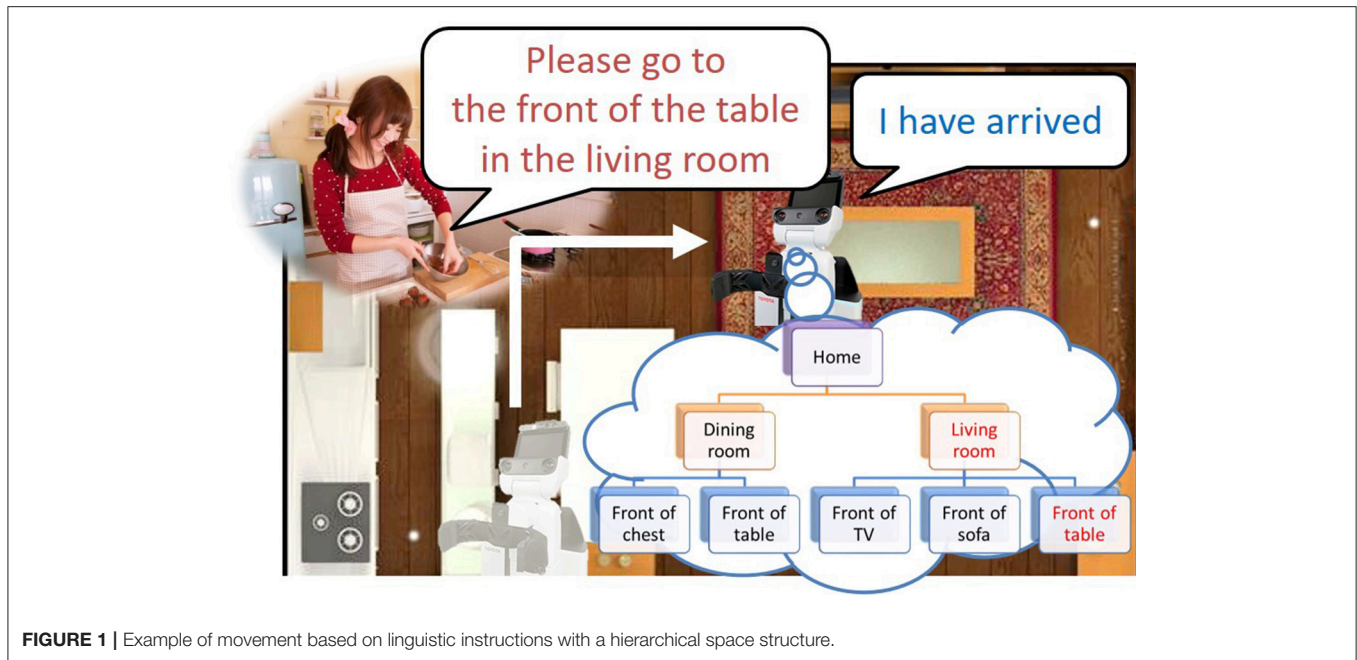
FIGURE 1 | Example of movement based on linguistic instructions with a hierarchical space structure.

The method enables a robot to predict a positional distribution from recognized human speech through formed spatial concepts. Ishibushi et al. (2015) proposed a method of learning the spatial regions at each place by stochastically integrating image recognition results and estimated self-positions. In these studies, it was possible to form a spatial concept conforming to human perception such as an entrance and a corridor by inferring the parameters of the model.

However, these studies did not focus on the hierarchical structure of spatial concepts. In particular, the features of the higher layer, such as the living space, are included in the features of the lower layer, such as the front of the television, and it was difficult to form the spatial concept in the abstract layer. Furthermore, the ability to understand and describe a place linguistically in different layers is an important function in robots that provide services through linguistic communication with humans. Despite the importance of the hierarchical structure of spatial concepts, a method that enables such concept formation has not been proposed in previous studies. We propose a method that forms a hierarchical spatial concept in a bottom-up manner from multimodal information and demonstrate the effectiveness of the formed spatial concepts in predicting location names and positions.

## 3. HIERARCHICAL SPACE CONCEPT FORMATION METHOD

3.1. Overview

An overview of the proposed method of forming hierarchical spatial concepts is shown in **Figure 2**. First, a robot was controlled manually in an environment based on a map generated by simultaneous localization and mapping

TABLE 1 | Definition of location names with two layers in RoboCup@Home.

Name (1st layer)	Name (2nd layer)
Living room	Bar
Living room	TV stand
Living room	Center table
Office	Drawer
Office	Desk
Kitchen	Bar
Kitchen	Cupboard
Bathroom	Cupboard

(SLAM) (Durrant-Whyte and Bailey, 2006) and acquires multimodal information, i.e., vision, position, and word information from attached sensors. Vision information is acquired as a feature vector generated by a convolutional neural network (CNN) (Krizhevsky et al., 2012). Position information is acquired as coordinate values in the map estimated by Monte Carlo localization (MCL) (Dellaert et al., 1999). Word information is acquired as set of words by word recognition. Text input is used for word recognition in this study. Second, acquired vision, position, and word information is represented as histograms. The histograms are utilized as observations in each modality. Third, the formation of hierarchical spatial concepts is performed by using hierarchical multimodal latent Dirichlet allocation (hMLDA) (Ando et al., 2013) on the observations. The proposed method enables a robot to form hierarchical spatial concepts in a bottom-up manner based on observed multimodal information. Therefore, it is possible to adaptively learn location names and the hierarchical structure of a space, which depend on the environment.

FIGURE 2 | Overview of the proposed method for hierarchical spatial concept formation.
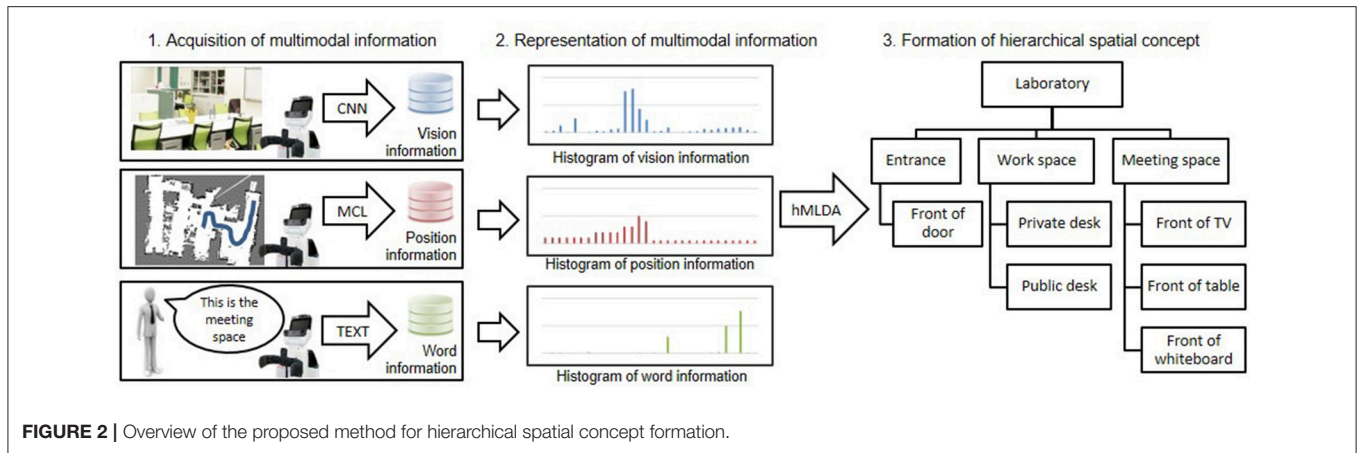
3.2. Acquisition and Feature Extraction of Multimodal Information

3.2.1. Vision Information
Vision information was acquired as the object recognition results of a captured image by Caffe (Jia et al., 2014), which is a framework of CNN (Krizhevsky et al., 2012) provided by Berkeley Vision and Learning Center. The parameters of CNN were trained by using the dataset from the ImageNet Large Scale Visual Recognition Challenge 2013[2], which comprises 1,000 object classes, e.g., television, cup, and desk. The output of Caffe is given as a probability $p(a_i)$ at an object class $a_i \in \{a_1, a_2, ..., a_I\}$ where I is the number of object classes and was set to 1,000. The probability $p(a_i)$ was represented as a 1,000-dimensional histogram of vision information $\boldsymbol{w}^{(v)} = (w_1^{(v)}, w_2^{(v)}, \cdots, w_{1,000}^{(v)})^T$ by the following equation:

$$w_i^{(v)} = p(a_i) * 10^2. \tag{1}$$

3.2.2. Position Information
The position information (x, y) in the map generated by SLAM was estimated by MCL (Dellaert et al., 1999). It is assumed that the observed information is generated from a multinomial distribution in hMLDA. For this reason, the observed information with a continuous value is generally converted into a finite dimensional histogram by vector quantization. Ando et al. (2013) replaced the observed information with typical patterns by k-means clustering to form a finite dimensional histogram. The proposed method converts a position information (x, y) into a finite dimensional histogram of position information w^p by hierarchical k-means clustering. The positional information (x, y) was classified hierarchically into 2, 4, 8, 16, 32, and 64 clusters with six layers by applying k-means clustering with $k = 2$ six times. If a position (x, y) was classified into a cluster $c_i \in \{0, 1\}$ at the ith layer, a path for the position information was described as $C = \{c_1, c_2, c_3, c_4, c_5, c_6\}$. The path C has the structure of a binary tree with six layers. The number of nodes at the 6th layer is $2^6 = 64$. The position information (x, y)

is represented as a 64-dimensional histogram of the position information $\boldsymbol{w}^{(p)} = (w_1^{(p)}, w_2^{(p)}, \cdots, w_{64}^{(p)})^T$ by incrementing $w_i^{(p)}$ based on the path C. For example, in a path C of position information (x, y), when $c_1 = 0$, $w_1^{(p)}$ to $w_{32}^{(p)}$ corresponding to nodes at the 6th layer are incremented, and when $c_1 = 1$, $w_{33}^{(p)}$ to $w_{64}^{(p)}$ are incremented. Similarly, $w^{(p)}$ corresponding to nodes at the 6th layer below it are incremented in each layer.

3.2.3. Word Information
The voice information uttered by a user is converted manually into text data, which is then used as word information. In section 5, rospeex (Sugiura and Zettsu, 2015) is used to convert human speech into text data. The location names are manually extracted from the text data. The word information is described as a set of location names, which is a Bag of Words (Harris, 1954) with a location name as a word. The user could give not only one name but also several names to a robot at a given position. The given word information was represented as a histogram of word information $\boldsymbol{w}^{(w)} = (w_1^{(w)}, w_2^{(w)}, \cdots, w_J^{(w)})^T$. J is the dimension of $\boldsymbol{w}^{(w)}$, and depends on the number of location names in a dictionary $S = \{s_1, s_2, \cdots, s_J\}$, which was obtained through the training phase. $w_j^{(w)}$ was incremented when a location name s_j was taught from user. J is the number of location names. Histograms of vision, position, and word information were used as observations in hMLDA.

3.3. Hierarchical Categorization by hMLDA
The hierarchical structure of spatial concepts is supported by nested Chinese restaurant process (nCRP) (Blei et al., 2010) in hMLDA (Ando et al., 2013). nCRP is an extended model of the Chinese restaurant process (CRP) (Aldous, 1985), which is a Dirichlet process used to generate multinomial distribution with infinite dimensions. nCRP stochastically calculates the hierarchical structure based on the idea that there are infinite Chinese restaurants with infinite number of tables. **Figure 3** shows the overview of nCRP. A box and a circle represent a restaurant and a customer, respectively. The customer
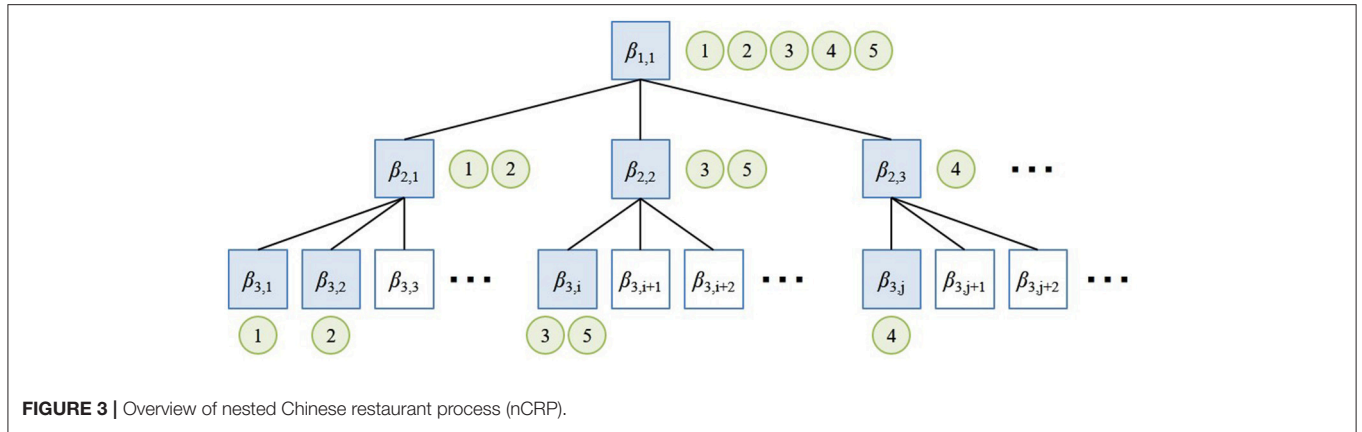
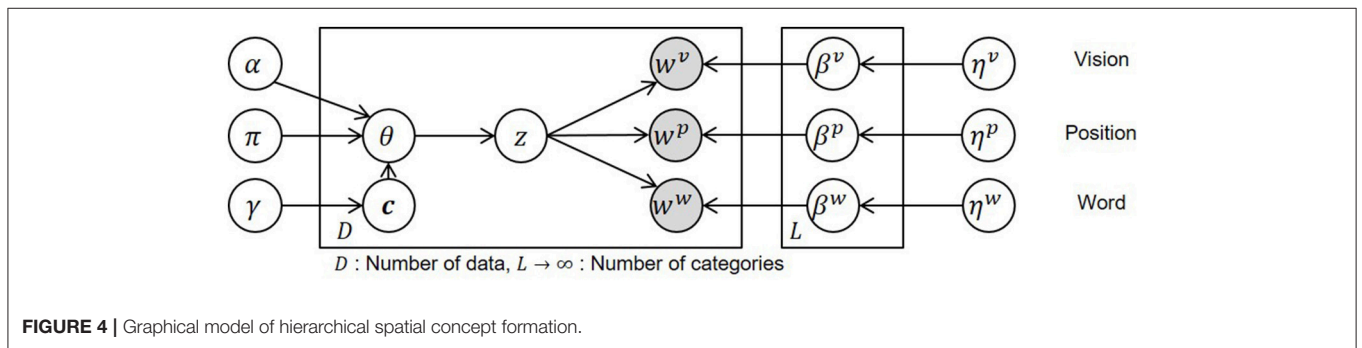**FIGURE 3 |** Overview of nested Chinese restaurant process (nCRP).

FIGURE 4 | Graphical model of hierarchical spatial concept formation.

stochastically decides the restaurant to visit. In the proposed method, a box and a circle mean a spatial concept and data, respectively. Data is stochastically allocated to a spatial concept in each layer by the nCRP. In hMLDA, each spatial concept has a probability distribution with parameter $\beta_{l,i}$ to generate data. The proposed method forms a hierarchical spatial concept by hierarchical probabilistic categorization using nCRP. In the non-hierarchical approach, a place called "meeting space" and its partial places called "front of the table" and "front of the TV" are formed in the same layer. Therefore, the meeting space is learned as a place different from places called "front of the table" and "front of the TV." The proposed method enables the robot to learn the meeting space as a upper concept encompassing places called "front of the table" and "front of the TV" as lower concepts.

The graphical model of hMLDA in the proposed method and the definition of the variables are shown in **Figure 4** and **Table 2**, respectively. In **Figure 4**, c is a tree-structured path generated by nCRP with a parameter γ and z is a category index for a spatial concept that is generated by a stick-breaking process (Pitman, 2002) with parameters α and π. w^υ, w^p, w^w are acquired vision, position, and word information generated by multinomial distributions with a parameter β^m at a modality m ($m \in \upsilon, p, w$). β^m was determined according to a Dirichlet prior distribution with a parameter η^m. D and L written on plates are the number of acquired data and the number of categories, respectively.

The generation process of the model is described as follows:

$$\beta_k^m \sim \text{Dirichlet}(\eta^m) \tag{2}$$
$$c_d \sim \text{nCRP}(\gamma) \tag{3}$$
$$\theta_d \sim \text{GEM}(\alpha, \pi) \tag{4}$$
$$z_{d,n}^m \sim \text{Multi}(\theta_d) \tag{5}$$
$$w_d^m \sim \text{Multi}(\beta_{c_d}[z_{d,n}^m]), \tag{6}$$

where:

- The parameter β_k^m of a multinomial distribution is generated by a Dirichlet prior distribution with a parameter η^m in a table $k(k \in T)$, e.g., $\beta_{1,1}$ and $\beta_{2,1}$ in **Figure 3**.
- The path c_d in a tree structure for the data d ($d \in 1, 2, ..., D$) is decided by nCRP with a parameter γ. c_d is represented by the sequence of numbers assigned to each node in the path, e.g., $\{(1, 1), (2, 1), (3, 2)\}$ at data 2 in **Figure 3**.
- The parameter θ_d of a multinomial distribution is generated by the stick-breaking process based on a GEM distribution which forms θ_d from a Beta$(\alpha\pi, (1 - \alpha)\pi)$ distribution with the parameters $\alpha(0 \leq \alpha \leq 1)$ and $\pi(\pi > 0)$ (Pitman, 2006). θ_d represents the selection probability of a layer in a path c_d and corresponds to the generation probability of a category index z in a path c_d.
- $z_{d,n}^m$, which is a category index at the nth feature of the observed information w_d^m, is generated by a multinomial distribution with a parameter θ_d.

TABLE 2 | Definition of variables in the graphical model.

Variable	Definition
w^v, w^p, w^n	Observation of vision, position and word information
z	Index of category
$\beta^v, \beta^p, \beta^n$	Parameter of multinomial distribution in vision, position and word information
θ	Parameter of multinomial distribution in category
c	Path of tree structure
η^v, η^v, η^w	Parameter of Dirichlet prior distribution
γ	Hyper parameter of c
α, π	Hyper parameter of θ

- w_d^m is the observed information generated by a multinomial distribution with a parameter β from a category $z_{d,n}^m$ at a path c_d.

In this study, z is equivalent to a spatial concept expressed by the location name such as "the living room" or "front of the table."

Model parameter learning was performed by a Gibbs sampler. Parameters were calculated by alternately sampling a path c_d for each datum and a category $z_{d,n}^m$ assigned to the nth feature value of a modality m of the data d in the path. Category $z_{d,n}^m$ was sampled according to the following formula.

$$z_{d,n}^m \sim p(z_{d,n}^m | z_{-(d,n)}^m, c, w^m, \alpha, \pi, \eta^m)$$
$$\propto p(z_{d,n}^m, z_{-(d,n)}^m, c, w^m | \alpha, \pi, \eta) \tag{7}$$
$$\propto p(z_{d,n}^m | z_{d,-n}^m, \alpha, \pi) p(w_{d,n}^m | z, c, w_{-(d,n)}^m, \eta^m),$$

where $-(d,n)$ means excluding the nth feature value of the data d. $p(z_{d,n}^m | z_{d,-n}^m, \alpha, \pi)$ is a multinomial distribution generated by the stick-breaking process. The probability, that k is assigned to a category of the n-th feature of modality m of the d-th data, was calculated by the following formula.

$$p(z_{d,n}^m = k | z_{d,-n}^m, \alpha, \pi) = E\left[V_k \prod_{j=1}^{k-1}(1 - V_j) | z_{d,-n}^m, \alpha, \pi\right]$$
$$= E\left[V_k | z_{d,-n}^m, \alpha, \pi\right] \prod_{j=1}^{k-1} E\left[1 - V_j | z_{d,-n}^m, \alpha, \pi\right] \tag{8}$$
$$= \frac{(1-\alpha)\pi + \#[z_{d,-n}^m = k]}{\pi + \#[z_{d,-n}^m \geq k]} \prod_{j=1}^{k-1} \frac{\alpha\pi + \#[z_{d,-n}^m > j]}{\pi + \#[z_{d,-n}^m \geq j]},$$

where $\#[\cdot]$ is a number that satisfies a given condition and V_k and V_j are values that determine the rate of folding a branch in categories k and j by the stick-breaking process, respectively.

In Formula (7), $p(w_{d,n}^m | z, c, w_{-(d,n)}^m, \eta^m)$ is the probability that a feature value is generated from a path c_d and a category $z_{d,n}^m$. Since it is assumed that the parameters of the multinomial distribution that generates a feature value are generated from a Dirichlet prior distribution, the following formula is obtained.

$$p(w_{d,n}^m | z, c, w_{d,n}^m, \eta^m) \propto \#[z_{-(d,n)}^m = z_{d,n}^m, c_{z_{d,n}^m} = c_{d,z_{d,n}^m}, w_{-(d,n)}^m$$
$$= w_{d,n}^m] + \eta^m \tag{9}$$

This gives the number of times that a category $z_{d,n}^m$ is assigned to a feature value $w_{d,n}^m$ in a path c_d. A path c_d was sampled by the following formula.

$$c_d \sim p(c_d | w^v, w^p, w^w, c_{-d}, z, \eta^v, \eta^p, \eta^w, \gamma)$$
$$\propto p(c_d | c_{-d}, \gamma) p(w_d^v | c, w_{-d}^v, z^v, \eta^v) p(w_d^p | c, w_{-d}^p, z^p, \eta^p)$$
$$p(w_d^w | c, w_{-d}^w, z^w, \eta^w), \tag{10}$$

where c_{-d} is a set of paths excluding c from c_d. Sampling based on Formulas (9) and (10) was repeated for each training datum $d \in \{d_1, d_2, \cdots, d_D\}$. In this process, paths and categories for all observed data converge to $\hat{c}$ and $\hat{z}$.

3.4. Name Prediction and Position Category Prediction

If vision information w_t^v and position information w_t^p are observed at a time t, then the posterior probability of word information w_t^w can be calculated with estimated parameters $\hat{c}$ and $\hat{z}$ by the following formula.

$$p(w_t^w | \hat{z}, \hat{c}, w^w, w^v, w^p, c_t, w_t^v, w_t^p, \alpha, \pi, \eta^n, \eta^v, \eta^p) =$$
$$\sum_{z_t} p(w_t^w | z_t, \hat{z}^w, \hat{c}, w^w, \eta^w) \tag{11}$$
$$p(z_t | \hat{z}^v, \hat{z}^p, \hat{c}, w^v, w^p, c_t, w_t^v, w_t^p, \alpha, \pi, \eta^v, \eta^p)$$

The location name $\hat{n}$ can be predicted by the maximum value of the calculated posterior probability.
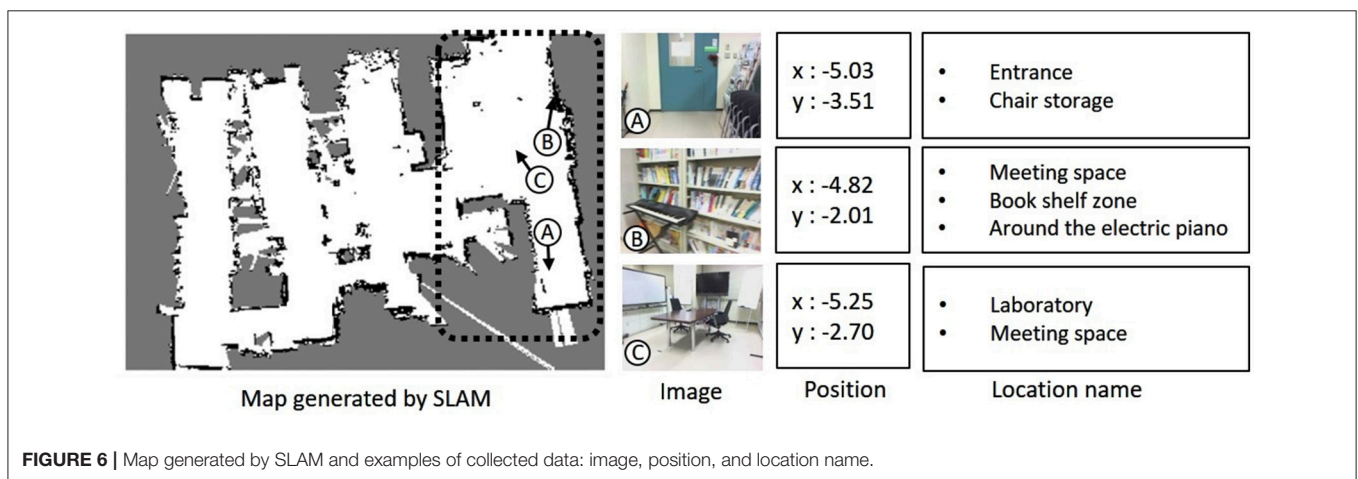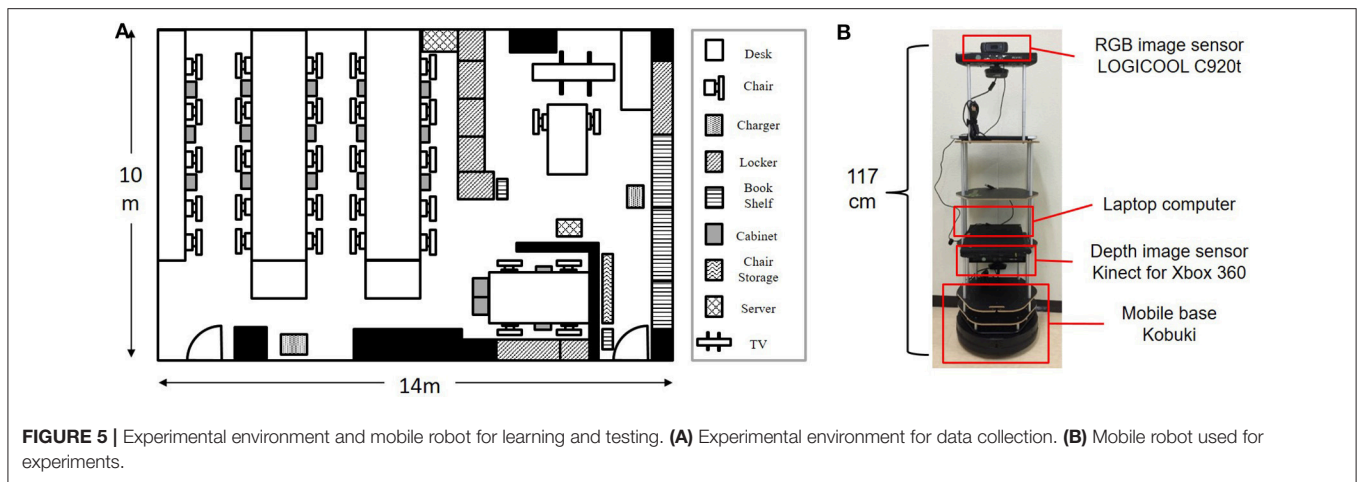
If word information w_t^w is obtained at a time t, then a category z_t^w can be predicted by Formula (12) and selecting position $\hat{p}$ randomly from dataset $D_{z_t^w}$, which is a set of position data categorized into z_t^w. $D_{z_t^w}$ was automatically generated by the robot itself as a part of the categorization process.

$$z_t^w \sim p(z_t^w | z_{-t}^w, w_t^w, \hat{c}, w^w, w^v, w^p, \eta^w, \eta^v, \eta^p, \alpha, \pi) \tag{12}$$

4. EXPERIMENT

4.1. Purpose

We conducted experiments to verify whether the proposed method can form hierarchical spatial concepts, which enable a robot to predict location names and position categories close to predictions made by humans. In the experiment, (1) the influence of multimodal information, i.e., words, on the formation of a hierarchical spatial concept was evaluated by comparing the space categorization results of using the proposed method and those of hierarchical latent Dirichlet allocation (hLDA) (Blei et al., 2010), which is a hierarchical categorization method with single modality; (2) the similarity between the hierarchical spatial concepts formed by the proposed method and those made by humans was evaluated in terms of predicting location names and position categories.

FIGURE 5 | Experimental environment and mobile robot for learning and testing. (A) Experimental environment for data collection. (B) Mobile robot used for experiments.

FIGURE 6 | Map generated by SLAM and examples of collected data: image, position, and location name.

4.2. Experimental Conditions

Figure 5A shows an experimental environment which includes furniture, e.g., tables, chairs, and a book shelf, in order to collect training and test data. **Figure 5B** shows a mobile robot, which consists of a mobile base, a depth sensor, an image sensor, and a computer, used to generate a map and collect multimodal information in the test environment. The height of the camera attached to the robot was 117 cm in consideration of the typical eye level in the human body. This is equivalent to the average height of a 5-year-old boy in Japan. The Navigation Stack package[3] was used with ROS Hydro[4] for mapping, localization, and moving in the experiment. The robot was manually controlled to collect data from the environment. The orientation of the robot was controlled in as many different orientations as possible.

Figure 6 shows a map generated in the environment by the robot using SLAM and examples of the collected data. Collected data consisted of image, position, and word information as shown in the samples of collected data at A, B, and C. In the experiment,

900 data points were used for training and 100 data points were used for testing from a total of 1,000 data points collected in the area surrounded by a dotted line in the map. The robot simultaneously acquired images and self-position data (x, y) at times of particle re-sampling for MCL. Words were given as location names by a user who was familiar with the experimental environment. The user gave one or more location names suitable for the place at a data point during the training. In example A, not only a name such as "front of the door" but also a name representing a space such as "entrance" and a name meaning a room such as "laboratory" were given as word information. Word information was partially supplied as training data. Five training data sets were prepared to evaluate robustness of the naming rate in training data as 1, 2, 5, 10, and 20%.

The similarity between the spatial concepts formed by the proposed method and those made by humans was evaluated in experiments of location name prediction and position category prediction based on the ground truth. The ground truth information was given for 100 test data points according to the agreement of three experts who were familiar with the environment. The hierarchy of the space in the experimental environment was defined as global, intermediate, and local.

TABLE 3 | List of location names and ground truth in the hierarchy.

Global	Laboratory	
Intermediate	Entrance	Meeting space
Local	Front of the door	Umbrella storage	Magazine rack zone
	Chair storage	Book shelf zone	Around Skype PC
	Around the charger	Around the electric piano	Locker zone
	Front of the white board	Front of the display	Front of the table

Location names assigned to each hierarchy are shown in **Table 3**. As the ground truth for name prediction, three location names were uniformly given to each test datum considering the hierarchy to evaluate the accuracy of name prediction. As the ground truth for the position category prediction, regions corresponding to the 15 location names in **Table 3** were decided on the map. **Figure 7** shows the three regions of the "laboratory," "entrance," and "front of the table." The environment was divided into a grid of 50 units in length and 25 units in width, and the gray grids show the ground truth.

In the name prediction experiment, the accuracy of name prediction compared with the ground truth was calculated as an index of similarity. Formula (11) was used to predict names using the proposed method. The accuracy of name prediction at global, intermediate, and local levels was calculated by the following formula.

$$Accuracy = \frac{M_l}{D}, \tag{13}$$

where M_l is a number matching the predicted names with the ground truth at layer l in the test dataset and D is the number of test data. In the experiment, l was set as ($l \in \{global, intermediate, local\}$) and D was 100.

In the position category prediction experiment, the precision, recall, and F-measure of the predicted position categories compared with the ground truth were calculated as an index of similarity. In the proposed method, a position (x, y) sampled multiple times for each location name by Formula (12). The precision, recall, and F-measure of position category prediction were calculated by the following formulas.

$$Precision = \frac{T_n}{T_n + F_n} \tag{14}$$

$$Recall = \frac{T_n}{G_n} \tag{15}$$

$$F\text{-}measure = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision}, \tag{16}$$

where T_n is a number matching the position with the ground truth for location name n, F_n is a number that does not match the position with the ground truth, and G_n is the number of

grids for the ground truth. In the experiment, n was set as ($n \in \{1, 2, \cdots, 15\}$).

In the proposed method, the hyper-parameters $\alpha, \pi, \gamma, \eta$ were set as $\alpha = 0.5, \pi = 100, \gamma = 1.0, \eta^v = 1.0 \times 10^{-1}, \eta^p = 1.0 \times 10^{-3}, \eta^w = 1.0 \times 10^{-2}$, respectively. The path c and category z of each data were trained with the hyper-parameters. In the experiment, the dimensions of the information vectors w^v, w^p, and w^w were 1,000, 64, and 15, respectively.

4.3. Baseline Methods

The most frequent class, nearest neighbor method, multimodal hierarchical Dirichlet process (HDP), and spatial concept formation model were used as baseline methods for evaluating the performance of the proposed method in the name prediction and position category prediction experiments. In the latter, the sampling of position for each location name was performed 100 times.

4.3.1. Most Frequent Class

The training dataset $D = \{d_1, d_2, \cdots, d_I\}$ is used in this method. The datum d_i consists of the position information $p_i = (x_i, y_i)$ and the word information w_i, which is a set of location names. The frequency cnt_{n_j} of each location name $n_j (j \in \{1, 2, \cdots, 15\})$ is counted in the training dataset D. The location name n_j is classified into three clusters by k-means ($k = 3$) based on cnt_{n_j}. The three clusters of location names are $C_{global}, C_{intermediate}$, and C_{local} in descending order of the frequency of the location name based on the assumption that global location names are more frequent than local location names. In the training dataset D, if a datum d_i includes a location name in $C_{global}, C_{intermediate}$, and C_{local}, the datum d_i is set as a global dataset D_g, an intermediate dataset D_i, and a local dataset D_l. The location names in the global, intermediate, and local levels are predicted as the most frequent location name in each dataset D_g, D_i, and D_l, respectively.

In the position category prediction, the positions are predicted by sampling the position information $\hat{p}$ randomly from the datasets $D_{g,f}, D_{i,f}$, and $D_{l,f}$, which have the most frequent location names in each dataset D_g, D_i, and D_l, respectively. The sampling of position information for each location name was performed 100 times.

4.3.2. Nearest Neighbor (Position and Word)

The nearest neighbor method (Friedman et al., 1977) discriminates data based on Euclidean distance. A datum d_i involves position information $p_i = (x_i, y_i)$ and word information w_i. w_i consists of a set of location names that obtained at a position p_i in the training. For example, w_i at data point B in **Figure 6** contains the following location names: "Meeting space," "Book shelf zone," and "Around the electric piano." If position information p_t is observed, then word information $\hat{w}_t$ is calculated with the training dataset $D = \{(p_1, w_1), (p_2, w_2), \cdots, (p_I, w_I)\}$ by the following formulas.

$$k = \operatorname*{arg\,min}_{1 \leq i \leq I} |p_t - p_i| \tag{17}$$
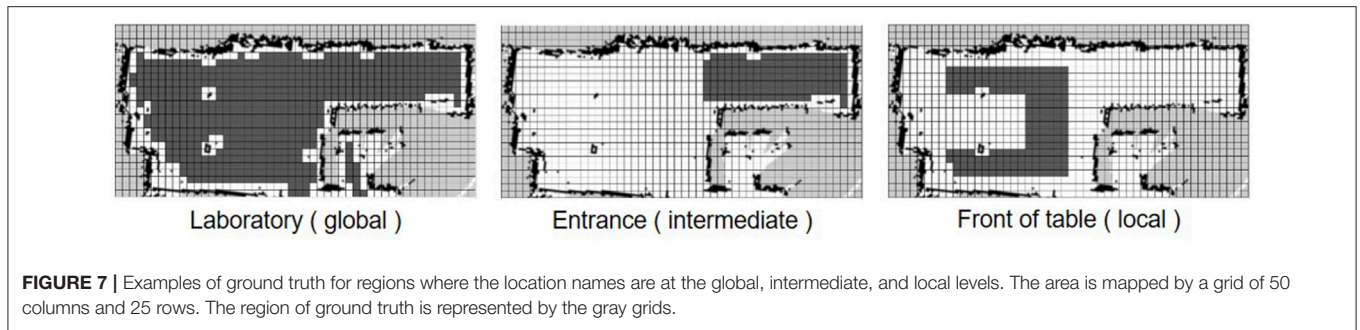
$$\hat{w}_t = w_k \tag{18}$$

FIGURE 7 | Examples of ground truth for regions where the location names are at the global, intermediate, and local levels. The area is mapped by a grid of 50 columns and 25 rows. The region of ground truth is represented by the gray grids.
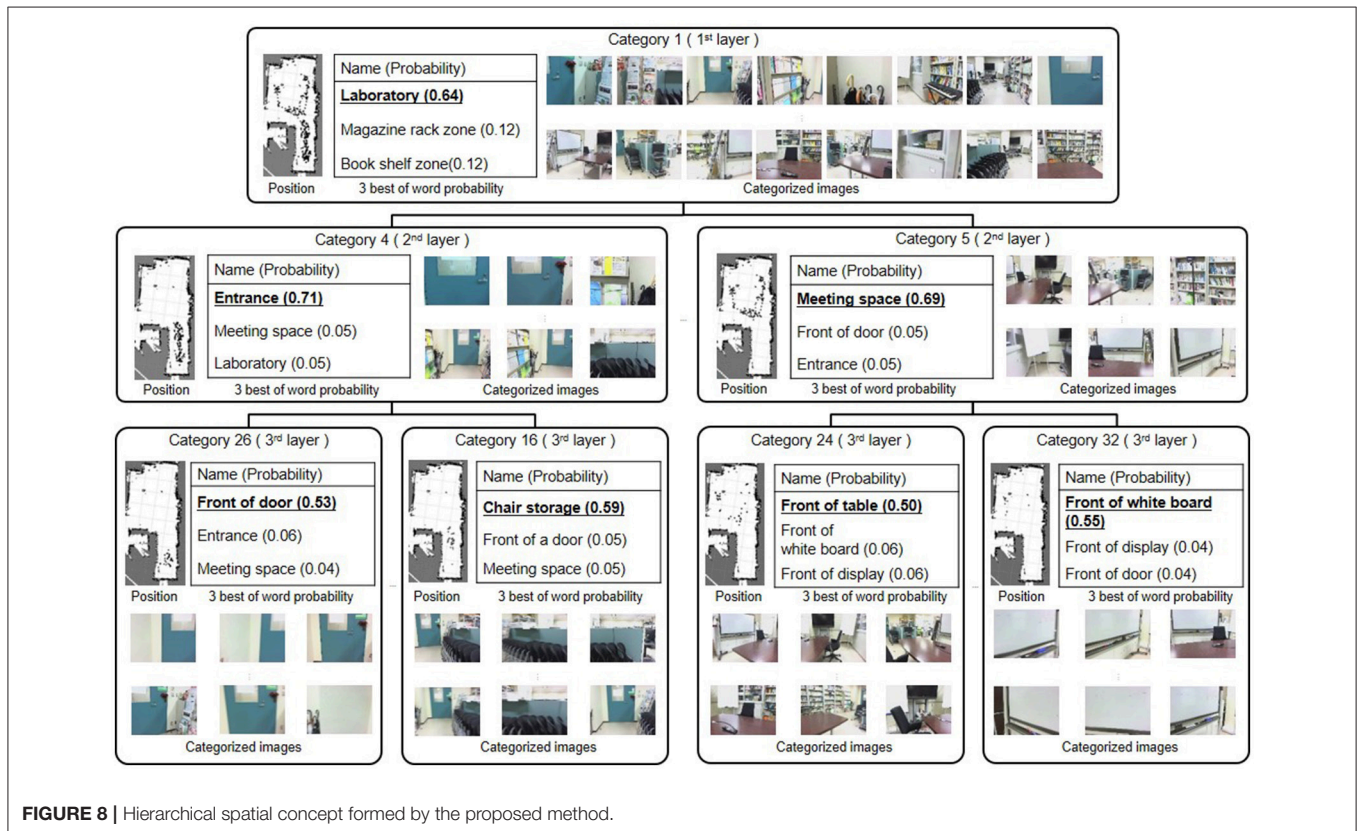

FIGURE 8 | Hierarchical spatial concept formed by the proposed method.

The location name $\hat{n}$ can be predicted by randomly selecting a location name from location names in $\hat{w}_t$ of the nearest data point.

If word information w_t is observed, then position information $\hat{p}_t$ is randomly selected from dataset D_{n_t}, which is a set of data $d_i = (p_i, w_i)$ satisfying the formula $w_i \in w_t$. The sampling of position information for each location name was performed 100 times.

4.3.3. Nearest Neighbor (Vision, Position and Word)

This method is used only in the name prediction experiment. A datum d_i includes vision information v_i, position information $p_i = (x_i, y_i)$ and word information w_i. v_i is a value calculated by Formula (1) at a position p_i during training. w_i consists of a set of location names that are obtained at a

position p_i during the training. If the vision information v_t and the position information p_t are observed, then the word information $\hat{w}_t$ can be calculated with the training dataset $D = \{(v_1, p_1, w_1), (v_2, p_2, w_2), \cdots, (v_I, p_I, w_I)\}$ by using the following formulas.

$$k = \underset{1 \leq i \leq I}{\arg \min}(\alpha|v_t - v_i| + (\alpha - 1)|p_t - p_i|) \qquad (19)$$

$$\hat{w}_t = w_k \qquad (20)$$

where α is the weight coefficient between vision and position information. α was set as 0.3 in the validation dataset empirically. The location name $\hat{n}$ can be predicted by randomly selecting a location name from the location names in $\hat{w}_t$ of the nearest data point.

4.3.4. Multimodal HDP

Multimodal HDP (Nakamura et al., 2011) enables the multimodal handling of HDP (Teh et al., 2005), which is a method of categorizing observed data based on a Bayes generative model, in the topic distribution of latent Dirichlet allocation (LDA) as HDP. The graphical model and definition of variables in the multimodal HDP are shown in the Supplementary Material. Here, multimodal HDP was trained using vision, position, and word information. If vision information w_t^v and position information w_t^p are observed at a time t, then the posterior probability of word information w_t^w can be calculated by the following formula:

$$p(w_t^w|\hat{z}, w^w, w^v, w^p, w_t^v, w_t^p, \pi, \eta^w, \eta^v, \eta^p) =$$
$$\sum_{z_t} p(w_t^w|z_t, \hat{z}^w, \hat{c}, w^w, \eta^w)p(z_t|\hat{z}^v, \hat{z}^p, w^v, w^p, w_t^v, w_t^p, \pi, \eta^v, \eta^p)$$

$$(21)$$

The location name $\hat{n}$ can be predicted by the maximum value of the calculated posterior probability.

If word information w_t^w is obtained at a time t, then a category z_t^w can be predicted by Formula (22) and selecting position information $\hat{p}$ randomly from dataset $D_{z_t^w}$, which is a set of position data categorized into z_t^w.

$$z_t^w \sim p(z_t^w|z_{-t}^w, w_t^w, w^w, w^v, w^p, \eta^w, \eta^v, \eta^p, \pi) \quad (22)$$

The sampling of position information for each location name was performed 100 times. In the multimodal HDP, the hyper-parameters π, η were set as $\pi = 50, \eta^v = 5.0 \times 10^{-1}, \eta^p = 1.0 \times 10^{-1}, \eta^w = 1.0 \times 10^{-1}$ in the validation dataset. The category z of each data is trained with the hyper-parameters.

4.3.5. Spatial Concept Formation

Spatial concept formation (SpCoFo)[5] is a model that integrates name modalities into the spatial region learning model (Ishibushi et al., 2015). The model forms concepts from multimodal information and predicts unobserved information. The graphical model and definition of variables in the spatial concept formation model are shown in the Supplementary Material. The posterior probability of word information w_t^n after obtaining vision information w_t^v and position information p_t was calculated by the following formula:

$$p(w_t^n|p_t, w_t^v) = \sum_{z_t} p(w_t^n|z_t)p(z_t|p_t, w_t^v)$$
$$= \sum_{z_t} p(w_t^n|\beta_{z_t}^n)p(p_t|\mu_{z_t}, \Sigma_{z_t})p(w_t^v|\beta_{z_t}^v) \quad (23)$$

The location name $\hat{n}$ can be predicted by the maximum value of the calculated posterior probability.

[5]Spatial Concept Formation: https://github.com/EmergentSystemLabStudent/Spatial_Concept_Formation

The prediction of position $\hat{p}_t$ after obtaining word information w_t^n was calculated by estimating a category z_t and sampling position information $\hat{p}$ using the following formulas.

$$z_t = \arg\max_{z_t} p(z_t|w_t^n)$$
$$\hat{p}_t \sim p(p_t|\mu_{z_t}, \Sigma_{z_t}) \quad (24)$$

The sampling of position information for each location name was performed 100 times. In the spatial concept formation, the hyper-parameters π, η, μ_0, κ_0, ψ_0, and ν_0 were set as $\pi = 50$, $\eta^v = 5.0 \times 10^{-1}$, $\eta^w = 1.0 \times 10^{-1}$, $\mu_0 = (x_{center}, y_{center})$, $\kappa_0 = 3.0 \times 10^{-2}$, $\psi_0 = diag[0.05, 0.05, 0.05, 0.05]$, and $\nu_0 = 15$ in the validation dataset, respectively. (x_{center}, y_{center}) indicates the center of the map. The category z of each data is trained with the hyper-parameters.

4.4. Experimental Results
4.4.1. Hierarchical Space Categorization

Figure 8 shows some categories formed by the proposed method. Categorized training data at each category are shown by positions, images, and the best three examples from the word probability. The category corresponds to the formed spatial concept. Each category was classified into an appropriate layer in the hierarchy of spatial concepts. One, four, and 28 categories were classified into the 1st, 2nd, and 3rd layers, respectively. The number of categories in each layer was determined by the nCRP based on the model parameter γ, which controls the probability that the data is allocated to a new category.

The 1st layer included only category 1, into which 900 data were allocated. The high-probability word of category 1 was "laboratory," which referred to the entire experimental environment. Since category 1 contains all the location names, the probabilities for location names becomes relatively low. Nonetheless, the proposed method was able to learn "laboratory," which was given only about 10% to the training dataset, with high probability compared to the second candidate. In the 2nd layer, 343 data in the vicinity of the entrance in the experimental environment were allocated into category 4. The location name of category 4 with the greatest probability was "entrance." The 389 data in the region deeper than the entrance in the experimental environment were categorized into category 5, in which "meeting space" had the greatest probability. In the 3rd layer, the data categorized into categories 4 and 5 in the second layer were further, more finely categorized. In categories 26 and 16, which were formed under category 4, "front of the door" and "front of the chair storage" had the greatest probabilities, respectively. 53 and 81 data were allocated into categories 26 and 16, respectively. Position and image data corresponding to "front of the door" and "front of the chair storage" were finely allocated. These results demonstrated that the proposed method can form not only categories in a lower layer such as "front of the chair storage" and "front of the door" but also categories at higher layers such as "entrance" and "laboratory," and can form its inclusion relations as a hierarchical structure.

TABLE 4 | Mutual information for categorization of location names when changing the number of layers in hLDA with word information and the proposed method with vision, position, and word information.

Method	Modality	2 layers	3 layers	4 layers	5 layers
hLDA	Word	0.87	0.71	0.44	0.41
Proposed method	Vision, position, and word	**0.97**	**1.28**	**0.94**	**0.89**

Mutual information was calculated by Formula 25. Underlined and bold values mean the maximum value in the experimental parameter.

4.4.2. Evaluation of Categorization

To evaluate the effectiveness of multimodal information on hierarchical space categorization, we compared the categorization results of using the proposed method and those obtained using hLDA, which is a hierarchical categorization method with single modality, i.e., based only on word information. Although the number of layers in ground truth in this experiment is 3, robots can not know the number of hierarchies of the spatial concepts in advance. Therefore, in the proposed method and hLDA, categorization was performed with the number of layers changed from 2 to 5. The accuracy of space categorization was evaluated by calculating mutual information between the ground truth, which consisted of a location name given by humans, and the estimated name, which was the best item in the word probability at a category allocated by the proposed method or by hLDA. Mutual information $I(E; G)$ between estimated name E and ground truth G in each layer i and j was calculated by the following formula:

$$I(E; G) = \sum_{g_j \in G} \sum_{e_i \in E} p(e_i, g_j) \log \frac{p(e_i, g_j)}{p(e_i)p(g_j)}. \tag{25}$$

When the mutual information become high, the dependency of e_i and g_j can be regarded as high. By using mutual information, accuracy of categorization can be evaluated when the number of layers on ground truth and estimation result is different. **Table 4** shows the mutual information for categorization results between hLDA with word information and the proposed method with vision, position, and word information in the training data set. The effectiveness of multimodal information in space categorization was clarified, since the proposed method had a high level of mutual information in all layers. In addition, mutual information was maximized when using the same hierarchical number as in the ground truth. In the subsequent evaluations, the number of layers of the proposed method is set to 3.

4.4.3. Evaluation of Name Prediction and Position Category Prediction

We conducted experiments to verify whether or not the proposed method could form hierarchical spatial concepts, which enable a robot to predict location names and position categories similar to predictions made by humans. In the experiment, (1) the influence of multimodal information on the formation of a hierarchical spatial concept was evaluated by comparing the space-categorization results obtained using the proposed method

and using hLDA, which is a hierarchical categorization method with single modality; (2) the similarity between the hierarchical spatial concepts formed by the proposed method and those of humans was evaluated in predicting location names and position categories. The evaluation experiments were performed by cross verification with three data sets that consist of 900 training data and 100 test data with ground truth. The experimental results are indicated by the mean and standard deviation in the three data sets.

To verify whether or not the proposed method can form hierarchical spatial concepts, accuracy evaluation of name prediction and position category prediction through spatial concept use was performed. In the evaluation of name prediction, vision, position, and word information were given to the robot at the training data points. In the test data points, only vision and position information were given. Therefore, the robot has to predict the unobserved word information from the observed vision and position information. **Table 5** shows the accuracy of name prediction using the baseline methods, the proposed method, and those made by humans. The most frequent class, nearest neighbor (position and word), nearest neighbor (vision, position, and word), multimodal HDP, and spatial concept formation model were used as the baseline methods. The accuracy of name prediction was calculated by Formula (13) at global, intermediate, and local layers in ground truth. The proposed method and humans predicted location names in three layers. The results of humans consisted of the average accuracy of three subjects familiar with the experimental environment.

Compared with the accuracy obtained using the baseline methods, higher accuracies were obtained by the proposed method in the 1st, 2nd, and 3rd layers. It was assumed that weak features buried in the lower layer in the baseline methods were categorized as features of the higher layer in the proposed method. The proposed method enabled a robot to predict location names close to predictions made by humans by selecting the appropriate layer depending on the situation.

Table 6 shows the evaluation results of position category prediction using the baseline methods, the proposed method, and those made by humans. In the evaluation, the most frequent class, nearest neighbor (position and word), multimodal HDP, and spatial concept formation model were used as the baseline methods. The position category prediction was evaluated in terms of precision, recall, and F-measure, which were calculated by Formula (14).

Compared with results obtained by the baseline methods, higher values of precision and recall were obtained by the proposed method in the global and intermediate layers. In the local layer, higher values of precision and recall were obtained by Nearest neighbor and Spatial Concept Formation (SpCoFo), respectively. However, in the F-measure, which is a harmonic mean between precision and recall, the proposed method has the largest values in the global, intermediate, and local layers. The reason why the recall and F-measure values were lower than the precision is that only 100 data points were predicted and plotted for regions with 100 grids or more, as shown in **Figure 7**. In the result of F-measure, independent *t*-tests were performed in nine samples consisting of three data

TABLE 5 | Accuracy of name prediction using the baseline methods, the proposed method, and those made by humans; the accuracy was calculated by using Formula (13).

Method	Modality	Layer	Mean (s.d.)		
			Global	Intermediate	Local
Most frequent class	Position and word		**1.00 (0.00)**	0.18 (0.32)	0.09 (0.02)
Nearest neighbor	Position and word		0.12 (0.01)	0.24 (0.02)	0.20 (0.03)
Nearest neighbor	Vision, position and word		0.18 (0.03)	0.28 (0.04)	0.31 (0.04)
Multimodal HDP	Vision, position, and word		0.13 (0.02)	0.54 (0.06)	0.24 (0.07)
SpCoFo	Vision, position, and word		0.25 (0.13)	0.23 (0.15)	0.36 (0.13)
		1st	**1.00 (0.00)**	0.00 (0.00)	0.00 (0.00)
Proposed method	Vision, position, and word	2nd	0.00 (0.00)	**0.96 (0.04)**	0.01 (0.02)
		3rd	0.00 (0.00)	0.04 (0.04)	**0.55 (0.07)**
		1st	1.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Humans		2nd	0.00 (0.00)	0.98 (0.02)	0.00 (0.00)
		3rd	0.00 (0.00)	0.03 (0.04)	0.74 (0.10)

The accuracy is indicated by the mean and standard deviation (s.d.). Underlined and bold values mean the maximum value in the experimental parameter.

sets with three types of ground truth: global, intermediate, and local. In the proposed method, the p-values of the Most frequent class, Nearest neighbor, multimodal HDP, and SpCoFo were 0.00012, 0.00004, 0.00003, and 0.00051, respectively, and significant differences were observed with ($p < 0.05$). As the reason why the result of humans were not perfect, some errors were found in the boundary of the place. For example, the boundary between "Book shelf zone" and "front of the table," and the edge of the region called "front of the door" were different depending on the human. The centricity of the place is consistent, but the region includes ambiguity even among humans. The experimental results show that the proposed method enabled a robot to predict position categories closer to predictions made by humans than possible using the baseline methods.

In the experiments for location name and position category prediction, the proposed method showed higher performance than the baseline methods. In the baseline methods, i.e., multimodal HDP and SpCoFo, since the feature space is classified uniformly, the location concepts are formed non-hierarchically. For example, an upper concept, e.g., meeting space, is embedded in the lower concepts, e.g., front of the table and front of the display. Therefore, the place called "Meeting space" is learned as a place different from the places called "front of the table" and "front of the display." Since the proposed method forms concepts by extracting the similarity of knowledge in the upper concept, it is possible to form an upper concept without interfering with the formation of the lower concept. For this reason, the proposed method was able to show high performance in the experiments of name and position category prediction with global, intermediate, and local.

In human-robot interactions in home environments, location names as word information are given to only a part of the training data from a user. We evaluated the robustness of the proposed method in terms of the naming rate in order to verify how name and position category prediction performance changes with decreasing naming rate. In this experiment, the formation of spatial concepts using the proposed method was performed

using the training data with the naming rate changed to 1, 2, 5, 10, and 20% successively. The naming rates of 1 or 20% mean that 9 or 180 of the 900 training data contained location names, while the remaining data did not contain any location name. **Table 7** shows the accuracy of name prediction and the F-measure of position category prediction for each naming rate. In the results of name prediction and position category prediction, it was confirmed that learning progresses in the global layer earlier than in the intermediate and local layers. It was clarified that overall prediction ability did not decrease greatly owing to the decreased naming rate, but gradually decreased from the lower layer. In this experiment, we performed spatial concept formation without prior knowledge in only one environment, but it is possible to increase learning efficiency by giving parameters of models estimated in other environments as prior probabilities. The transfer learning of spatial concepts will be performed in the future.

5. APPLICATION EXAMPLES FOR HUMAN SUPPORT ROBOTS

Application examples of the hierarchical spatial concept using the proposed method are demonstrated in this section. We implemented the proposed method for the Toyota human support robot (HSR)[6] and created application examples in which the robot moves based on human linguistic instructions and describes its self-position linguistically in an experimental field assuming a home environment.

The home environment and the robot used are shown in **Figure 9**. There were two tables as shown in **Figure 9A**, A and B. In the environment, whether the robot could move based on linguistic instructions including the hierarchical structure of spaces such as "front of the table in the living room" and "front of the table in the dining room" was verified. In **Figure 9B**, an

---

[6]Toyota Global Site—Partner Robot Family: http://www.toyota-global.com/innovation/partner_robot/family_2.html

TABLE 6 | Precision, recall, and F-measure evaluation of position category prediction using the baseline methods, the proposed method, and those made by humans in global, intermediate, and local; the precision, recall, and F-measure were calculated by using Formula (14).

Method	Precision			Recall			F-measure		
	Global	Intermediate	Local	Global	Intermediate	Local	Global	Intermediate	Local
Most frequent class	**1.00 (0.01)**	0.49 (0.01)	0.37 (0.03)	0.12 (0.02)	0.17 (0.02)	0.15 (0.03)	0.22 (0.03)	0.25 (0.02)	0.20 (0.02)
Nearest neighbor	**1.00 (0.00)**	0.93 (0.03)	**0.67 (0.03)**	0.12 (0.03)	0.26 (0.04)	0.23 (0.04)	0.22 (0.04)	0.41 (0.05)	0.33 (0.03)
Multimodal HDP	**1.00 (0.00)**	0.95 (0.02)	0.53 (0.03)	0.12 (0.01)	0.26 (0.04)	0.26 (0.02)	0.21 (0.02)	0.40 (0.05)	0.33 (0.02)
SpCoFo	0.82 (0.00)	0.62 (0.04)	0.35 (0.04)	0.16 (0.01)	0.32 (0.02)	**0.38 (0.04)**	0.27 (0.02)	0.42 (0.01)	0.35 (0.04)
Proposed method	**1.00 (0.00)**	**0.96 (0.03)**	0.59 (0.05)	**0.18 (0.01)**	**0.34 (0.02)**	0.36 (0.04)	**0.30 (0.02)**	**0.50 (0.02)**	**0.43 (0.01)**
Humans	1.00	0.99	0.76	0.19	0.50	0.49	0.32	0.65	0.56

In the experiment, the modalities of the nearest neighbor were position and word. The results are indicated by the mean and standard deviation as mean (s.d.). Underlined and bold values mean the maximum value in the experimental parameter.

TABLE 7 | Robustness evaluation of the proposed method with respect to naming rate: accuracy in name prediction indicates the maximum value of the three layers.

Naming rate	Name prediction (accuracy)			Position prediction (F-measure)		
	Global	Intermediate	Local	Global	Intermediate	Local
1%	1.00	0.68	0.14	0.29	0.46	0.30
2%	1.00	0.77	0.26	0.29	0.47	0.31
5%	1.00	0.92	0.35	0.28	0.36	0.37
10%	1.00	0.92	0.58	0.30	0.46	0.37
20%	1.00	0.92	0.63	0.31	0.50	0.44
Humans	1.00	0.96	0.76	0.32	0.65	0.56
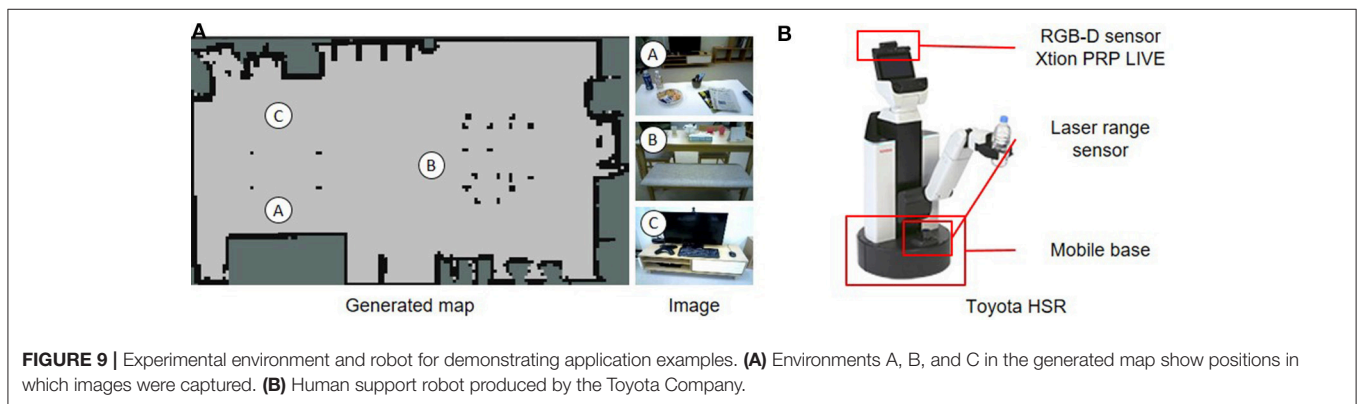
FIGURE 9 | Experimental environment and robot for demonstrating application examples. **(A)** Environments A, B, and C in the generated map show positions in which images were captured. **(B)** Human support robot produced by the Toyota Company.

RGB-D sensor and a laser range sensor were used to capture images and to estimate self-position, respectively. The packages[7]: hector_slam and omni_base for mapping, localization, and moving were used with ROS Indigo[8] to navigate the robot to the predicated position.

The robot collected 715 training data consisting of images, positions, and word information and formed a hierarchical spatial concept using the proposed method. Location names were given to 20% of total training data. Rospeex (Sugiura and Zettsu, 2015) was used to recognize human speech instructions and convert them into text information. In the experiment, the dimensions of the information vectors w^v, w^p, and w^w were 1,000, 64, and 16, respectively.

The two places predicted by Formula (12) based on the speech instructions, i.e., "go to the front of the table in the living room" and "go to the front of the table in the dining room" are shown in **Figures 10A,B**, respectively. Predicted position categories indicated by red dots show that the "front of the table in the living room" and the "front of the table in the dining room" were recognized as different places using the space concept in the higher layer.

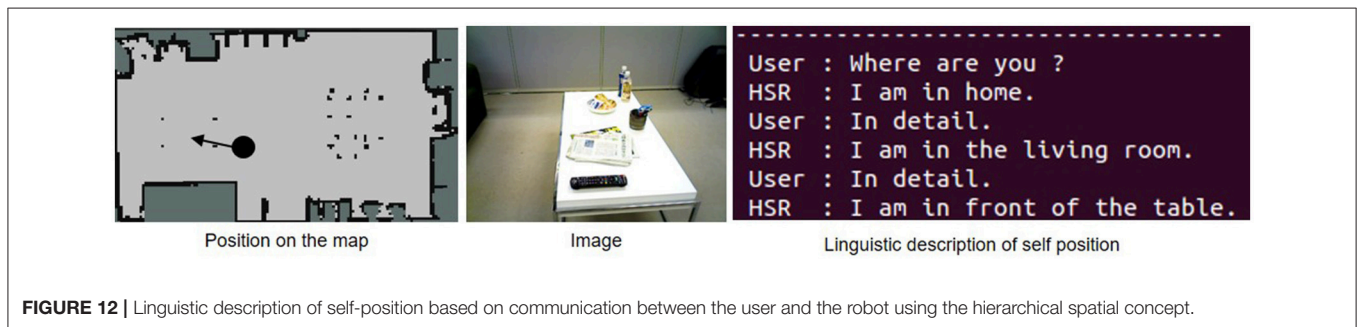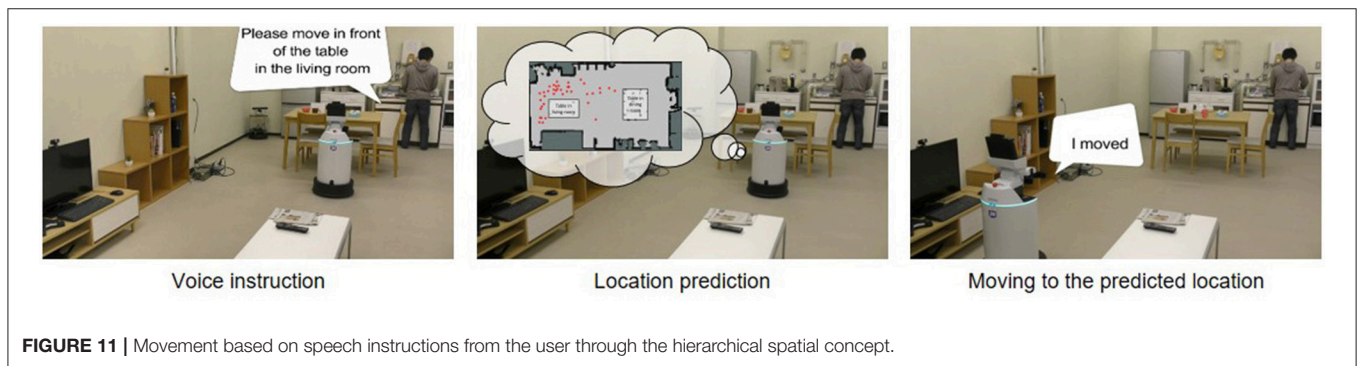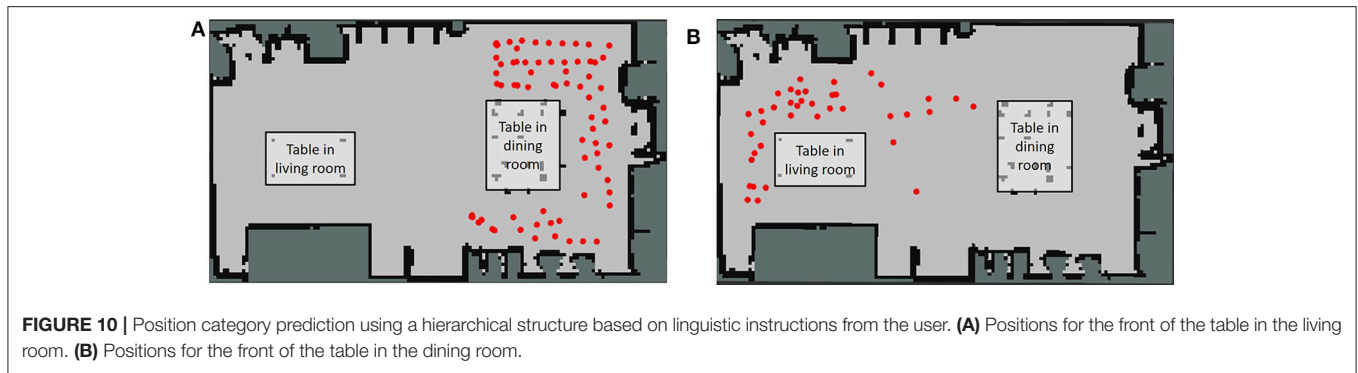[7]hector_slam: http://wiki.ros.org/hector_slam
[8]ROS Indigo: http://wiki.ros.org/indigo

FIGURE 10 | Position category prediction using a hierarchical structure based on linguistic instructions from the user. **(A)** Positions for the front of the table in the living room. **(B)** Positions for the front of the table in the dining room.

FIGURE 11 | Movement based on speech instructions from the user through the hierarchical spatial concept.

FIGURE 12 | Linguistic description of self-position based on communication between the user and the robot using the hierarchical spatial concept.

Figure 11 shows how the robot moved based on human speech instructions in the experiment. The robot recognized human speech instructions using rospeex and predicted position categories with the Formula (12) using a hierarchical spatial concept. It moved to the instructed place by sampling randomly from the predicted positions. **Figure 12** shows an application example in which the robot described its self-position linguistically. The robot observed its self-position and image and predicted the name of its self-position by calculating Formula 11 using the hierarchical spatial concept. As shown in the left side of **Figure 12**, the proposed method enabled the robot to describe its self-position linguistically with different layers. We demonstrated application examples using the formed hierarchical spatial concept in the service scene in a home environment. The movie of the demonstration and training dataset can be found at the URL[9].

---

[9]Multimedia - emlab page: https://emlab.jimdo.com/multimedia/

6. CONCLUSIONS

We assumed that a computational model that considers the hierarchical structure of space enables robots to predict the name and position of a space close to the corresponding prediction by humans. In our assumptions, we proposed a hierarchical spatial concept formation method based on a Bayesian generative model with multimodal information, i.e., vision, position, and word information, and developed a robot that can predict unobserved location names and position categories based on observed information using the formed hierarchical spatial concept. We conducted experiments to form a hierarchical spatial concept using a robot and evaluated its ability in name prediction and position category prediction.

The experimental results for name and position category prediction demonstrated that, relative to baseline methods, the proposed method enabled the robot to predict location names and position categories closer to predictions made by

humans. Application examples using the hierarchical spatial concept in a home environment demonstrated that a robot could move to an instructed place based on human speech instructions and describe its self-position linguistically through the formed hierarchical spatial concept. The experimental results and application example demonstrated that the proposed method enabled the robot to form spatial concepts in abstract layers and use the concepts for human-robot communications in a home environment. This study showed that it the name and position of a location could be predicted, even in a home, using generalized spatial concepts. Furthermore, by conducting additional learning in each house, a spatial concept adapted to the environment can be formed.

The limitation of this study is as follows. In the feature extraction of the position information, hierarchical k-means method was utilized to convert the position information (x, y) into the position histogram. In the experiment, 389 and 511 data were allocated to two clusters at the top layer c_1. In the bottom layer c_6, the number and standard deviation of the data allocated to each of the 64 clusters were 14.1 and 12.2, respectively. There is some bias between the clusters. The hierarchical k-means makes it possible to convert the position information into the position histogram including hierarchical spatial features. However, nearby data points at a classification boundary, which are classified into different clusters on a high level, are regarded as very different. We are considering a method to reduce bias in space while maintaining hierarchical features of space. As for the number of location names, at section 4 and 5 in the experiments, the numbers of location names were 15 and 16, respectively. The number of location names increases with increase in the numbers of teachings and users. If the robot learns the location names from several users over a long term, an algorithm to remove location names with low probability of observation is needed in order to improve the learning efficiency.

As future work, we will generalize the spatial concepts for various environments and perform transition learning of spatial concepts with the generalized spatial concepts as prior knowledge.

AUTHOR CONTRIBUTIONS

YH designed the study, and wrote the initial draft of the manuscript. HK and MI contributed to analysis and interpretation of data, and assisted in the preparation of the manuscript. TT has contributed to data collection and interpretation, and critically reviewed the manuscript. All authors approved the final version of the manuscript, and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

FUNDING

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnbot.2018.00011/full#supplementary-material

REFERENCES

Aldous, D. J. (1985). "Exchangeability and related topics," in *École d'Été de Probabilités de Saint-Flour XIII–1983*, Lecture Notes in Mathematics, Vol. 1117 (Berlin; Heidelberg: Springer), 1–198.

Ando, Y., Nakamura, 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 (IROS)* (Tokyo), 2272–2279.

Blei, D. M., Griffiths, T. L., and Jordan, M. I. (2010). The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *J. ACM* 57, 7:1–7:30. doi: 10.1145/1667053.1667056

Dellaert, F., Fox, D., Burgard, W., and Thrun, S. (1999). "Monte Carlo localization for mobile robots," in *Proceedings of 1999 IEEE International Conference on Robotics and Automation* (Detroit, MI), 1322–1328.

Durrant-Whyte, H., and Bailey, T. (2006). Simultaneous localization and mapping: Part I. *IEEE Robot. Autom. Mag.* 13, 99–110. doi: 10.1109/MRA.2006.1638022

Espinace, P., Kollar, T., Roy, N., and Soto, A. (2013). Indoor scene recognition by a mobile robot through adaptive object detection. *Robot. Auton. Syst.* 61, 932–947. doi: 10.1016/j.robot.2013.05.002

Friedman, J. H., Bentley, J. L., and Finkel, R. A. (1977). An algorithm for finding best matches in logarithmic expected time. *ACM Trans. Math. Softw.* 3, 209–226. doi: 10.1145/355744.355745

Galindo, C., Madrigal, J. A., Gonzalez, J., and Saffiotti, A. (2008). Robot task planning using semantic maps. *Robot. Auton. Syst.* 56, 955–966. doi: 10.1016/j.robot.2008.08.007

Granda, N. C., Rogers, J. G., Trevor, A. J., and Christensen, H. I. (2010). "Semantic map partitioning in indoor environments using regional analysis," in *IEEE International Conference on Intelligent Robots and Systems* (Taipei), 1451–1456.

Gu, Z., Taguchi, R., Hattori, K., Hoguro, M., and Umezaki, T. (2016). "Learning of relative spatial concepts from ambiguous instructions," in *13th IFAC/IFIP/IFORS/IEA Symposium on Analysis, Design, and Evaluation of Human-Machine Systems*, Vol. 49 (Kyoto), 150–153.

Hagiwara, Y., Inoue, M., and Taniguchi, T. (2016). "Place concept learning by hMLDA based on position and vision information," in *13th IFAC/IFIP/IFORS/IEA Symposium on Analysis, Design, and Evaluation of Human-Machine Systems*, Vol. 49 (Kyoto), 216–220.

Harris, Z. (1954). Distributional structure. *Word* 10, 146–162. doi: 10.1080/00437956.1954.11659520

Heath, S., Ball, D., and Wiles, J. (2016). Lingodroids: cross-situational learning for episodic elements. *IEEE Trans. Cogn. Dev. Syst.* 8, 3–14. doi: 10.1109/TAMD.2015.2442619

Ishibushi, S., Taniguchi, A., Takano, T., Hagiwara, Y., and Taniguchi, T. (2015). "Statistical localization exploiting convolutional neural network for an autonomous vehicle," in *41th Annual Conference of the IEEE Industrial Electronics Society (IECON)* (Yokohama), 1369–1375.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., et al. (2014). "Caffe: convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia* (ACM) (Orlando, FL), 675–678.

Kostavelis, I., and Gasteratos, A. (2013). Learning spatially semantic representations for cognitive robot navigation. *Robot. Auton. Syst.* 61, 1460–1475. doi: 10.1016/j.robot.2013.07.008

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* (Lake Tahoe, NV), 1097–1105.

Mozos, O. M., and Burgard, W. (2006). "Supervised learning of topological maps using semantic information extracted from range data," in *IEEE International Conference on Intelligent Robots and Systems*, 2772–2777.

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

Nielsen, C. W., Ricks, B., Goodrich, M. A., Bruemmer, D., Few, D., and Few, M. (2004). "Snapshots for semantic maps," in *IEEE International Conference on Systems*, Vol. 3 (The Hague), 2853–2858.

Pitman, J. (2006). *Combinatorial Stochastic Processes*. Lecture Notes in Mathematics. Berkeley, CA: Springer-Verlag.

Pitman, J. (2002). *Combinatorial Stochastic Processes*. Technical Report of Department of Statistics, UC Berkeley, 2002. Lecture notes for St. Flour Course, 621.

Pronobis, A., Caputo, B., Jensfelt, P., and Christensen, H. I. (2006). "A discriminative approach to robust visual place recognition," in *IEEE International Conference on Intelligent Robots and Systems* (Beijing), 3829–3836.

Pronobis, A., and Jensfelt, P. (2012). "Large-scale semantic mapping and reasoning with heterogeneous modalities," in *IEEE International Conference on Robotics and Automation* (Saint Paul, MN), 3515–3522. doi: 10.1109/ICRA.2012.6224637

Ranganathan, A., and Dellaert, F. (2007). "Semantic modeling of places using objects," in *Proceedings of the 2007 Robotics: Science and Systems Conference*, Vol. 3 (Atlanta, GA), 27–30.

Rangel, J. C., Martínez-Gómez, J., García-Varea, I., and Cazorla, M. (2017). Lextomap: lexical-based topological mapping. *Adv. Robot.* 31, 268–281. doi: 10.1080/01691864.2016.1261045

Rusu, R. B., Marton, Z. C., Blodow, N., Holzbach, A., and Beetz, M. (2009). "Modelbased and learned semantic object labeling in 3d point cloud maps of kitchen environments," in *IEEE International Conference on Intelligent Robots and Systems* (St. Louis, MO), 3601–3608.

Rusu, R. B., Marton, Z. C., Blodow, N., Dolha, M., and Beetz, M. (2008). Towards 3d point cloud based object maps for household environments. *Robot. Auton. Syst.* 56, 927–941. doi: 10.1016/j.robot.2008.08.005

Sugiura, K., and Zettsu, K. (2015). "Rospeex: a cloud robotics platform for human-robot spoken dialogues," in *Proceedings of 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems* (Hamburg), 6155–6160.

Taniguchi, A., Taniguchi, T., and Inamura, T. (2016). Spatial concept acquisition for a mobile robot that integrates self-localization and unsupervised word discovery from spoken sentences. *IEEE Trans. Cogn. Dev. Syst.* 8, 285–297. doi: 10.1109/TCDS.2016.2565542

Taniguchi, A., Yoshizaki, H., Inamura, T., and Taniguchi, T. (2014). Research on simultaneous estimation of self-location and location concepts. *Trans. Inst. Syst. Control Inform. Eng.* 27, 166–177. doi: 10.5687/iscie.27.166

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 17*, eds L. K. Saul, Y. Weiss, and L. Bottou (Long Beach, CA: MIT Press), 1385–1392.

Zender, H., Mozos, O. M., Jensfelt, P., Kruijff, G. J., and Burgard, W. (2008). Conceptual spatial representations for indoor mobile robots. *Robot. Auton. Syst.* 56, 493–502. doi: 10.1016/j.robot.2008.03.007