

Learning Relationships Between Objects and Places by Multimodal Spatial Concept with Bag of Objects

Shota Isobe^(✉), Akira Taniguchi, Yoshinobu Hagiwara,
and Tadahiro Taniguchi

Ritsumeikan University, 1-1-1 Noji Higashi, Kusatsu, Shiga 525-8577, Japan
{isobe.shota,a.taniguchi,yhagiwara,taniguchi}@em.ci.ritsumei.ac.jp

Abstract. Human support robots need to learn the relationships between objects and places to provide services such as cleaning rooms and locating objects through linguistic communications. In this paper, we propose a Bayesian probabilistic model that can automatically model and estimate the probability of objects existing in each place using a multimodal spatial concept based on the co-occurrence of objects. In our experiments, we evaluated the estimation results for objects by using a word to express their places. Furthermore, we showed that the robot could perform tasks involving cleaning up objects, as an example of the usage of the method. We showed that the robot correctly learned the relationships between objects and places.

Keywords: Spatial concept · Place categorization · Object detection · Human Support Robot · Relationship between objects and places

1 Introduction

Home service robots are required to be able to understand and carry out tasks of cleaning or picking up objects through communications with people. We use various commands, such as “please fetch the cup” and “please put it back”. In the above task, it is necessary to estimate where the object indicated by the word “cup” is located. We consider that robots can communicate effectively with people using vocabularies representing locations, to efficiently perform tasks for estimating where the robot should go to pick up a cup when the cup can be located in multiple places. Furthermore, for the task of “please put it back”, it is necessary to estimate the place in which the presented object should be placed and the vocabulary expressing this. Therefore, we consider that robots should be able to learn the relationships between objects and places in order to carry out such tasks.

In addition, understanding human social interactions and developing a robot that can smoothly communicate with human users in the long term, requires an understanding of the dynamics of symbol systems, such as multimodal categorization [1]. Multimodal categorization involves forming categories based on

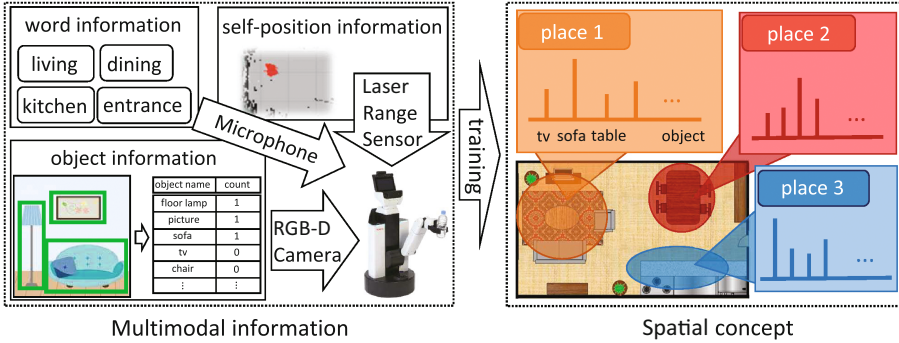


Fig. 1. Overview of the learning relationships between places and objects by the proposed method

sensorymotor information acquired by a robot, including visual information, haptic information, and auditory information. By forming categories by multimodal information, it is possible to classify observation information using each modal information category, and to estimation other modal information from one item of modal information.

Regarding related work on place categorization, Taniguchi et al. proposed a nonparametric Bayesian spatial concept acquisition method (SpCoA) on the basis of unsupervised word segmentation and a nonparametric Bayesian generative model that integrates self-localization and clustering in both words and places [2]. Hagiwara et al. proposed a method that enables robots to autonomously form place concepts using hierarchical multimodal latent Dirichlet allocation (hMLDA) [3], based on position and visual information [4]. In that study, robots are enabled to autonomously form hierarchical place concepts using hMLDA. Further, Ishibushi et al. proposed a method that statistically integrates position information obtained by Monte Carlo localization (MCL) [5] and visual information obtained by a convolutional neural network (CNN) [6, 7]. In that study, the authors demonstrated an ability to converge the positions and orientations of particles using their method, and reduced global positional errors. Further, Espinace et al. proposed a generative probabilistic hierarchical model, where object category classifiers are used to associate low-level visual features to objects, and contextual relations are used to associate objects to scenes [8]. In that study, common objects such as doors and furniture are used as distinguishing features of indoor scenes, as a key intermediate representation for recognizing indoor scenes. Rusu et al. proposed a method of acquisition of semantic 3D object maps that contain those parts of the environment with fixed positions and utilitarian functions for indoor household environments, in particular kitchens, from sensed 3D point cloud data [9]. However, their method cannot perform tasks such as “please fetch the cup” and “please put it back”, because the relationships between objects such as “cup” and places such as “kitchen” are not learned.

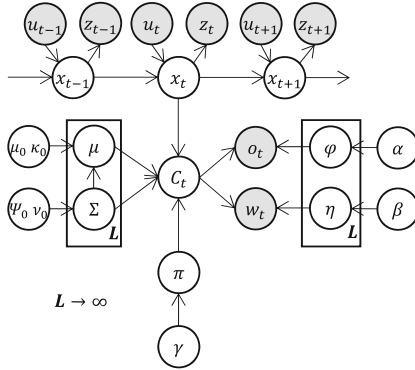


Fig. 2. Graphical model of the proposed method

Table 1. Definitions of variables in the graphical model

x_t	Self-position of a robot
z_t	Sensor data
u_t	Control data
o_t	Object information
w_t	Word information
C_t	Index of spatial concepts
μ, Σ	Normal distribution as a position distribution
φ, η, γ	Parameters of multinomial distribution
π	Multinomial distribution of index of spatial concepts
$\mu_0, \kappa_0, \psi_0, \nu_0$	Hyperparameters of normal-inverse-Wishart prior distribution
α, β	Hyperparameter of Dirichlet prior distribution

An overview of the learning of relationships between places and objects using the proposed method is shown in Fig. 1. In our study, we propose a model that learns the relationships between objects and places from multimodal information of self-localization, object information, and word information. Vocabulary expressing a place is adopted as vocabulary information. Word information constitutes a word expressing a place. Object information is a feature vector expressing a Bag-of-Objects (BoO) representation, detecting an object’s label using an object detection method.

In our experiments, we quantitatively evaluate the estimation results for objects from words expressing their places and estimation results for words expressing places from images.

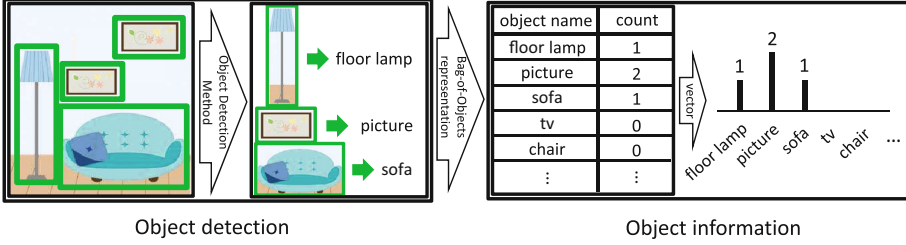


Fig. 3. Object information representing BoO using result of detected object’s label by object detection method

2 Learning of Multimodal Spatial Concepts Based on Co-Occurrences of Objects

In this study, we propose a method that learns the relationships between objects and places using self-position, object, and word information. We define that relationship between an object and a place as the probability the object existing in that place. A graphical model of proposed method is shown in Fig. 2, the definitions of the variables in the graphical model are given in Table 1, and the generative model for the proposed method is given in Eqs. (1)–(10).

$$\pi \sim \text{GEM}(\gamma) \tag{1}$$

$$C_t \sim p(C_t|x_t, \mu, \Sigma, \pi) \propto \frac{\mathcal{N}(\mathbf{x}_t|\mu_{C_t}, \Sigma_{C_t}) \text{Mult}(C_t|\pi)}{\sum_{c'} \mathcal{N}(\mathbf{x}_t|\mu_{c'}, \Sigma_{c'}) \text{Mult}(c'|\pi)} \tag{2}$$

$$\Sigma \sim \mathcal{IW}(\Sigma|\psi_0, \nu_0) \tag{3}$$

$$\mu \sim \mathcal{N}(\mu|\mu_0, (\Sigma/\kappa_0)) \tag{4}$$

$$\varphi \sim \text{Dir}(\alpha) \tag{5}$$

$$\eta \sim \text{Dir}(\beta) \tag{6}$$

$$o_t \sim \text{Mult}(o_t|\varphi_{C_t}) \tag{7}$$

$$w_t \sim \text{Mult}(w_t|\eta_{C_t}) \tag{8}$$

$$\mathbf{x}_t \sim p(\mathbf{x}_t|\mathbf{x}_{t-1}, u_t) \tag{9}$$

$$z_t \sim p(z_t|\mathbf{x}_t) \tag{10}$$

Here, $\text{GEM}(\cdot)$ is the prior distribution configured using a stick breaking process (SBP) [10], $\mathcal{IW}(\cdot)$ is the inverse-Wishart distribution, $\mathcal{N}(\cdot)$ is a multivariate normal distribution, $\text{Mult}(\cdot)$ is a multinomial distribution, and $\text{Dir}(\cdot)$ is a Dirichlet distribution. Robots estimate their self-position with MCL, using a map created by simultaneous localization and mapping (SLAM) [5]. Moreover, in order to detect objects from images, we use you only look once (YOLO) [11] which is an object detection method. A method of acquiring object information is illustrated in the Fig. 3. Robots acquire object information o_t representing BoO using the

result of the object’s label detected by YOLO from images acquired at time t . Furthermore, using the bounding box acquired by YOLO, object information is weighted according to Eq. (11) using the obtained depth information. Because learning is performed based on the position of the robot, weighted is accordingly performed to avoid the influences of distant objects.

$$weight(d) = \exp\left\{-\frac{\zeta d}{(D-d)}\right\} \quad (11)$$

Here, d is the depth information for each object observed by the robot, D is a value for setting a convergence point where the weight becomes zero, and ζ is the damping factor. From Eq. (11), the attenuation factor is increased as the value of the distance increases.

Self-position information is defined as $\mathbf{x}_t = (x_t, y_t, \sin \theta_t, \cos \theta_t)$, where (x, y) is the self-position value of the robot in two-dimensional coordinates, and θ_t is the direction of the robot. The angle θ_t is such that the angle to the x axis is 0° and angle to the y axis is 90° . Furthermore, u_t and z_t represent the control information of the robot and observation information from the distance sensor, respectively. The object information o_t is defined as $o_t = (o_t^1, o_t^2, \dots, o_t^I)$, where I is the number of categories of objects that can be detected by the object detection method. A human gives the name of the location corresponding to the self-position information \mathbf{x}_t using vocabulary information w_t . The number of spatial concepts is determined stochastically by SBP.

For learning spatial concepts in the proposed method, each parameter is estimated using Gibbs sampling. The procedure for sampling each parameter using the Gibbs sampling is shown in the Eqs. (12)–(15), where $\mathcal{NIW}(\cdot)$ is the normal-Inverse-Wishart distribution; $\psi_{n_l}, \nu_{n_l}, \mu_{n_l}, \kappa_{n_l}$ are hyperparameters after updating; and x_l, o_l, w_l are sets of self position information, object information, and word information data at $C_t = l$, respectively. Furthermore, $C_t, \mu, \Sigma, \varphi, \eta$ are parameters estimated by Gibbs sampling.

$$\begin{aligned} C_t &\sim p(C_t = l | \mathbf{x}_t, \mu, \Sigma, \pi, \varphi, \eta) \\ &\propto \mathcal{N}(\mathbf{x}_t | \mu_{C_t}, \Sigma_{C_t}) \text{Mult}(o_t | \varphi_{C_t}) \\ &\quad \times \text{Mult}(w_t | \eta_{C_t}) \text{Mult}(C_t | \pi) \end{aligned} \quad (12)$$

$$\begin{aligned} \mu_l, \Sigma_l &\sim \mathcal{N}(x_l | \mu_{C_t}, \Sigma_{C_t}) \mathcal{NIW}(\mu_l, \Sigma_l | \psi_0, \nu_0, \mu_0, \kappa_0) \\ &\propto \mathcal{NIW}(\mu_l, \Sigma_l | \psi_{n_l}, \nu_{n_l}, \mu_{n_l}, \kappa_{n_l}) \end{aligned} \quad (13)$$

$$\varphi_l \sim \text{Multi}(o_l | \varphi_l) \text{Dir}(\varphi_l | \alpha) \quad (14)$$

$$\eta_l \sim \text{Multi}(w_l | \eta_l) \text{Dir}(\eta_l | \beta) \quad (15)$$

3 Experiment

An experiment is performed using the proposed model to estimate objects from vocabulary information expressing places, and to estimate the vocabulary expressing places from images, and a quantitative evaluation makes it possible

to judge the relevance relations between objects and places. In addition, we show the usefulness of our proposed model by actually carrying out the task of having the robot clear up an object using the proposed model.

3.1 Experimental Condition

In this experiment, we conduct experiments using TOYOTA’s Human Support Robot (HSR) [12]. The experimental environment is a home environment in the house owned by our laboratory. The layout of the experimental environment is illustrated in Fig. 4. It is assumed that the map is generated by SLAM in advance, using a laser range sensor, and that the robot has a map. Self-position estimation is performed using the amcl (adaptive MCL) package of Robot Operating System (ROS) [13]. The dictionary of the obtained word information contains the following: “The front of dining table”, “The front of TV”, “The front of trash box”, “The front of microwave rack”, “The front of sink”, “The front of bookshelf”, “The front of refrigerator”, “The front of living table”, and “The front of sofa”. Word information is allocated to 10% of the data of the self position information. Because we used the pre-learned darknet 19 model [11] in the dataset MS-COCO¹ for YOLO, the object information consists of 80 dimensions. The parameters for weight calculation are $D = 4$ and $\zeta = 0.7$. The other parameters for this experiment are $\alpha = 0.1$, $\beta_0 = 0.1$, $\gamma_0 = 10$, $\mu_0 = (-0.05, -0.74, -0.01, -0.27)$, $\kappa_0 = 1.0$, $\nu_0 = 15$, and $\psi_0 = \text{diag}(0.05, 0.05, 0.05, 0.05)$. In addition, the number of iterations used for Gibbs sampling is 100. In order to verify the validity of the learned relationships between objects and places, three objects are selected from 80 objects that can be detected and set as correct labels. A correct label is created by a person who knows the experimental environment.

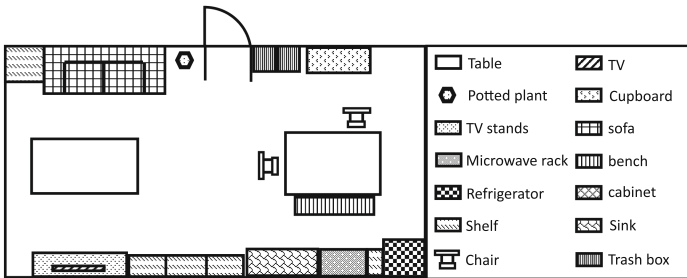


Fig. 4. Layout of the experimental environment

¹ MS-COCO: <http://mscoco.org/>.

3.2 Experimental Procedure

While estimating its self-position using MCL, a robot moves in the environment by operation of the joy stick, and acquires self-position information and images at each self-position. In this experiment, word information gives data by typing in order to eliminate the problem of speech recognition error. The amount data giving word information is randomly determined. We use YOLO to detect an object in the image and acquire object information representing BoO. Furthermore, using a bounding box acquired by YOLO, obtained object information is weighted according to the Eq. (11), using the obtained depth information. Subsequently, a robot learns the relationships between objects and places by the proposed method using self-position, object, and word information. We confirm the spatial regions of each learned place by drawing a normal distribution on the map. We evaluate the proposed model by a quantitative evaluation that estimates the results for objects from the vocabulary information expressing places, and the results for the vocabulary expressing places from specified objects. Equation (16) is used to estimate the word W expressing a place by the occurrence probability of the feature vector O obtained from the presented image. We compare this with a model designed to handle word information in Ishibushi’s method [7]. Furthermore, we perform a comparison with the number of image features used in Ishibushi’s model in constructing the final and middle (fc6) layers of CNN. Equation (17) is used to estimate objects O at a place from the word W expressing that place. We compare the results with multimodal HDP-LDA. Multimodal HDP-LDA enables the multimodal handling of HDP-LDA in the topic distribution of LDA as Hierarchical Dirichlet Process (HDP) [14]. Here, HDP-LDA was learned using object information and word information. Finally, we demonstrate the usefulness of the proposed model by applying it to the task of the robot actually clearing up objects.

$$\begin{aligned}
 W &= \arg \max_{w_t} p(w_t | o_t, \eta, \varphi, \pi) \\
 &= \arg \max_{w_t} \sum_{C_t} p(w_t | \eta_{C_t}) p(o_t = O | \varphi_{C_t}) p(C_t | \pi) \quad (16) \\
 O &\sim p(o_t | w_t, \varphi, \eta, \pi) \\
 &\propto \sum_{C_t} p(o_t | \varphi_{C_t}) p(w_t = W | \eta_{C_t}) p(C_t | \pi) \quad (17)
 \end{aligned}$$

3.3 Experimental Result

An example of a position distribution formed for each place and an image classified is presented in the Fig. 5. The left side shows the position distribution learned when the object information is represents the BoO, and the right side shows position distribution when the object information is weighted according to the Eq. (11) from the obtained depth information. In Fig. 5, 10 spatial regions estimated by learning are shown, and can be identified by color. The position and

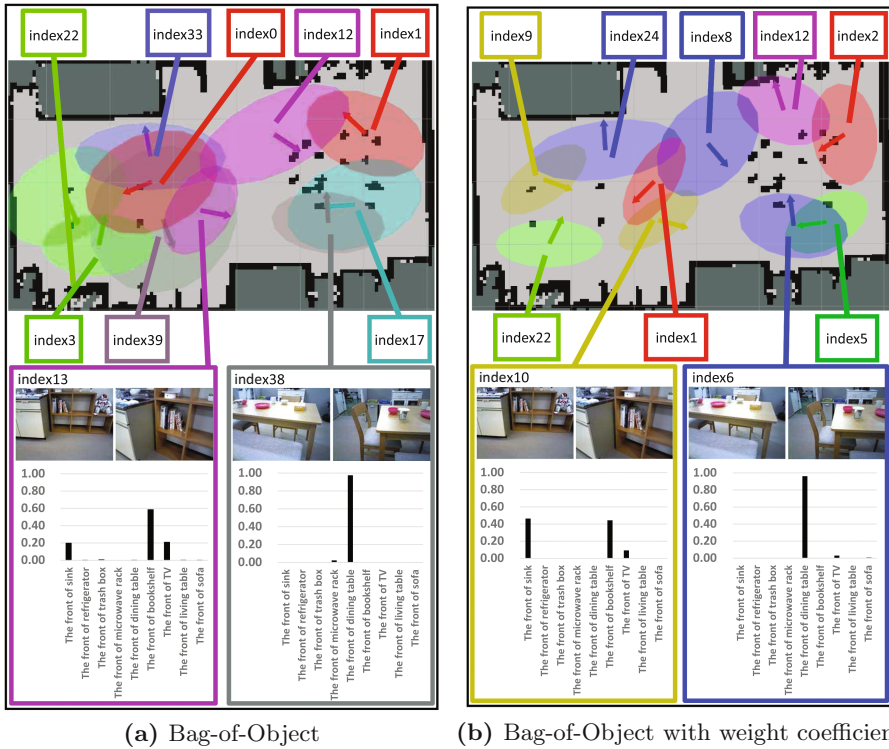


Fig. 5. An example of a position distribution formed for each place and classified image when object information is represented by BoO (a), and its weight (b) (Color figure online)

direction of the arrow shows center and direction of the location area, and the translucent circle represents the covariance matrix of the spatial region. Although we need to discuss whether to include the direction in the spatial region estimated by learning, in this experiment we used data including directions to learn spatial concepts, so we showed arrows as a result of learning. The color of a circle identifies the spatial region, and does not indicate any relationship. Each image is an example of an image assigned to a spatial region. Furthermore, each histogram represents the probability of the occurrence of a word expressing the place, as obtained by learning. From this point, each spatial region will be indicated by its index. As can be seen from Fig. 5, since the indexes 0 and 33 differ in the directions in which visible objects are facing, they are distinct from each other. Moreover, when object information is weighted in the BoO representation, the range of spatial regions became smaller, because the learned spatial concept involves approaching the place where an object exists using the place in which the object can be seen. Table 2 presents the results of estimating words expressing places from images, and the accuracy, from more than 50 test data points. Table 2 shows that despite the fact that a few objects can be detected,

Table 2. An example of the results of estimating words expressing places




Input image	Our method (BoO representation)	Our method (weighted BoO representation)	[7]+word (final layer)	[7]+word (middle layer)	Ground truth
	The front of living table	The front of living table	The front of living table	The front of living table	The front of living table
	The front of bookshelf	The front of sink	The front of bookshelf	The front of sink	The front of bookshelf
	The front of microwave rack	The front of microwave rack	The front of microwave rack	The front of sink	The front of microwave rack
Accuracy rate	0.75	0.64	0.71	0.82	

Table 3. Examples of object estimation results

Word information	Proposed method (BoO representation)	Proposed method (Weighted BoO representation)	HDP-LDA	Ground truth
The front of refrigerator	refrigerator skis microwave	refrigerator microwave bowl	book chair refrigerator	refrigerator microwave dining table
The front of trash box	chair bowl cup	microwave oven bowl	book refrigerator potted plant	chair dining table cup
The front of dining table	chair bowl cup	bowl chair dining table	book potted plant couch	bowl chair dining table
The front of bookshelf	book potted plant refrigerator	book refrigerator potted plant	chair bowl cup	potted plant book tv
The front of sofa	couch potted plant chair	couch potted plant refrigerator	refrigerator chair microwave	couch laptop potted plant
Accuracy rate	0.52	0.48	0.11	

our proposed method did not differ much from the final layer of CNN in terms of accuracy. Table 3 shows the results of estimating objects from words expressing places. From the Table 3, it can be seen that the accuracy of the proposed method is better than that of HDP-LDA. There was a bookshelf presents just inside the kitchen, and so the probability that a book exists is high. In addition, using the proposed model, we confirmed the task of the robot actually carrying out the task of clearing up could be performed. The movie that performed the task to clean up the object, the source code of our proposed method, and dataset are publicly available².

² movie, source code, and dataset: <https://emlab.jimdo.com/multimedia/>.

4 Conclusion and Future Work

In this study, we have proposed a method that learns the relationships between objects and places using self-position, object, and word information. Experimental results showed that the proposed method can estimate objects from words expressing their places, and estimate the words expressing places from images. In our experiment to estimate words expressing places from image, our method achieved an equivalent performance to Ishibushi's method, but we consider that our method is more useful than that method, in that we can learn the relationships between objects and places. Furthermore, in the experiment to estimate objects from the words expressing their places, when using BoO a distant chair could be located, but in the case of weighting it was estimated to be located on the wrong side of a table. From this result, we consider that the learned spatial concept involves approaching the place where an object exists by using the place where it can be seen. From a quantitative evaluation of the results of estimating objects from words expressing their places and the correct labels, the validity of the relationships between objects and places obtained by the learning of this proposed model was demonstrated.

In this study, we conducted experiments on the relationships between objects and places. In the future, we will conduct experiments on the estimation of positions and movement, and we will further verify the effectiveness of proposed method. In addition, it is necessary to fine-tune YOLO, so that it is possible to detect objects in the home. In future work, we are considering conducting relative spatial concept learning [15], such as with the phrase “the front of”.

References

1. Taniguchi, T., Nagai, T., Nakamura, T., Iwahashi, N., Ogata, T., Asoh, H.: Symbol emergence in robotics: a survey. *Adv. Robot.* **30**(11–12), 706–728 (2016)
2. Taniguchi, A., Taniguchi, T., Inamura, T.: Spatial concept acquisition for a mobile robot that integrates self-localization and unsupervised word discovery from spoken sentences. *IEEE Trans. Cogn. Dev. Syst.* **8**(4), 285–297 (2016)
3. Ando, Y., Nakamura, T., Araki, T., Nagai, T.: Formation of hierarchical object concept using hierarchical latent Dirichlet allocation. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2272–2279. IEEE (2013)
4. Hagiwara, Y., Masakazu, I., Tadahiro, T.: Place concept learning by hMLDA based on position and vision information. *IFAC-PapersOnLine* **49**(19), 216–220 (2016)
5. Thrun, S., Burgard, W., Fox, D.: *Probabilistic Robotics*. MIT press, Cambridge (2005)
6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*, pp. 1097–1105 (2012)
7. Ishibushi, S., Taniguchi, A., Takano, T., Hagiwara, Y., Taniguchi, T.: Statistical localization exploiting convolutional neural network for an autonomous vehicle. In: *IECON 2015–41st Annual Conference of the IEEE Industrial Electronics Society* pp. 001369–001375. IEEE (2015)

8. Espinace, P., Kollar, T., Roy, N., Soto, A.: Indoor scene recognition by a mobile robot through adaptive object detection. *Robot. Auton. Syst.* **61**(9), 932–947 (2013)
9. Rusu, R.B., Marton, Z.C., Blodow, N., Holzbach, A., Beetz, M.: Model-based and learned semantic object labeling in 3D point cloud maps of kitchen environments. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2009*, pp. 3601–3608. IEEE (2009)
10. Ishwaran, H., James, L.F.: Gibbs sampling methods for stick-breaking priors. *J. Am. Stat. Assoc.* **96**(453), 161–173 (2001)
11. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. arXiv preprint [arXiv:1612.08242](https://arxiv.org/abs/1612.08242) (2016)
12. Personal Assist Robot, Human Support Robot (HSR). http://www.toyota-global.com/innovation/partner_robot/family_2.html
13. ROS.org. <http://wiki.ros.org/>
14. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Sharing clusters among related groups: hierarchical Dirichlet processes. In: *Advances in Neural Information Processing Systems*, pp. 1385–1392 (2005)
15. Gu, Z., Taguchi, R., Hattori, K., Hoguro, M., Umezaki, T.: Learning of relative spatial concepts from ambiguous instructions. *IFAC-PapersOnLine* **49**(19), 150–153 (2016)