

Simultaneous Estimation of Self-position and Word from Noisy Utterances and Sensory Information

Akira Taniguchi* Tadahiro Taniguchi* Tetsunari Inamura**

* *Ritsumeikan University, 1-1-1 Noji Higashi, Kusatsu, Shiga
 525-8577, Japan (e-mail: a.taniguchi@em.ci.ritsumei.ac.jp;
 taniguchi@em.ci.ritsumei.ac.jp).*

** *National Institute of Informatics/The Graduate University for
 Advanced Studies, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430,
 Japan (e-mail: inamura@nii.ac.jp).*

Abstract: In this paper, we propose a novel learning method that can simultaneously estimate the self-position of a robot and place names. The robot moves in a room environment and performs probabilistic self-localization based on noisy sensory information. Speech recognition results include the uncertainty of phonemes or syllables, because the robot does not have lexical knowledge in advance. The purpose of this study is to reduce the uncertainty of both self-position and speech recognition using knowledge about place names, which is obtained from human speech. The proposed method integrates ambiguous speech recognition results with the self-localization method, i.e., Monte Carlo localization, using a Bayesian approach. Probability distributions over places and the speech recognition error are modeled using the proposed method. We implemented the proposed method in SIGVerse, which is a simulation environment. Experimental results showed that the robot can acquire the names of several places and use this knowledge to reduce the uncertainty of estimation in its position in a self-localization task. In addition, we evaluated the performance of the lexical acquisition task for the names of places and showed its effectiveness. Results showed that the robot could acquire spatial concepts by integrating noisy information from sensors and speech.

© 2016, IFAC (International Federation of Automatic Control) Hosting by Elsevier Ltd. All rights reserved.

Keywords: ambiguous speech recognition, lexical acquisition, self-localization.

1. INTRODUCTION

It is important for mobile robots operating in the human living environment, e.g., homes and offices, to estimate their positions and to learn words related to places through human-robot interaction. The robot should be able to recognize a variety of objects, places, and situations in its immediate environment. The robot handles noisy inputs observed from various sensors, such as visual sensors, odometers, and auditory sensors [Thrun et al. (2005)].

The purpose of this study is to reduce the uncertainty of both the self-position and the speech recognition using knowledge about place names that is obtained from user speech. To achieve this purpose, we propose a method in which the robot can learn the names of places and spatial areas from noisy sensory information and speech signals. We assume that the robot can recognize user speech at the unit of phonemes or syllables, i.e., the robot does not have word knowledge in advance. This study requires self-localization and lexical acquisition.

There are several problems associated with self-localization and lexical acquisition. The central problem in global self-localization is that the self-localization results are uncertain, e.g., the self-localization results exist for multiple remote locations. The robot estimates self-position based on control data and local sensor data (containing noise)

using a probabilistic method. If the robot's position on the global map is uncertain, the identification of the self-position using only local sensor data becomes a very difficult problem. On the other hand, the central problem of lexical acquisition is that the speech recognition results are uncertain, e.g., /*oranji*/, /*ourenji*/, and /*orenjye*/ are all incorrect phoneme recognition results for *orange*. It is very difficult for the robot to determine the identity of the word from these different phoneme sequences (and other various phoneme sequences) using speech recognition without prior knowledge of the set of words, e.g., a word dictionary.

The main contributions of this paper are as follows:

- (i) We simultaneously obtained solutions to the above two problems by integrating self-localization using noisy sensory information with ambiguous speech recognition results. The proposed method is based on a Bayesian probability approach for simultaneous estimation of self-position and words from noisy utterances and sensory information.
- (ii) We developed a robot with learning skills for the discovery and grounding of words related to places. We show that the robot can learn the names of places and spatial areas, i.e., spatial concepts. In addition, we show that the robot can reduce the estimation error of self-position by utilizing word information.

2. BACKGROUND

Most studies of lexical acquisition focus on words related to objects [Roy and Pentland (2002); Iwahashi (2003); Hörnstein et al. (2010); Nakamura et al. (2011)]. Roy and Pentland (2002) proposed a computational model that enables a robot to learn the names of objects from object images and spontaneous infant-directed utterances. Iwahashi (2003) reported that a robot can understand situations and acquire the relationships between object behaviors and sentences. Hörnstein et al. (2010) proposed a method for language acquisition by humanoid robots. Nakamura et al. (2011) proposed a method that enables the acquisition of object concepts and word meanings from multimodal information. The methods in these studies do not require pre-programmed lexical knowledge; words and phonemes are acquired by pattern recognition, clustering methods, etc. However, these studies did not address the lexical acquisition of space and place names in a way that can also tolerate the uncertainty of speech recognition. For the introduction of robots into the human living environment, robots need to acquire a lexicon related not only to objects but also to places. Our study focuses on the lexical acquisition of place names.

Taguchi et al. (2011) proposed a method for the simultaneous categorization of self-position coordinates and lexical learning. These experimental results showed that, in some cases, it was possible to learn place names from utterances and to output words corresponding to places in a different location from that used for learning. However, their study is not able to use the learned words for self-localization of the robot. In this study, the proposed method can utilize words related to places for self-localization.

3. SIMULTANEOUS ESTIMATION OF SELF-POSITION AND WORDS

We propose a method that simultaneously estimates the self-position of a robot and place names. In this study, we define a *spatial concept* as a cluster of place names and a position distribution. The *position distribution* represents the spatial region in the environment. The self-localization method adopts MCL (Monte Carlo localization) [Thrun et al. (2005)], a method that is generally used for the self-localization of mobile robots. We assume that the robot can recognize phonemes or syllables.

3.1 Overview of the tasks

A schematic representation depicting the target task of this study is shown in Fig. 1. Fig. 1 (a) shows the locations of the learning targets. Fig. 1 (b) shows the situations of teaching of place names by the user. When the robot arrives at the designated location of the learning target, the user says the name of the current place. The speech recognition results include phoneme recognition errors because the robot does not have word knowledge in advance. After multiple teachings by the user, the robot learns the spatial concepts, as shown in Fig. 1 (c). After learning the spatial concepts, the robot performs self-localization while the robot moves around in the environment. Fig. 1 (d) shows a scene in which the robot is in front of a TV. The self-localization result exists for two

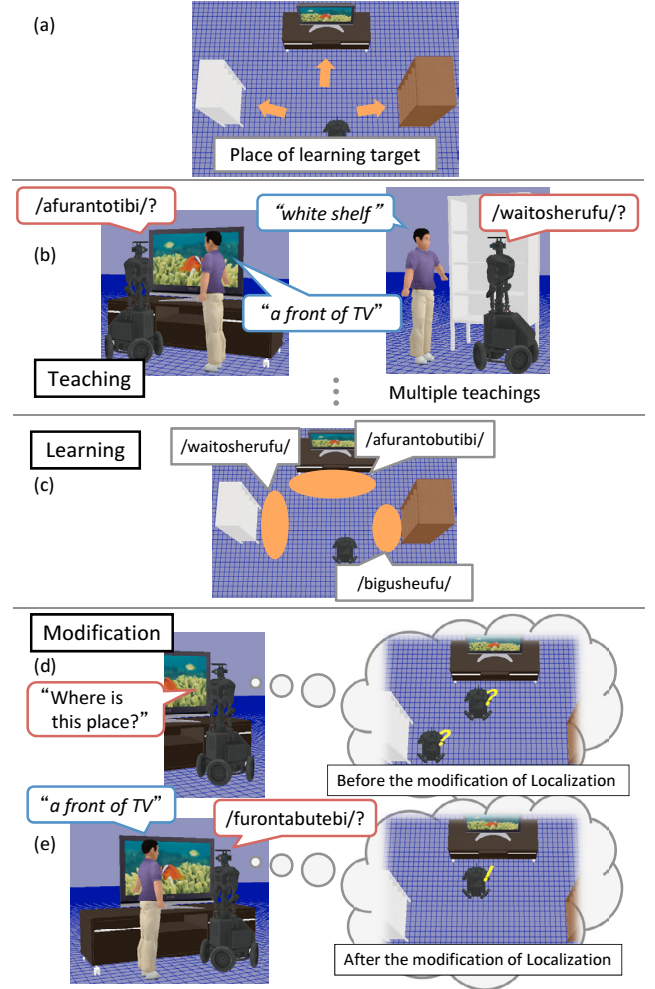


Fig. 1. Schematic diagram of our research

remote places. The user speaks the current place name, and the robot recognizes the user's speech, as shown in Fig. 1 (e). The robot can narrow down the self-localization result to being in front of TV by utilizing spatial concepts and the speech recognition result.

3.2 The definition of spatial concepts

The set of learned spatial concepts λ is shown in Eq. (1). The number of spatial concepts is denoted as L . We define the spatial concept of index i as shown in Eq. (2).

$$\lambda = \{\lambda_1, \lambda_2, \dots, \lambda_L\} \quad (1)$$

$$\lambda_i = \{W_i, \mu_i, \Sigma_i\} \quad (2)$$

Then, the learned place name of index i is denoted as W_i , and the position distribution of index i is denoted as μ_i, Σ_i . The set of place names \mathbf{W} is shown in Eq. (3). We define the position distribution as a multivariate Gaussian distribution. Therefore, the set of mean vectors is denoted as μ , and the set of covariance matrices is denoted as Σ .

$$\mathbf{W} = \{W_1, W_2, \dots, W_L\} \quad (3)$$

$$\mu = \{\mu_1, \mu_2, \dots, \mu_L\} \quad (4)$$

$$\Sigma = \{\Sigma_1, \Sigma_2, \dots, \Sigma_L\} \quad (5)$$

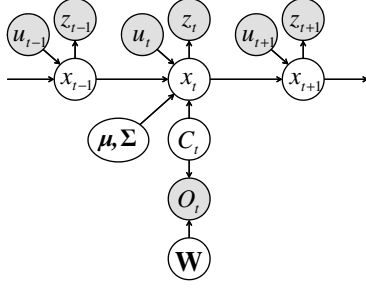


Fig. 2. Graphical model of the proposed method

Table 1. Each element of the graphical model

x_t	Robot position
u_t	Control data
z_t	Sensor data
C_t	State of spatial concept
O_t	Recognized word
\mathbf{W}	Set of place names
μ, Σ	Set of position distributions (mean vectors, covariance matrices)

A state of the spatial concept is denoted as $C_t \in \{1, 2, \dots, L\}$. The state C_t represents an index of the spatial concept, which the robot obtains by listening to speech information from a user.

3.3 Self-localization using spatial concepts

We describe the self-localization method using spatial concepts λ and a speech signal O_t from the user. Fig. 2 shows the graphical model of the proposed method, which integrates the state of spatial concepts C_t , position distributions (μ, Σ) , the place names \mathbf{W} , and the words in the speech recognition results O_t of MCL. Table 1 shows each variable in the graphical model.

The posterior probability that provides O_t to the conditional part of MCL is calculated as follows:

$$p(x_{0:t} | z_{1:t}, u_{1:t}, O_{1:t}, \lambda) \propto p(z_t | x_t) p(O_t | x_t, \lambda) p(x_t | x_{t-1}, u_t) p(x_{0:t-1} | z_{1:t-1}, u_{1:t-1}, O_{1:t-1}, \lambda) \quad (6)$$

where $p(z_t | x_t)$ is the measurement model of MCL and $p(x_t | x_{t-1}, u_t)$ is the motion model of MCL [Thrun et al. (2005)]. We describe the probability distribution of $p(O_t | x_t, \lambda)$ using Eq. (7). If the robot detects a word O_t in a position x_t , Eq. (7) calculates the likelihood.

$$p(O_t | x_t, \lambda) = \sum_{C_t} p(O_t | C_t, \lambda) p(C_t | x_t, \lambda) \propto \sum_{C_t} p(O_t | W_{C_t}) p(x_t | \mu_{C_t}, \Sigma_{C_t}) p(C_t) \quad (7)$$

We assume that the probability distribution representing the distance of the learned place name and the recognized sequence of phonemes or syllables is calculated as follows:

$$p(O_t | W_{C_t}) \propto \exp(-\beta \text{LD}(O_t, W_{C_t})) \quad (8)$$

where $\text{LD}(\cdot)$ denotes the function calculating the Levenshtein distance. The parameter representing the extent of the effect on the Levenshtein distance is denoted as β .

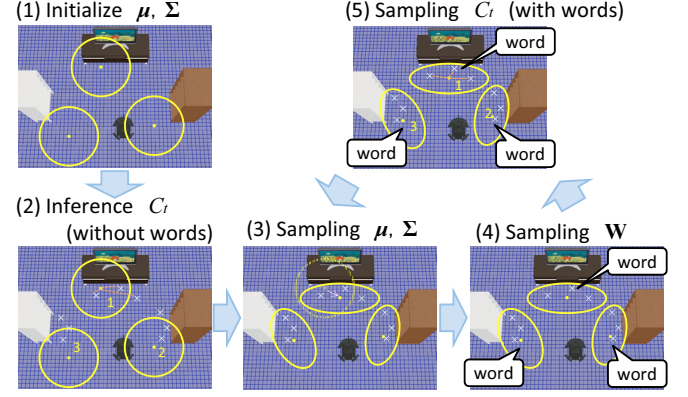


Fig. 3. Flow diagram of learning of spatial concepts using Gibbs sampling

The probability distribution of the robot position in the learned position distribution is calculated as follows:

$$p(x_t | \mu_{C_t}, \Sigma_{C_t}) = \frac{1}{(2\pi)^{D/2} |\Sigma_{C_t}|^{1/2}} \exp\left(-\frac{1}{2} (x_t - \mu_{C_t})^T \Sigma_{C_t}^{-1} (x_t - \mu_{C_t})\right) \quad (9)$$

where D denotes the number of dimensions.

In addition, we define the prior distribution of C_t as the uniform distribution $p(C_t) = 1/L$.

3.4 Learning spatial concepts

We propose a learning algorithm for spatial concepts that uses Gibbs sampling. Fig. 3 shows the flow diagram of the process of learning spatial concepts using Gibbs sampling. The algorithm for learning place names and position distributions is shown in Algorithm 1. The set of teaching times is denoted as $T_o = \{t_1, t_2, \dots, t_N\}$. The number of data points is N . The user sets the number of spatial concepts L in advance.

We describe the Gibbs sampling procedure as follows.

(1) Initialization of μ, Σ

We set the initial position distribution. The mean vector μ_c is the uniform random number in the given range, i.e., the range in which the robot can move on the map. The covariance matrix Σ_c is set as $\text{diag}(\sigma_{\text{initial}}, \sigma_{\text{initial}})$.

(2) Sampling of C_t without \mathbf{W}

The conditional posterior distribution of C_t samples the state of spatial concept C_t at each time $t \in T_o$. In this step, the robot does not estimate the place names \mathbf{W} . Therefore, the sampling of C_t is performed as follows:

$$C_t \sim p(C_t | x_t, \mu, \Sigma) \propto p(x_t | \mu_{C_t}, \Sigma_{C_t}) p(C_t) \quad (10)$$

(3) Sampling of μ, Σ

The conditional posterior distribution of μ_c, Σ_c samples the position distribution for each state of the spatial concept $c \in \{1, 2, \dots, L\}$. The sampling of μ_c, Σ_c is performed as follows:

Algorithm 1 Learning of spatial concepts

```

procedure Gibbs_sampling( $x_{T_o}, O_{T_o}, L, N$ )
   $N_{iteration}$  The number of the iteration
  // (1) Initialization of  $\mu, \Sigma$ 
  Initialize parameters  $\mu, \Sigma$ 
  // (2) Sampling of  $C_t$  without  $\mathbf{W}$ 
  for  $t = 1$  to  $N$  do
     $C_t \sim p(x_t | \mu_{C_t}, \Sigma_{C_t})$ 
  end for
  // (6) Multiple iterations of procedures (3)–(5)
  for  $i = 1$  to  $N_{iteration}$  do
    // (3) Sampling of  $\mu, \Sigma$ 
    for  $c = 1$  to  $L$  do
       $\mu_c, \Sigma_c \sim \left[ \prod_{\substack{c = C_t \\ t \in T_o}} p(x_t | \mu_{C_t}, \Sigma_{C_t}) \right] p(\mu_c, \Sigma_c)$ 
    end for
    // (4) Sampling of  $\mathbf{W}$ 
    for  $c = 1$  to  $L$  do
       $W_c \sim \left[ \prod_{\substack{c = C_t \\ t \in T_o}} p(O_t | W_{C_t}) \right] p(W_c)$ 
    end for
    // (5) Sampling of  $C_t$  with  $\mathbf{W}$ 
    for  $t = 1$  to  $N$  do
       $C_t \sim p(x_t | \mu_{C_t}, \Sigma_{C_t}) p(O_t | W_{C_t}) p(C_t)$ 
    end for
  end for
  return  $\mathbf{W}, \mu, \Sigma$ 
end procedure

```

$$\mu_c, \Sigma_c \sim \left[\prod_{\substack{c = C_t \\ t \in T_o}} p(x_t | \mu_{C_t}, \Sigma_{C_t}) \right] p(\mu_c, \Sigma_c) \quad (11)$$

We define the prior distribution $p(\mu_c, \Sigma_c)$ as a Gaussian-Wishart distribution $\mathcal{NW}(\mu_c, \Sigma_c | m_0, \beta_0, V_0, \nu_0)$. The hyperparameters of the prior distribution are denoted as m_0, κ_0, V_0 , and ν_0 . Then, the posterior distribution of μ_c, Σ_c becomes the Gaussian-Wishart distribution. The Gaussian-Wishart posterior distribution is performed as follows:

$$(11) = \mathcal{NW}(\mu_c, \Sigma_c | m_{N_c}, \kappa_{N_c}, V_{N_c}, \nu_{N_c}) \quad (12)$$

where the hyperparameters of the posterior distribution are denoted as $m_{N_c}, \kappa_{N_c}, V_{N_c}$, and ν_{N_c} .

If an index c of the state of the spatial concept does not exist in the estimated C_{T_o} from teaching data, we set a uniform random number in the given range as the parameter m_{N_c} .

(4) Sampling of \mathbf{W}

The conditional posterior distribution of W_c samples the place name W_c for each state of the spatial concept $c \in \{1, 2, \dots, L\}$. The sampling of W_c is performed as follows:

$$W_c \sim \left[\prod_{\substack{c = C_t \\ t \in T_o}} p(O_t | W_{C_t}) \right] p(W_c) \quad (13)$$

where the prior distribution $p(W_c)$ denotes a uniform distribution, i.e., $p(W_c) = 1/N$. We choose W_c from the recognized words in the teaching data. Therefore, sampling of W_c performs selection based on the posterior probability of each recognized word $O_{t'}$ of time $t' \in T_o = \{t_1, t_2, \dots, t_N\}$.

(5) Sampling of C_t with \mathbf{W}

The conditional posterior distribution of C_t samples the state of spatial concept C_t for each time $t \in T_o$. The sampling of C_t is shown as follows:

$$C_t \sim p(C_t | x_t, O_t, \mathbf{W}, \mu, \Sigma) \propto p(x_t | \mu_{C_t}, \Sigma_{C_t}) p(O_t | W_{C_t}) p(C_t) \quad (14)$$

(6) Multiple iterations of procedure (3)–(5)

Multiple iterations were performed for the process described in steps (3) to (5) above.

4. EXPERIMENTS

We perform experiments on learning related to spatial concepts. In addition, we evaluate the accuracy of self-localization and word acquisition. In these experiments, we use SIGVerse¹ as the simulator of the robot and the environment [Inamura et al. (2010)], and Julius² as the Japanese speech recognition system [Lee et al. (2001)]. To recognize speech signals as syllable sequences, the dictionary of Julius uses Japanese syllables only. We use a SHURE PG27-USB as the microphone.

4.1 Learning spatial concept

Condition We performed an experiment on learning related to spatial concepts. The experimental environment is shown in Fig. 4. The four red triangles in Fig. 4 represent landmarks for use in self-localization. We set the number of spatial concepts to $L = 4$. The other model parameters are as follows: $\beta = 1$, $\kappa_0 = 0.001$, $m_0 = [0, 0]^T$, $V_0 = \text{diag}(1, 1)$, $\nu_0 = 1$, and $\sigma_{initial} = 10000$. The number of iterations of step (6) is 10. Four places were selected as learning targets. The user decides the teaching positions in a learning target. The teaching utterances are repeated 40 times. There were four types of place names. The recognition results of place names are shown in Table 2. The taught place names are /shiroitana/, /tsukue/, /gomibako/, and /terebimae/. In the learning phase, the self-position x_t is assumed to be an accurate estimation; it was obtained using Monte Carlo smoothing³ [Kitagawa (2014)]. In this experiment, we use the true coordinates of a robot to teaching data as an approximation.

Results Fig. 4 shows the learning results for spatial concepts. Each ellipse shows the learned position distribution. Each balloon shows the state number and the place name for each position distribution. As a result, the proposed method can learn the place names corresponding to each place that is a learning target.

4.2 Self-localization using learned spatial concepts

Condition We performed an experiment for the evaluation of self-localization using learned spatial concepts. We compared the estimation error of self-localization between MCL using the spatial concepts, i.e., the proposed method, and conventional MCL. We implemented the landmark-based MCL with a mobile robot in SIGVerse. The robot

¹ SIGVerse, <http://www.sigverse.org/wiki/en/>

² Julius dictation-kit-v4.2, <http://julius.sourceforge.jp/>

³ In general, the smoothing method can provide a more accurate estimation than the MCL of online estimation.

Table 2. Syllable recognition results using the Japanese syllable dictionary

Teaching word	Recognized words				
/shiroitana/ (white shelf)	shiroitana chiitana	shinitana tsutana	shiroitaga shinonitana	shinotaga shinotana	sinutana siruitana
/tsukue/ (desk)	tsukube tsukune	tsukune tsukuke	tsukuu tsutsune	tsukue tsukune	tsukume
/gomibako/ (trash bin)	komiwako komiwako	gubiwako gumiwako	gumiwako gumiwako	gubiwako gumibako	kumiwako gumibaku
/terebimae/ (in front of the TV)	terimae terenae	tebimae tezurimae	tegikunae terimae	terenae teteginae	terimae teribinae

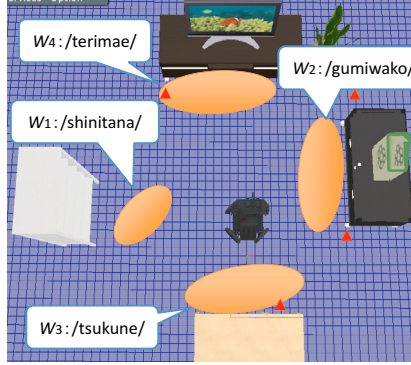


Fig. 4. Learning result of spatial concepts in SIGVerse

can recognize landmarks within the view angle of the robot's camera. The robot observes a distance and an angle from the recognized landmark as sensor information. The view angle of the camera is 45 degrees centered at the front face of the robot. The number of particles is set to $M = 1000$. The initial particles are sampled from the uniform distribution in the given range. The robot performs the actions, one at a time, using the control data for each step. When the robot arrives at the learning target, the user says the place name of the target place for the robot. The taught words are randomly selected from recognized words of table 2 in this experiment.

We present the evaluation of the estimation error in MCL as follows. While the robot performs self-localization, the estimation error on the xy plane of the floor for each time step is calculated as follows:

$$e_t = \sqrt{(\bar{x}_t - x_t^*)^2 + (\bar{y}_t - y_t^*)^2} \quad (15)$$

where $\bar{x}_t = \sum_{i=1}^M w_t^{(i)} x_t^{(i)}$ and $\bar{y}_t = \sum_{i=1}^M w_t^{(i)} y_t^{(i)}$ denote the weighted mean values of particle coordinates, and x_t^* and y_t^* denote the robot's true position. The normalized weight of MCL is denoted as $w_t^{(i)}$. The x and y coordinates of the particle index i are denoted as $x_t^{(i)}, y_t^{(i)}$. After performing self-localization, we calculate the average of e_t as e_m . We denote the minimum value of the estimation error γ as e_r . The value of γ is determined by the interval $[0, \gamma]$, which includes values greater than or equal to 95% of e_t in all teaching times. We regard e_r as the evaluation indicator for the stability of self-localization.

Result Fig. 5 shows the comparison result using e_m , e_r of the average of ten trials. The estimation errors of the proposed method were smaller than that of MCL for two evaluation indicators. We used t -tests of level of significance of 5% for both e_m and e_r . The results of t -tests showed significant differences in the two evaluation

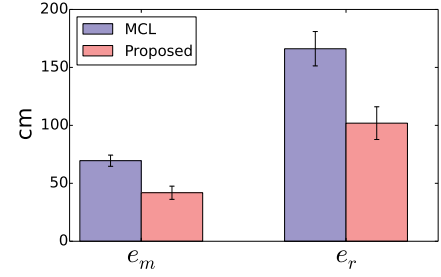


Fig. 5. Comparison of the estimated error due to MCL and the proposed method

indicators. As a result, we showed that the spatial concepts enable the modification of the global self-localization error. We believe that the robot can narrow particles to the accurate place using the spatial concept estimated from the user utterance, even when particles exist for multiple discrete places in the self-localization.

4.3 Evaluation for word clustering

Condition We compared the matching rate for the estimated states of the spatial concept C_t of teaching utterances and the classification results of ground truth by a human. Fig. 6 (a) shows the classification results of the ground truth. The axis of ordinate of Fig. 6 is the teaching data number. The axis of abscissas of Fig. 6 is the recognized state of the spatial concept. Then, the pairs of indices of C_t and teaching words are 1:/shiroitana/, 2:/tsukue/, 3:/gomibako/, and 4:/terebimae/. The white boxes represent the classification of the teaching data according to the state of the spatial concept.

The EAR (estimation accuracy rate) of spatial concepts is calculated as follows:

$$\text{EAR} = \frac{\text{The number of estimated correct data}}{\text{The number of all teaching data}} \quad (16)$$

We compare the estimation accuracy rate of spatial concepts between the proposed method, i.e., clustering using position data x_t and word data O_t , and word clustering using only O_t . In other words, this experiment considers the comparative evaluation of the accuracy of the lexical acquisition based on word information and the lexical acquisition using word information when adding position information. We describe a clustering method using word data only O_t as follows.

(1) Initialization of μ, Σ

This step is the same as step (1) in the proposed method.

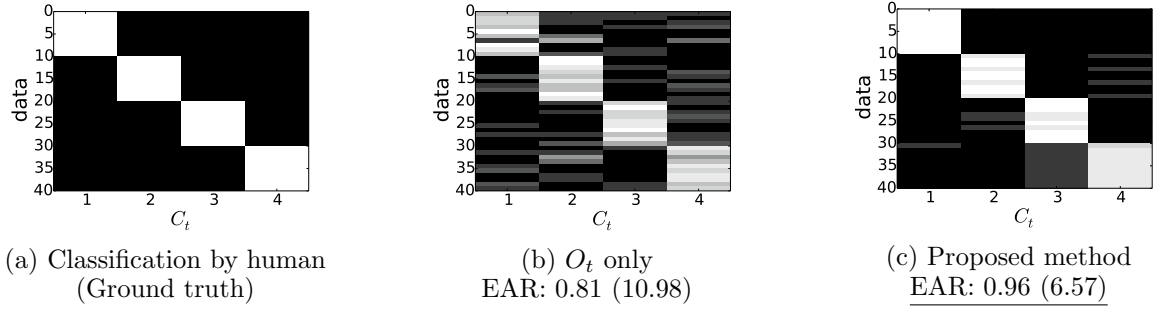


Fig. 6. Correct classification by human and estimation results of C_t via two methods; the average value and standard deviation of EAR (estimation accuracy rate) of ten trials

- (2) Sampling of C_t without \mathbf{W}
The sampling of C_t is performed as follows:

$$C_t \sim p(C_t) \quad (17)$$

- (3) Sampling of \mathbf{W}
This step is the same as step (4) in the proposed method.
- (4) Sampling of C_t with \mathbf{W}
The sampling of C_t is performed as follows:

$$C_t \sim p(O_t | W_{C_t})p(C_t) \quad (18)$$

- (5) Multiple iterations of procedure (3) and (4).

Result We perform an experiment for the learning of spatial concepts in ten trials for each method. Fig. 6 (b) shows the estimation result and the EAR for the clustering of only word data O_t . Fig. 6 (c) shows the estimation result and the EAR for the proposed method. The brightness of the gray scale represents the number of the estimated state of spatial concept C_t for each datum over ten trials. The experimental results showed that the proposed method, i.e., the clustering of both position information x_t and word data O_t , had higher estimation accuracy than the clustering of only word data O_t . As a result, we showed that it is possible to improve the lexical acquisition accuracy by performing clustering that considered position and word information.

5. CONCLUSION AND FUTURE WORK

In this paper, we proposed a simultaneous estimation method for self-position and words from noisy sensory information and utterances by a user. We performed experiments and evaluations in a simulator environment. In the experiment for learning of spatial concepts, we showed that the robot was able to accurately learn the spatial concepts of the target places. In the experiment for self-localization, we showed that the robot was able to estimate self-position by employing spatial concepts with lower error than when using MCL. In the experiment for the evaluation of word clustering, we showed that the proposed method can improve the estimation accuracy of spatial concepts, relative to clustering of only word information.

As a similar approach to the proposed method, Ishibushi et al. (2015) proposed a self-localization which exploits image features using CNN (convolutional neural network). We believe that the robot can learn multimodal spatial concepts from positions, images, and words by integrating

image features using our proposed method. Furthermore, online spatial concept learning during the mapping is the object of future work.

REFERENCES

- Hörnstein, J., Gustavsson, L., Santos-Victor, J., and Lacerda, F. (2010). Multimodal language acquisition based on motor learning and interaction. In *From Motor Learning to Interaction Learning in Robots*, 467–489. Springer.
- Inamura, T., Shibata, T., Sena, H., Hashimoto, T., Kawai, N., Miyashita, T., Sakurai, Y., Shimizu, M., Otake, M., Hosoda, K., et al. (2010). Simulator platform that enables social interaction simulation –SIGVerse: SocioIntelliGenesis simulator–. In *IEEE/SICE International Symposium on System Integration*, 212–217.
- Ishibushi, S., Taniguchi, A., Takano, T., Hagiwara, Y., and Taniguchi, T. (2015). Statistical localization exploiting convolutional neural network for an autonomous vehicle. In *Industrial Electronics Society, IECON 2015-41st Annual Conference of the IEEE*, 1369–1375. IEEE.
- Iwahashi, N. (2003). Language acquisition through a human–robot interface by combining speech, visual, and behavioral information. *Information Sciences*, 156(1), 109–121.
- Kitagawa, G. (2014). Computational aspects of sequential Monte Carlo filter and smoother. *Annals of the Institute of Statistical Mathematics*, 66(3), 443–471.
- Lee, A., Kawahara, T., and Shikano, K. (2001). Julius—an open source real-time large vocabulary recognition engine. In *European Conference on Speech Communication and Technology (EUROSpeech)*.
- Nakamura, T., Araki, T., Nagai, T., and Iwahashi, N. (2011). Grounding of word meanings in latent Dirichlet allocation-based multimodal concepts. *Advanced Robotics*, 25(17), 2189–2206.
- Roy, D. and Pentland, A. (2002). Learning words from sights and sounds: A computational model. *Cognitive science*, 26(1), 113–146.
- Taguchi, R., Yamada, Y., Hattori, K., Umezaki, T., Hoguro, M., Iwahashi, N., Funakoshi, K., and Nakano, M. (2011). Learning place-names from spoken utterances and localization results by mobile robot. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 1325–1328.
- Thrun, S., Burgard, W., and Fox, D. (2005). *Probabilistic Robotics*. MIT Press.