2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
September 24–28, 2017, Vancouver, BC, Canada

Online Spatial Concept and Lexical Acquisition with Simultaneous Localization and Mapping

Akira Taniguchi[1], Yoshinobu Hagiwara[1], Tadahiro Taniguchi[1] and Tetsunari Inamura[2]

***Abstract*— In this paper, we propose an online learning algorithm based on a Rao-Blackwellized particle filter for spatial concept acquisition and mapping. We have proposed a nonparametric Bayesian spatial concept acquisition model (SpCoA). We propose a novel method (SpCoSLAM) integrating SpCoA and FastSLAM in the theoretical framework of the Bayesian generative model. The proposed method can simultaneously learn place categories and lexicons while incrementally generating an environmental map. Furthermore, the proposed method has scene image features and a language model added to SpCoA. In the experiments, we tested online learning of spatial concepts and environmental maps in a novel environment of which the robot did not have a map. Then, we evaluated the results of online learning of spatial concepts and lexical acquisition. The experimental results demonstrated that the robot was able to more accurately learn the relationships between words and the place in the environmental map incrementally by using the proposed method.**

I. INTRODUCTION

Robots coexisting with humans and operating in various environments are required to adaptively learn and use the spatial concepts and vocabulary related to different places. However, spatial concepts are such that their target domain may be unclear compared with object concepts and may differ according to the user and environment. Therefore, it is difficult to manually design spatial concepts in advance, and it is desirable for robots to autonomously learn spatial concepts based on their own experiences.

The related research fields of semantic mapping and place categorization [1], [2] have attracted considerable interest in recent years. However, most of these studies have consisted of separate independent methods of semantics of places and mapping using simultaneous localization and mapping (SLAM) [3]. In addition, the semantics of places, place categories, and names of places could only be learned from pre-set values. In this paper, we propose a novel unsupervised Bayesian generative model and an online learning algorithm that can perform simultaneous learning of the spatial concepts and an environmental map from multimodal information. The proposed method can automatically and sequentially perform place categorization and learn unknown words without prior knowledge.

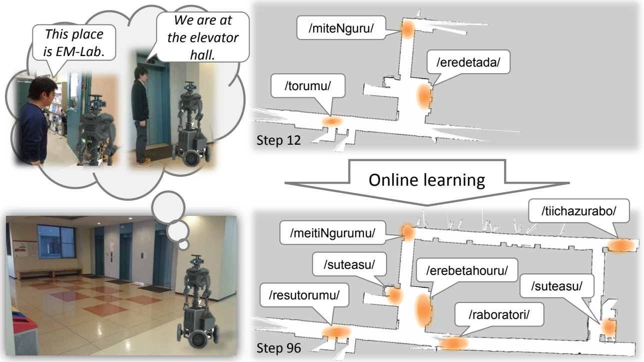


Fig. 1. Overview of online learning of spatial concepts and an environmental map; We aim to develop a method that enables mobile robots to learn spatial concepts, a lexicon and an environmental map sequentially from interaction with an environment and human, even in an unknown environment without prior knowledge.

Taniguchi et al. [4] proposed a method that integrated ambiguous speech-recognition results with the self-localization method for learning spatial concepts. In addition, Taniguchi et al. [5] proposed the nonparametric Bayesian spatial concept acquisition method (SpCoA) based on an unsupervised word-segmentation method known as latticelm [6]. On the other hand, Ishibushi et al. [7] proposed a self-localization method that exploits image features using a convolutional neural network (CNN) [8]. These methods [4], [5], [7] cannot cope with changes in the names of places and the environment because these methods use batch learning algorithms. In addition, these methods cannot learn spatial concepts from unknown environments without a map, i.e., the robot needs to have a map generated by SLAM beforehand. Therefore, in this paper, we develop an online algorithm that can sequentially learn a map, spatial concepts integrating positions, speech signals, and scene images.

FastSLAM [9], [10] has realized an on-line algorithm for efficient self-localization and mapping using a Rao-Blackwellized particle filter (RBPF) [11]. In this paper, we introduce a grid-based FastSLAM algorithm in the generative model for spatial concept acquisition. The graphical model of SpCoA has integrated spatial lexical acquisition into Monte Carlo localization (MCL), a particle-filter-based self-localization method. SpCoA can be extended naturally to SLAM. Therefore, we assume that the robot can learn vocabulary related to places and a map sequentially.

One of the important problems of our research is unsupervised lexical acquisition. There are research efforts on

*This work was partially supported by JST, CREST.

[1] Akira Taniguchi, Yoshinobu Hagiwara and Tadahiro Taniguchi are with Ritsumeikan University, 1-1-1 Noji-Higashi, Kusatsu, Shiga 525-8577, Japan {a.taniguchi, yhagiwara, taniguchi} @em.ci.ritsumei.ac.jp

[2] Tetsunari Inamura is with National Institute of Informatics / SOKENDAI (The Graduate University for Advanced Studies), 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan inamura@nii.ac.jp

978-1-5386-2681-8/17/$31.00 ©2017 IEEE

incremental spatial language acquisition through robot-to-robot interaction [12], [13]. However, these studies [12], [13] did not consider lexical acquisition through human-to-robot speech interactions (HRSI). For online unsupervised lexical acquisition by HRSI, it is necessary to deal with the problems of phoneme recognition errors and word segmentation of uttered sentences containing errors. SpCoA reduced phoneme recognition errors of word segmentation by using the weighted finite-state transducer (WFST)-based unsupervised word segmentation method latticelm [6]. Araki et al. [14] performed a pseudo-online algorithm using the nested Pitman–Yor language model (NPYLM) [15]. However, these studies [5], [14] have reported that word segmentation of speech recognition results including errors causes over-segmentation [16]. In this paper, we will improve the accuracy of speech recognition by updating the language models sequentially.

We assume that the robot has not acquired any vocabulary in advance, and can recognize only phonemes or syllables. We represent the spatial area of the environment in terms of a *position distribution*. Furthermore, we define a *spatial concept* as a place category that includes place names, scene image features, and the position distributions corresponding to those names.

The goal of this study is to develop a robot that learns spatial concepts incrementally from multimodal information obtained while moving in the environment. The main contributions of this paper are as follows.

- We propose an online algorithm based on RBPF for spatial concept acquisition. The proposed method integrates SpCoA and FastSLAM in the theoretical framework of the Bayesian generative model.
- We demonstrated that a robot without a pre-existing lexicon or map can learn spatial concepts and an environmental map incrementally.

II. Online Spatial Concept Acquisition

A. Overview

An overview of the proposed method is shown in Fig. 1. The proposed method is an online spatial concept acquisition and simultaneous localization and mapping (SpCoSLAM). The proposed method can learn sequential spatial concepts for unknown environments and unsearched regions without maps. In addition, it can mutually complement the uncertainty of information by using multimodal information. A pseudo-code for the online learning is given in Algorithm 1. The procedure of SpCoSLAM for each step is described as follows. 2) – 6) are performed for each particle.

1) A robot gets WFST speech recognition results of the user's speech signals using a language model of the previous step. (line 3 in Algorithm 1)
2) The robot gets the observation likelihood by performing a sample motion model and a measurement model of FastSLAM. (line 5-10)
3) The robot performs unsupervised word segmentation latticelm [6] using WFST speech recognition results. (line 11)

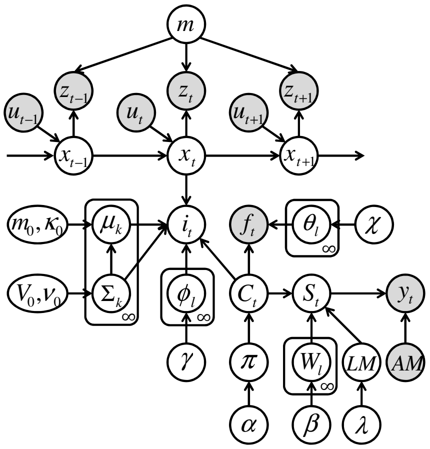


Fig. 2. Graphical model representation of SpCoSLAM; It expresses multimodal place categorization, lexical acquisition and SLAM as one Bayesian generative model. Gray nodes indicate observation variables.

TABLE I

Each element of the graphical model of SpCoSLAM

x_t	Self-position of a robot
z_t	Sensor data (depth data)
u_t	Control data
f_t	Image feature
y_t	Speech signal
S_t	Sequence of words (word segmentation result)
C_t	Index of spatial concepts
i_t	Index of position distributions
m	Environmental map
π	Multinomial distribution of index C_t of spatial concepts
ϕ_l	Multinomial distribution of index i_t of position distribution
μ_k, Σ_k	Position distribution (mean vector, covariance matrix)
θ_l	Multinomial distribution of image feature
W_l	Multinomial distribution of the names of places
LM	Language model (word dictionary)
AM	Acoustic model for speech recognition
α,β,γ,χ,λ m_0,κ_0,V_0,ν_0	Hyperparameters of prior distributions

4) The robot gets latent variables of spatial concepts by sampling. The details of this process are described in Section II-E. (line 12)
5) The robot gets the marginal likelihood of observation data as the importance weight. (line 13-15)
6) The robot updates an environmental map. (line 16)
7) The robot estimates the set of parameters of spatial concepts from data and sampled values. (line 17)
8) The robot updates a language model of the maximum weight for next step. (line 20-21)
9) The robot performs resampling of particles according to weights. (line 22-25)

B. Definition of generative model and graphical model

Figure 2 shows the graphical model of SpCoSLAM and Table I lists each variable of the graphical model. We describe the formulation of the generation process represented by the graphical model as follows:

$$
\begin{aligned}
\pi &\sim \mathrm{DP}(\alpha) && (1)\\
C_t &\sim \mathrm{Mult}(\pi) && (2)\\
\phi_l &\sim \mathrm{DP}(\gamma) && (3)\\
W_l &\sim \mathrm{Dir}(\beta) && (4)\\
LM &\sim p(LM \mid \lambda) && (5)\\
S_t &\sim p(S_t \mid \mathbf{W}, C_t, LM) && (6)\\
y_t &\sim p(y_t \mid S_t, AM) && (7)\\
\theta_l &\sim \mathrm{Dir}(\chi) && (8)\\
f_t &\sim \mathrm{Mult}(\theta_{C_t}) && (9)\\
\Sigma_k &\sim \mathrm{IW}(\Sigma \mid V_0, \nu_0) && (10)\\
\mu_k &\sim \mathrm{N}(\mu \mid m_0, (\Sigma_k/\kappa_0)) && (11)\\
x_t &\sim p(x_t \mid x_{t-1}, u_t) && (12)\\
z_t &\sim p(z_t \mid x_t, m) && (13)\\
i_t &\sim p(i_t \mid x_t, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\phi}, C_t) && (14)
\end{aligned}
$$

where $\mathrm{DP}()$ represents Dirichlet process, $\mathrm{Mult}()$ is multinomial distribution, $\mathrm{Dir}()$ is Dirichlet distribution, $\mathrm{IW}()$ is inverseWishart distribution, and $\mathrm{N}()$ is Gaussian distribution.

Equation (6) approximates by using unigram rescaling [17], as shown in (15). $\overset{\mathrm{UR}}{\approx}$ represents the approximation by unigram rescaling.

$$
\begin{aligned}
&p(S_t \mid \mathbf{W}, C_t, LM)\\
&\overset{\mathrm{UR}}{\approx} \quad p(S_t \mid LM) \prod_{B_t} \frac{\mathrm{Mult}(S_{t,b} \mid W_{C_t})}{\sum_{c'} \mathrm{Mult}(S_{t,b} \mid W_{c'})} \qquad (15)
\end{aligned}
$$

where B_t denotes the number of words in the sentence.

Then, the probability distribution for (14) can be defined as follows:

$$
\begin{aligned}
&p(i_t \mid x_t, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\phi}, C_t)\\
&= \frac{\mathrm{N}(x_t \mid \mu_{i_t}, \Sigma_{i_t}) \mathrm{Mult}(i_t \mid \phi_{C_t})}{\sum_{i_t = j} \mathrm{N}(x_t \mid \mu_j, \Sigma_j) \mathrm{Mult}(j \mid \phi_{C_t})} . \qquad (16)
\end{aligned}
$$

C. Formulation of the speech recognition and the unsupervised word segmentation

The 1-best speech recognition and the WFST speech recognition are represented as follows:

$$
\begin{aligned}
S_t^{(\text{1-best})} &= \underset{S_t}{\mathrm{argmax}}\, \mathrm{SR}(S_t \mid y_t, AM, LM) && (17)\\
\mathcal{L}_t &\approx \mathrm{SR}(\mathcal{L}_t \mid y_t, AM, LM) && (18)
\end{aligned}
$$

where $\mathcal{L}_t$ denotes the speech recognition result of WFST format, which is a word graph representing the speech recognition results. The unsupervised word segmentation of WFST by latticelm [6] is represented as follows:

$$
S_{T_o} \sim \mathit{latticelm}(S_{T_o} \mid \mathcal{L}_{T_o}, \lambda). \qquad (19)
$$

Algorithm 1 Online learning algorithm of SpCoSLAM

```
procedure SpCoSLAM($X_{t-1}, u_t, z_t, f_{1:t}, y_{1:t}$)
    $\bar{X}_t = X_t = \emptyset$
    $\mathcal{L}_{1:t} = \mathrm{SR}(\mathcal{L}_{1:t} \mid y_{1:t}, AM, LM_{t-1})$
    for $r = 1$ to $R$ do
        $\acute{x}_t^{[r]} =$ sample_motion_model$(u_t, x_{t-1}^{[r]})$
        $x_t^{[r]} =$ scan_matching$(z_t, \acute{x}_t^{[r]}, m_{t-1}^{[r]})$
        for $j = 1$ to $J$ do
            $x_j =$ sample_motion_model$(u_t, x_{t-1}^{[r]})$
        end for
        $\omega_z^{[r]} = \sum_{j=1}^{J}$ measurement_model$(z_t, x_j, m_{t-1}^{[r]})$
        $S_{1:t}^{[r]} \sim \mathit{latticelm}(S_{1:t} \mid \mathcal{L}_{1:t}, \lambda)$
        $i_t^{[r]}, C_t^{[r]} \sim p(i_t, C_t \mid x_{0:t}^{[r]}, i_{1:t-1}^{[r]}, C_{1:t-1}^{[r]}, S_{1:t}^{[r]}, f_{1:t}, \mathbf{h})$
        $\omega_f^{[r]} = p(f_t \mid C_{1:t-1}^{[r]}, f_{1:t-1}, \alpha, \chi)$
        $\omega_s^{[r]} = p(S_t^{[r]} \mid S_{1:t-1}^{[r]}, C_{1:t-1}^{[r]}, \alpha, \beta) / p(S_t^{[r]} \mid S_{1:t-1}^{[r]}, \beta)$
        $\omega_t^{[r]} = \omega_z^{[r]} \cdot \omega_f^{[r]} \cdot \omega_s^{[r]}$
        $m_t^{[r]} =$ updated_occupancy_grid$(z_t, x_t^{[r]}, m_{t-1}^{[r]})$
        $\Theta_t^{[r]} = E[p(\Theta \mid x_{0:t}^{[r]}, \mathbf{C}_{1:t}^{[r]}, f_{1:t}, \mathbf{h})]$
        $\bar{X}_t = \bar{X}_t \cup \langle x_{0:t}^{[r]}, \mathbf{C}_{1:t}^{[r]}, m_t^{[r]}, \Theta_t^{[r]}, \omega_t^{[r]} \rangle$
    end for
    $S_{1:t}^{*} = \mathrm{argmax}_{S_{1:t}^{[r]}} \sum_{r=1}^{R} \omega_t^{[r]} \delta(S_{1:t} - S_{1:t}^{[r]})$
    $LM_t = \mathrm{argmax}_{LM}\, p(LM \mid S_{1:t}^{*}, \lambda)$
    for $r = 1$ to $R$ do
        draw $i$ with probability $\propto \omega_t^{[i]}$
        add $\langle x_{0:t}^{[i]}, \mathbf{C}_{1:t}^{[i]}, m_t^{[i]}, \Theta_t^{[i]}, LM_t \rangle$ to $X_t$
    end for
    return $X_t$
end procedure
```

D. Online spatial concept acquisition and mapping

Here, we describe the derivation of formulas for the online algorithm. The online learning algorithm of the proposed method can be derived by introducing sequential update equations for estimating the parameters of the spatial concepts into the formulation of FastSLAM based on RBPF. The proposed method assumes grid-based FastSLAM 2.0 [9], [10] algorithm. Algorithm 1 is the online learning algorithm of SpCoSLAM. As an advantage of using a particle filter, parallel processing can be easily applied because each particle can be calculated independently.

In the formulation of FastSLAM, the joint posterior distribution is factorized as follows:

$$
\begin{aligned}
&p(x_{0:t}, m \mid u_{1:t}, z_{1:t})\\
&= \underbrace{p(m \mid x_{0:t}, z_{1:t})}_{\text{Mapping}} \underbrace{p(x_{0:t} \mid u_{1:t}, z_{1:t})}_{\text{Particle filter}} . \qquad (20)
\end{aligned}
$$

This factorization represents a decomposition into two calculations: the mapping and self-localization by RBPF.

In the formulation of SpCoSLAM, the joint posterior distribution can be factorized to the probability distributions of a language model LM, a map m, the set of model parameters of spatial concepts $\Theta = \{\mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \theta, \phi, \pi\}$, and the joint distribution of trajectory of self-position $x_{0:t}$ and the set of latent variables $\mathbf{C}_{1:t} = \{i_{1:t}, C_{1:t}, S_{1:t}\}$. We describe the joint posterior distribution of SpCoSLAM as follows:

$$\begin{aligned}
&p(x_{0:t}, \mathbf{C}_{1:t}, LM, \Theta, m \mid u_{1:t}, z_{1:t}, y_{1:t}, f_{1:t}, AM, \mathbf{h}) \\
&= p(LM \mid S_{1:t}, \lambda) p(m \mid x_{0:t}, z_{1:t}) \\
&\quad \cdot p(\Theta \mid x_{0:t}, \mathbf{C}_{1:t}, f_{1:t}, \mathbf{h}) \\
&\quad \cdot \underbrace{p(x_{0:t}, \mathbf{C}_{1:t} \mid u_{1:t}, z_{1:t}, y_{1:t}, f_{1:t}, AM, \mathbf{h})}_{\text{Particle filter}} \qquad (21)
\end{aligned}$$

where the set of hyperparameters is denoted as $\mathbf{h} = \{\alpha, \beta, \gamma, \chi, \lambda, m_0, \kappa_0, V_0, \nu_0\}$. Note that the speech signal y_t is not observed at all times. In this paper, the proposed method is equivalent to FastSLAM at the time when y_t is not observed.

The particle filter algorithm uses sampling importance resampling (SIR). We describe the importance weight $\omega_t^{[r]}$ for each particle as follows:

$$\begin{aligned}
\omega_t^{[r]} &= \frac{p(x_{0:t}^{[r]}, \mathbf{C}_{1:t}^{[r]} \mid u_{1:t}, z_{1:t}, y_{1:t}, f_{1:t}, AM, \mathbf{h})}{q(x_{0:t}^{[r]}, \mathbf{C}_{1:t}^{[r]} \mid u_{1:t}, z_{1:t}, y_{1:t}, f_{1:t}, AM, \mathbf{h})} \\
&= \frac{P_t^{[r]}}{Q_t^{[r]}} \qquad (22)
\end{aligned}$$

where the particle index is r. The number of particles is R. Henceforth, equations are also calculated for each particle r, but the subscripts representing the particle index are omitted.

We describe the target distribution P_t as follows:

$$\begin{aligned}
&p(x_{0:t}, \mathbf{C}_{1:t} \mid u_{1:t}, z_{1:t}, y_{1:t}, f_{1:t}, AM, \mathbf{h}) \\
&\approx p(z_t \mid x_t, m_{t-1}) p(f_t \mid C_{1:t-1}, f_{1:t-1}, \mathbf{h}) \\
&\quad \cdot p(i_t, C_t \mid x_{0:t}, i_{1:t-1}, C_{1:t-1}, S_{1:t}, f_{1:t}, \mathbf{h}) \\
&\quad \cdot p(x_t \mid x_{t-1}, u_t) p(S_t \mid S_{1:t-1}, y_{1:t}, AM, \lambda) \\
&\quad \cdot \frac{p(S_t \mid S_{1:t-1}, C_{1:t-1}, \alpha, \beta)}{p(S_t \mid S_{1:t-1}, \beta)} \cdot P_{t-1}. \qquad (23)
\end{aligned}$$

We describe the proposal distribution Q_t as follows:

$$\begin{aligned}
&q(x_{0:t}, \mathbf{C}_{1:t} \mid u_{1:t}, z_{1:t}, y_{1:t}, f_{1:t}, AM, \mathbf{h}) \\
&= \underbrace{q(x_t, \mathbf{C}_t \mid x_{0:t-1}, \mathbf{C}_{1:t-1}, u_{1:t}, z_{1:t}, y_{1:t}, f_{1:t}, AM, \mathbf{h})}_{q_t} \\
&\cdot \underbrace{q(x_{0:t-1}, \mathbf{C}_{1:t-1} | u_{1:t-1}, z_{1:t-1}, y_{1:t-1}, f_{1:t-1}, AM, \mathbf{h})}_{Q_{t-1}} \\
&= q_t Q_{t-1}. \qquad (24)
\end{aligned}$$

The weight ω_t is represented by (22), (23), and (24) as follows:

$$\begin{aligned}
\omega_t \approx\ & p(z_t \mid x_t, m_{t-1}) p(f_t \mid C_{1:t-1}, f_{1:t-1}, \mathbf{h}) \\
& \cdot p(i_t, C_t \mid x_{0:t}, i_{1:t-1}, C_{1:t-1}, S_{1:t}, f_{1:t}, \mathbf{h}) \\
& \cdot p(x_t \mid x_{t-1}, u_t) p(S_t \mid S_{1:t-1}, y_{1:t}, AM, \lambda) \\
& \cdot \frac{p(S_t \mid S_{1:t-1}, C_{1:t-1}, \alpha, \beta)}{p(S_t \mid S_{1:t-1}, \beta) q_t} \cdot \underbrace{\frac{P_{t-1}}{Q_{t-1}}}_{\omega_{t-1}}. \qquad (25)
\end{aligned}$$

We assume the proposal distribution q_t at time t as follows:

$$\begin{aligned}
q_t = \ & p(x_t \mid x_{t-1}, z_t, m_{t-1}, u_t) \\
& \cdot p(i_t, C_t \mid x_{0:t}, i_{1:t-1}, C_{1:t-1}, S_{1:t}, f_{1:t}, \mathbf{h}) \\
& \cdot p(S_t \mid S_{1:t-1}, y_{1:t}, AM, \lambda). \qquad (26)
\end{aligned}$$

Then, $p(x_t \mid x_{t-1}, z_t, m_{t-1}, u_t)$ is equivalent to the proposal distribution of FastSLAM 2.0.

The term of i_t and C_t is the marginal distribution regarding the set of model parameters Θ. This distribution can be calculated by a formula equivalent to collapsed Gibbs sampling. We describe the equation for sampling i_t and C_t simultaneously as follows:

$$\begin{aligned}
&p(i_t, C_t \mid x_{0:t}, i_{1:t-1}, C_{1:t-1}, S_{1:t}, f_{1:t}, \mathbf{h}) \\
&\propto p(S_{1:t} \mid C_{1:t}, \beta) p(f_{1:t} \mid C_{1:t}, \chi) p(x_{0:t} \mid i_{1:t}, \mathbf{h}) \\
&\quad \cdot p(i_t, C_t \mid i_{1:t-1}, C_{1:t-1}, \alpha, \gamma). \qquad (27)
\end{aligned}$$

The details of (27) are described in Section II-E.

We approximate the term of S_t by speech recognition using the language model LM_{t-1} and unsupervised word segmentation using the WFST speech recognition results $\mathcal{L}_{1:t}$ as follows:

$$\begin{aligned}
&p(S_t \mid S_{1:t-1}, y_{1:t}, AM, \lambda) \\
&\approx latticelm(S_{1:t} \mid \mathcal{L}_{1:t}, \lambda) \mathrm{SR}(\mathcal{L}_{1:t} \mid y_{1:t}, AM, LM_{t-1}). \qquad (28)
\end{aligned}$$

In the formulation of (21), it is desirable to estimate the language model LM_t for each particle. However, in this case, it is necessary to perform speech recognition of the number of data times the number of particles for each teaching utterance. In order to reduce the computational cost, we use a language model LM_t of a particle with the maximum weight for speech recognition of the next step.

Finally, ω_t is represented as follows:

$$\begin{aligned}
\omega_t \approx\ & p(z_t \mid m_{t-1}, x_{t-1}, u_t) p(f_t \mid C_{1:t-1}, f_{1:t-1}, \mathbf{h}) \\
& \cdot \frac{p(S_t \mid S_{1:t-1}, C_{1:t-1}, \alpha, \beta)}{p(S_t \mid S_{1:t-1}, \beta)} \cdot \omega_{t-1}. \qquad (29)
\end{aligned}$$

This is an equation obtained by multiplying the weight ω_{t-1} at a previous time with the marginal likelihoods for z_t, f_t, and S_t.

E. Simultaneous sampling of indices i_t and C_t

The proposed method uses the Chinese restaurant process (CPR) [18], which is one of the constitution methods of the Dirichlet process (DP). We describe the distribution of C_t using the CRP representation as follows:

$$p(C_t = l \mid C_{1:t-1}, \alpha) = \begin{cases} \frac{n_t^{(l)}}{n_t + \alpha} & (n_t^{(l)} > 0) \\ \frac{\alpha}{n_t + \alpha} & (l \text{ is new}) \end{cases} \qquad (30)$$

where $n_t^{(l)}$ denotes the number of data allocated to the l-th spatial concept in all data up to the time $t-1$. The number of data is $n_t = \sum_{l'} n_t^{(l')}$.

We describe the distribution of i_t by the CRP representation as follows:

$$\begin{aligned}
&p(i_t = k \mid i_{1:t-1}, C_{1:t-1}, C_t = l, \gamma) \\
&= \begin{cases} \frac{n_t^{(l,k)}}{n_t^{(l)} + \gamma} & (n_t^{(l,k)} > 0) \\ \frac{\gamma}{n_t^{(l)} + \gamma} & (k \text{ is new}) \end{cases} \qquad (31)
\end{aligned}$$

where $n_t^{(l,k)}$ denotes the number of data allocated to the k-th position distribution in data allocated to the l-th spatial concept.

Therefore, the joint prior distribution of i_t and C_t is represented as follows:

$$
\begin{aligned}
&p(i_t = k, C_t = l \mid i_{1:t-1}, C_{1:t-1}, \alpha, \gamma) \\
&= \begin{cases}
\frac{n_t^{(l,k)}}{n_t^{(l)}+\gamma} \frac{n_t^{(l)}}{n_t+\alpha} & (n_t^{(l,k)} > 0) \\
\frac{\gamma}{n_t^{(l)}+\gamma} \frac{n_t^{(l)}}{n_t+\alpha} & (n_t^{(l)} > 0 \ \cap \ k \text{ is new}) \\
\frac{\gamma}{n_t^{(l)}+\gamma} \frac{\alpha}{n_t+\alpha} & (l \text{ and } k \text{ are new})
\end{cases}
\end{aligned}
\tag{32}
$$

The probability of words S_t is represented as follows:

$$
\begin{aligned}
&p(S_{1:t} \mid C_{1:t-1}, C_t = l, \beta) \\
&= \prod_{B_t} p(S_{t,b} = s_g, S_{1:t-1} \mid C_{1:t-1}, C_t = l, \beta) \\
&\propto \begin{cases}
\prod_{B_t} \frac{n_t^{(l,g)}+\beta}{\sum_{g'=1}^{G}(n_t^{(l,g')}+\beta)} & (n_t^{(l)} > 0) \\
\frac{1}{G^{B_t}} & (l \text{ is new})
\end{cases}
\end{aligned}
\tag{33}
$$

where G denotes the number of types of words, i.e., the number of dimensions of the multinomial distribution of the names of places and $n_t^{(l,g)}$ denotes the total number of words s_g of the g-th dimension allocated to the l-th multinomial distribution of the names of the places in words $S_{1:t-1}$.

The probability of image features f_t is represented as follows:

$$
\begin{aligned}
&p(f_{1:t} \mid C_{1:t-1}, C_t = l, \chi) \\
&= \prod_{E} p(f_{t,e}, f_{1:t-1} \mid C_{1:t-1}, C_t = l, \chi) \\
&\propto \begin{cases}
\prod_{E} \Big(\frac{n_t^{(l,e)}+\chi}{\sum_{e'=1}^{E}(n_t^{(l,e')}+\chi)} \Big)^{f_{t,e}} & (n_t^{(l)} > 0) \\
\frac{1}{E^{F_t}} & (l \text{ is new})
\end{cases}
\end{aligned}
\tag{34}
$$

where E denotes the number of dimensions of image features, $n_t^{(l,e)}$ denotes the total number of image features of the e-th dimension allocated to the l-th multinomial distribution of image features in image features $f_{1:t-1}$, and $F_t = \sum_E f_{t,e}$.

The probability of self-position x_t of the robot is described as follows:

$$
\begin{aligned}
&p(x_t, x_{0:t-1} \mid i_{1:t-1}, i_t = k, \mathbf{h}) \\
&\propto \mathrm{St}(x_t \mid m_k, \frac{V_q(\kappa_k + 1)}{\kappa_k(\nu_k - d + 1)}, \nu_k - d + 1)
\end{aligned}
\tag{35}
$$

where the function St() denotes the multivariate Student's t-distribution [19]. Then, the posterior parameters in (35) are represented as follows:

$$
\begin{aligned}
\bar{x}_k &= \frac{1}{n_t^{(k)}} \sum_{x_j \in \mathbf{x}_k} x_j & (36) \\
m_k &= \frac{n_t^{(k)} \bar{x}_k + \kappa_0 m_0}{n_t^{(k)} + \kappa_0} & (37) \\
\kappa_k &= n_t^{(k)} + \kappa_0 & (38) \\
\nu_k &= \nu_0 + n_t^{(k)} & (39) \\
V_q &= V_0 + \sum_{x_j \in \mathbf{x}_k} x_j x_j^{\mathrm{T}} + \kappa_0 m_0 m_0^{\mathrm{T}} - \kappa_k m_k m_k^{\mathrm{T}} & (40)
\end{aligned}
$$

where $n_t^{(k)}$ and $\mathbf{x}_k$ are the number of data and the set of position data, respectively, allocated to the position distribution of $i_t = k$ in data up to the time $t - 1$.

From the above, (27) can be expressed as follows:

$$
\begin{aligned}
&p(i_t = k, C_t = l \mid x_{0:t}, i_{1:t-1}, C_{1:t-1}, S_{1:t}, f_{1:t}, \mathbf{h}) \\
&\propto \begin{cases}
\prod_{B_t} \frac{n_t^{(l,g)}+\beta}{\sum_{g'=1}^{G}(n_t^{(l,g')}+\beta)} \prod_{E} \Big(\frac{n_t^{(l,e)}+\chi}{\sum_{e'=1}^{E}(n_t^{(l,e')}+\chi)} \Big)^{f_{t,e}} \\
\cdot \mathrm{St}(x_t \mid m_k, \frac{V_q^{-1}(\kappa_k+1)}{\kappa_k(\nu_k-d+1)}, \nu_k - d + 1) \\
\cdot \frac{n_t^{(l,k)}}{n_t^{(l)}+\gamma} \frac{n_t^{(l)}}{n_t+\alpha} \\
\qquad\qquad (n_t^{(l,k)} > 0) \\
\prod_{B_t} \frac{n_t^{(l,g)}+\beta}{\sum_{g'=1}^{G}(n_t^{(l,g')}+\beta)} \prod_{E} \Big(\frac{n_t^{(l,e)}+\chi}{\sum_{e'=1}^{E}(n_t^{(l,e')}+\chi)} \Big)^{f_{t,e}} \\
\cdot \mathrm{St}(x_t \mid m_0, \frac{V_0^{-1}(\kappa_0+1)}{\kappa_0(\nu_0-d+1)}, \nu_0 - d + 1) \\
\cdot \frac{\gamma}{n_t^{(l)}+\gamma} \frac{n_t^{(l)}}{n_t+\alpha} \\
\qquad\qquad (n_t^{(l)} > 0 \ \cap \ k \text{ is new}) \\
\frac{1}{G^{B_t}} \frac{1}{E^{F_t}} \\
\cdot \mathrm{St}(x_t \mid m_0, \frac{V_0^{-1}(\kappa_0+1)}{\kappa_0(\nu_0-d+1)}, \nu_0 - d + 1) \\
\cdot \frac{\gamma}{n_t^{(l)}+\gamma} \frac{\alpha}{n_t+\alpha} \\
\qquad\qquad (l \text{ and } k \text{ are new})
\end{cases}
\end{aligned}
\tag{41}
$$

III. Experiments

We performed experiments for online learning of spatial concepts from a novel environment. In addition, we performed evaluations of place categorization and lexical acquisition related to place. We compare the performance of four methods as follows:

(A) SpCoSLAM
(B) Online SpCoA based on RBPF
(C) Online SpCoA
(D) SpCoA (Batch learning) [5]

Methods (A), (B), and (C) performed online learning algorithms based on the CRP representation. Methods (B), (C), and (D) based on SpCoA did not perform the update of a language model and did not use image features. Method (D) performed Gibbs sampling based on a weak-limit approximation [20] of the stick-breaking process (SBP) [21], i.e., the upper limit numbers of spatial concepts and position distributions were set as $L = 100$ and $K = 100$ respectively. In the batch learning (D), we performed Gibbs sampling for 100 iterations.

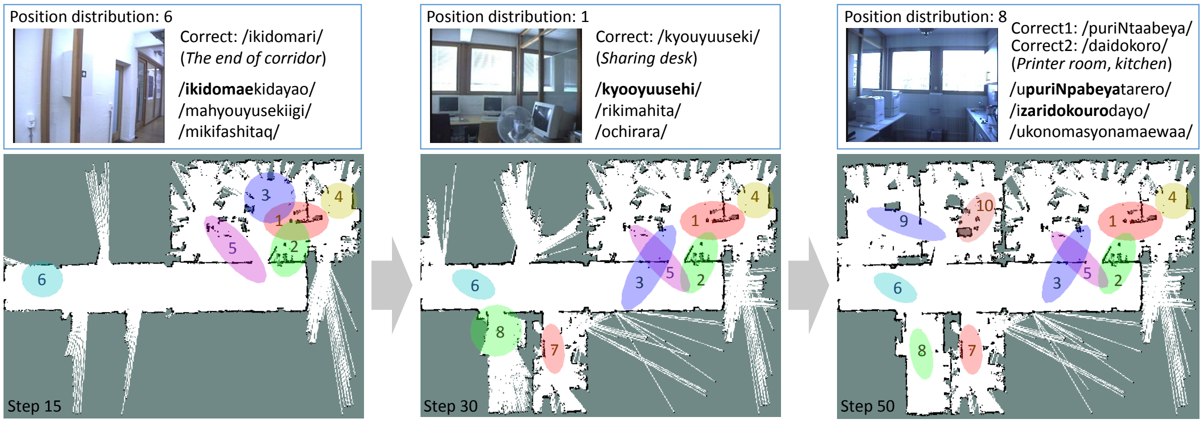


Fig. 3. Learning results of each position distribution in a generated map; Ellipses denoting the position distributions are drawn on the map at steps 15, 30, and 50. The colors of the ellipses were determined randomly. Furthermore, each index number is denoted as $i_t = k$.

A. Online learning

We conducted experiments for online spatial concept acquisition in a real environment. We extended the gmapping package, implementing the grid-based FastSLAM 2.0 [9], [10] in the robot operating system (ROS). We used an open dataset (albert-b-laser-vision) containing a rosbag file in which the odometry, laser range data, and vision data were recorded. This dataset was obtained from the Robotics Data Set Repository (Radish) [22]. The authors thank Cyrill Stachniss for providing this data. We prepared a Japanese speech signal data corresponding to the movement of the robot of the above dataset because it did not include speech signal data. The number of teaching places was 10 and there were nine place names. The teaching utterances included 10 types of various phrases. The total number of utterances was 50. The employed microphone was a SHURE PG27-USB. The speech recognition system uses Julius dictation-kit-v4.3.1-linux (GMM-HMM decoding) [23]. The initial word dictionary of the Julius system contains 115 Japanese syllables. The unsupervised word segmentation system uses latticelm [6]. We used a deep learning framework Caffe [24] for CNNs as an image feature extractor. We used a pre-trained CNN, i.e., Places205-AlexNet trained on 205 scene categories of Places Database with 2.5×10^6 images [25]. The map resolution was 0.05 m/grid. The number of particles was $R = 30$. The hyperparameters were set as follows: $\alpha = 20$, $\gamma = 10$, $\beta = 0.2$, $\chi = 0.2$, $m_0 = [0, 0]^{\mathrm{T}}$, $\kappa_0 = 0.001$, $V_0 = \mathrm{diag}(2, 2)$, and $\nu_0 = 3$. The above parameters were set so that all methods in the comparison were tested under the same conditions.

Fig. 3 shows the position distributions in the environmental maps at steps 15, 30, and 50. The upper part of this figure shows an example of the image corresponding to each position distribution, the correct phoneme sequence of the name of the place, and the upper three words of the probability value estimated by the probability distribution $p(S_t \mid i_t, \Theta_t, LM_t)$ at step t. As a result, Fig. 3 shows how the spatial concepts are acquired while sequentially mapping. Details on online learning experiment can be seen in the video attachment.

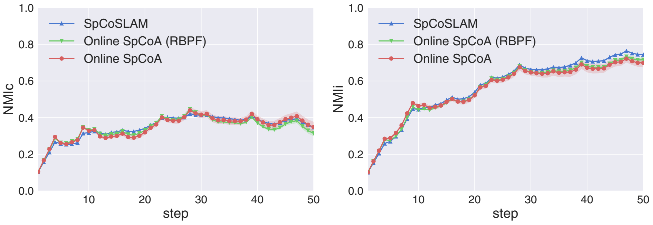
B. Estimation accuracy of spatial concepts

We compare the matching rate for the estimated index C_t of the spatial concept of each teaching utterance and the classification results of correct answers by a person. In this experiment, the evaluation metric uses the normalized mutual information (NMI), which is a measure of the degree of similarity between two clustering results. The estimated index i_t of the position distributions is also evaluated in the same manner. In addition, we evaluate the estimated number of spatial concepts L and position distributions K by using the estimation accuracy rate (EAR). The EAR was calculated as follows:

$$\mathrm{EAR} = \min(1 - \frac{\mid n_{\mathrm{T}} - n_{\mathrm{E}} \mid}{n_{\mathrm{T}}}, 0) \tag{42}$$

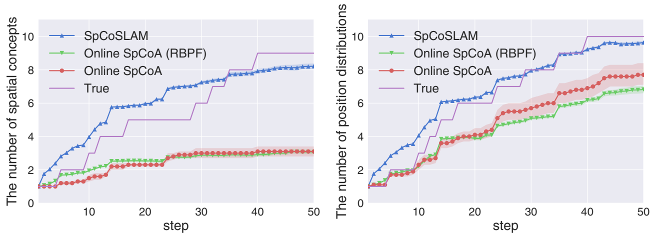
where n_{T} is the correct number and n_{E} is the estimated number.

Table II lists the evaluation-value averages calculated using the metrics NMI and EAR at step 50. Fig. 4 shows the average of the NMI values in 10 trials by online learning. In both C_t and i_t, the NMI values tended to rise at the beginning. The NMI values of C_t were similar for methods (A), (B), and (C). In the NMI values for i_t, the proposed method (A) showed higher values than the other methods after step 30. We consider a major possible reason for the clustering results of spatial concepts. In online lexical acquisition, the word segmentation results cannot be obtained stably when training dataset is small. We consider that stable words can be obtained by further increasing the number of training steps. Fig. 5 shows the average of the number of spatial concepts and the number of position distributions in 10 trials by online learning. The average values of the estimated results of method (D) were $L = 18.9$, $K = 13.1$. True data was determined by a user based on teaching data. The experimental results show that the proposed method (A) was closer to the true data than other methods for both L and K.



(a) NMI values of $C_{1:t}$ (b) NMI values of $i_{1:t}$

Fig. 4. Accuracy rates of the estimation results for (a) index of spatial concepts $C_{1:t}$ and (b) index of position distributions $i_{1:t}$



(a) Number of spatial concepts (b) Number of position distributions

Fig. 5. Estimation results for (a) the number of spatial concepts L and (b) the number of position distributions K

C. Comparison of the number of segmented words

We show whether a phoneme sequence including the name of a place is properly segmented. Fig. 6 shows the number of segmented words. The morphological segmentation (purple line) was suitably segmented into Japanese morphemes using MeCab, which is an off-the-shelf Japanese morphological analyzer that is widely used for natural language processing. The phrase segmentation (yellow line) was the number of words in the case of segmenting words only before and after the name of the place, i.e., we assume that a phrase other than the name of the place is one word. Table III presents examples of the word segmentation results of the four methods. Method (A) was similar to the phrase segmentation. On the other hand, methods (B) and (C) showed results of over-segmentation. In addition, the average value of the number of segmented words of method (D) was 391.4, i.e., it was similar to methods (B) and (C) at step 50. The results indicate that method (A) improved the problem of over-segmentation by updating the language model sequentially.

D. Place recognition using a speech signal

When the robot hears a user's speech signal y_t including the name of a place, the robot estimates a position $x_t^{(\mathrm{best})}$ indicated by the uttered sentence. The user says *"** ni iqte."* (which means *"Go to **."* in English). The estimation of a position was calculated as follows:

$$x_t^{(\mathrm{best})} = \operatorname*{argmax}_{x_t} p(x_t \mid y_t, \Theta, AM, LM). \tag{43}$$

TABLE II
EVALUATION VALUES OF NMI AND EAR FOR EACH METHOD

Methods	NMI		EAR	
	C_t	i_t	L	K
(A) SpCoSLAM	0.347	0.744	**0.913**	**0.964**
(B) Online SpCoA (RBPF)	0.314	0.716	0.341	0.682
(C) Online SpCoA	0.348	0.699	0.344	0.770
(D) SpCoA [5]	**0.805**	**0.856**	0.000	0.690

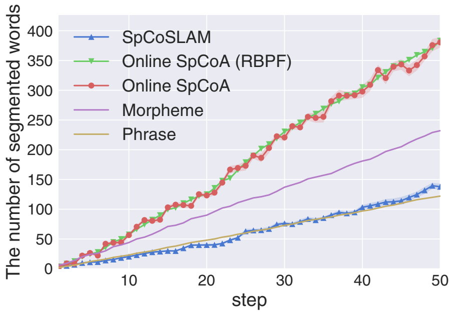


Fig. 6. Number of segmented words

In this experiment, (43) was approximated by using the speech recognition results $S_t^{(1:10\text{-best})}$ from 1-best to 10-best as follows:

$$\begin{aligned} S_t^{(1:10\text{-best})} &\sim \mathrm{SR}(S_t \mid y_t, AM, LM), & (44)\\ x_t^{(\mathrm{best})} &= \operatorname*{argmax}_{x_t} p(x_t \mid S_t^{(1:10\text{-best})}, \Theta). & (45) \end{aligned}$$

It is difficult to calculate (45) for all of the possible positions. Therefore, we use 10 position coordinates sampled for each position distribution as candidates for $x_t^{(\mathrm{best})}$. As a justification for this, we consider that positions near the mean values of position distributions become possible candidates for calculating (45). In this experiment, we decided to correct the position within the rectangular area surrounding the position coordinates taught as the same place (including 0.5 m margins to the right, left, above, and below). The place recognition rate (PRR) is calculated as follows:

$$\mathrm{PRR} = \frac{n_{\mathrm{C}}}{n_{\mathrm{U}}}, \tag{46}$$

where n_{U} denotes the number of utterances and n_{C} denotes the number of correct positions. The number of utterances is nine.

Fig. 7 shows the average of the PRR values in 10 trials. The average value of PRR of Method (D) was 0.500. Method (A) showed the highest overall evaluation values of the online methods. The experimental results show that the robot was able to more accurately learn the relationships between words and the position in the map incrementally by using method (A).

IV. CONCLUSION

This paper discussed online learning methods of spatial concepts and an environmental map by a mobile robot. The proposed method integrated the spatial concept acquisition into SLAM by an RBPF-based approach. In the experiments, we conducted online learning in a novel environment by the robot without a pre-existing lexicon and map. The

TABLE III

EXAMPLES OF WORD SEGMENTATION RESULTS OF UTTERED SENTENCES. "|" DENOTES A WORD SEGMENT POSITION

English	*"We come to* **the end of corridor**."	"**The faculty laboratory** *is here."*	*"This place name is* **the break room**."
Morpheme	**ikidomari**|ni|ki|mashi|ta	**kyouiNkeNkyuushitsu**|wa|kochira|desu	kono|basyo|no|namae|wa|**kyuukeijo**
Phrase	**ikidomari**|nikimashita	**kyouiNkeNkyuushitsu**|wakochiradesu	konobasyononamaewa|**kyuukeijo**
(A)	**aaerikidomari**|nikeiwasuta	**kyoiiNiNteNkyushitsu**|waqgochigadesu	ukonomasyonamaewaa|**kyuuqkirijo**
(B), (C), (D)	**pikido**|**ma**|**e**|ni|ki|ma|sya	**kyoo**|**i**|**N**|**teN**|**kyu**|**shi**|**su**|wa|ko|chi|ga|desu	kono|basyo|no|nama|e|wa|**kyuu**|**ke**|**i**|**jo**

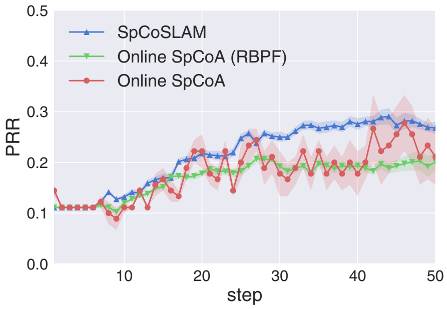


Fig. 7. Results of PRR values

experimental results demonstrated that SpCoSLAM enhances the performance of place recognition using a speech signal in online learning methods. SpCoSLAM improved over-segmentation problem in lexical acquisition by updating the language model sequentially. We consider that incorporating forgetting [14] and rejuvenation [26] into SpCoSLAM could further improve estimation accuracy.

One of the advantages of online learning is that it can deal with changes in the environment and place names. Moreover, we consider that the spatial concepts the robot mistakenly learns can also be corrected sequentially. In this way, it will be possible to acquire spatial concepts that flexibly respond to changes in the environment, which could not be done so far. We expect this work to contribute greatly to the realization of long-term spatial language interaction between people and robots. In the future, we would like to perform continuous online learning of spatial concepts in a long-term dynamic environment and incremental transfer learning of spatial concepts to other novel environments.

REFERENCES

[1] I. Kostavelis and A. Gasteratos, "Semantic mapping for mobile robotics tasks: A survey," *Robotics and Autonomous Systems*, vol. 66, pp. 86–103, 2015.

[2] C. Landsiedel, V. Rieser, M. Walter, and D. Wollherr, "A review of spatial reasoning and interaction for real-world robotics," *Advanced Robotics*, vol. 31, no. 5, pp. 222–242, 2017.

[3] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*. MIT Press, 2005.

[4] A. Taniguchi, T. Taniguchi, and T. Inamura, "Simultaneous estimation of self-position and word from noisy utterances and sensory information," in *IFAC-PapersOnLine*, vol. 49, no. 19. Elsevier, 2016, pp. 221–226.

[5] ——, "Spatial concept acquisition for a mobile robot that integrates self-localization and unsupervised word discovery from spoken sentences," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 8, no. 4, pp. 285–297, 2016.

[6] G. Neubig, M. Mimura, and T. Kawahara, "Bayesian learning of a language model from continuous speech," *IEICE Transactions on Information and Systems*, vol. 95, no. 2, pp. 614–625, 2012.

[7] S. Ishibushi, A. Taniguchi, T. Takano, Y. Hagiwara, and T. Taniguchi, "Statistical localization exploiting convolutional neural network for an autonomous vehicle," in *Proceedings of IECON*, 2015, pp. 1369–1375.

[8] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of NIPS*, 2012, pp. 1097–1105.

[9] G. Grisetti, C. Stachniss, and W. Burgard, "Improving grid-based SLAM with Rao-Blackwellized particle filters by adaptive proposals and selective resampling," in *Proceedings of ICRA*, 2005.

[10] ——, "Improved techniques for grid mapping with Rao-Blackwellized particle filters," *IEEE Transactions on Robotics*, vol. 23, pp. 34–46, 2007.

[11] A. Doucet, N. De Freitas, K. Murphy, and S. Russell, "Rao-Blackwellised particle filtering for dynamic Bayesian networks," in *Proceedings of the Conference on UAI*, 2000, pp. 176–183.

[12] M. Spranger, "Incremental grounded language learning in robot-robot interactions-examples from spatial language," in *Proceedings of ICDL-EpiRob*, 2015, pp. 196–201.

[13] S. Heath, D. Ball, and J. Wiles, "Lingodroids: Cross-situational learning for episodic elements," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 8, no. 1, pp. 3–14, 2016.

[14] T. Araki, T. Nakamura, T. Nagai, S. Nagasaka, T. Taniguchi, and N. Iwahashi, "Online learning of concepts and words using multimodal LDA and hierarchical Pitman-Yor Language Model," in *Proceedings of IROS*, 2012, pp. 1623–1630.

[15] D. Mochihashi, T. Yamada, and N. Ueda, "Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling," in *Proceedings of ACL-IJCNLP*, 2009, pp. 100–108.

[16] S. Goldwater, T. L. Griffiths, and M. Johnson, "A Bayesian framework for word segmentation: Exploring the effects of context," *Cognition*, vol. 112, no. 1, pp. 21–54, 2009.

[17] D. Gildea and T. Hofmann, "Topic-based language models using em," in *In Proceedings of EUROSPEECH*, 1999.

[18] D. Aldous, "Exchangeability and related topics," *École d'Été de Probabilités de Saint-Flour XIII-1983*, pp. 1–198, 1985.

[19] K. P. Murphy, *Machine learning: a probabilistic perspective*, 2012.

[20] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "A sticky HDP-HMM with application to speaker diarization," *The Annals of Applied Statistics*, pp. 1020–1056, 2011.

[21] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[22] A. Howard and N. Roy, "The robotics data set repository (radish)," 2003. [Online]. Available: http://radish.sourceforge.net/

[23] T. Kawahara, T. Kobayashi, K. Takeda, N. Minematsu, K. Itou, M. Yamamoto, A. Yamada, T. Utsuro, and K. Shikano, "Sharable software repository for Japanese large vocabulary continuous speech recognition," in *Proceedings of ICSLP*, 1998.

[24] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[25] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proceedings of NIPS*, 2014, pp. 487–495.

[26] K. R. Canini, L. Shi, and T. L. Griffiths, "Online inference of topics with latent Dirichlet allocation." in *Proceedings of AISTATS*, vol. 9, 2009, pp. 65–72.