2016 IEEE-RAS 16th International Conference on
Humanoid Robots (Humanoids)
Cancun, Mexico, Nov 15-17, 2016

Transformable Semantic Map Based Navigation using Autonomous Deep Learning Object Segmentation

Yuki Furuta[1], Kentaro Wada[1], Masaki Murooka[1],
Shunichi Nozawa[1], Yohei Kakiuchi[1], Kei Okada[1] and Masayuki Inaba[1]

***Abstract*— For daily assistive robots working in home environment, it is important to use geometry-free representation for navigation which can deal with dynamic environmental changes. In this paper we propose semantic map based navigation which consists of 1) generating deep learning enabled semantic map from annotated world and 2) object based navigation using learned semantic map representation. One point of our proposed framework is to let robots autonomously generate a dataset for deep learning method and transfer existing geometric map based task execution system to semantic map based one which is invariant to changes of object location. Since deep learning for object segmentation technique enables end-to-end learning of object features, it is not necessary to design segmentation and labeling methods for each objects manually. We confirmed the effectiveness of our approach by performing task in dynamic environment and adaptability for two type of robots with experiments.**

I. INTRODUCTION

Daily assistive robots are expected to perform a wide range of tasks including object manipulation and navigation. To manipulate objects, robots have to know not only geometric models for the target objects but also locations to approach them enough close in order to register with high accuracy.

Recent progress of SLAM (Simultaneous Localization and Mapping) technology enables to let robot agents navigate around home environment through building geometric map from odometry. Representation of geometric maps generated with SLAM depends on kinds of sensors on robots and navigation with them works well on static environment. On dynamic environment it would be efficient for navigation to build maps where rather consist of relationship among objects than grid based representation, though segmentation and recognition of each objects greatly depended on hard-coded object recognition methods.

In this paper, we are introducing a solution for the problem by integrating deep learning architecture. With deep learning object segmentation which enables end-to-end learning without designing object specific feature descriptor, we can generate semantic maps with which robots can identify target objects and navigate robots to them.

Deep learning based methods require obtaining enormous annotated image sensor data in a real environment as input data. To avoid annotating all objects in images by manual, we instead use annotated world representation that robot have used for map-based navigation and autonomously create annotated image dataset for learning from it. This turns a supervised learning problem into a self-supervised one, as all dataset are annotated by robots themselves. The difference between well annotated dataset used in the field of image recognition and dataset generated from real robot sensor data is that they are always subject to noise from sensor device and environment. This difference affects performance of object segmentation significantly, so our goal is to clarify the method to acquire dataset of good quality for learning from noisy data.

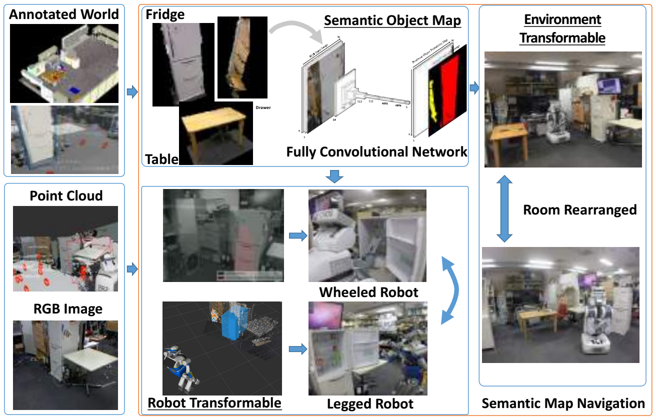


Fig. 1: Overview of transformable task executive system
The robots use semantic map for navigation based on deep learning enabled object segmentation and labeling learnt from local dataset autonomously generated during task performance.

We also propose a method to transform task from existing geometric map based task executive system into semantic map based navigation. One benefit of this approach is that the robot agents can acquire feature of objects for navigation end-to-end from generated dataset, so we no more need to formulate any hypothesis for identifying objects.

Through experiments, we confirmed that robots with our proposed method can deal with daily assistive tasks even in dynamic environment. Also we confirmed that our framework is independent to specific robots by applying to a different humanoid robot.

Our contribution is summarized as follows:

1) Autonomous generation of dataset for deep learning object segmentation from image sensor data in deployed environment
2) Robots can perform tasks in dynamic environment where location of entities is transformable by building semantic map from geometric map

[1] Y.Furuta, K.Wada, M.Murooka, S.Nozawa, Y.Kakiuchi, K.Okada and M.Inaba are with Graduate School of Information Science and Technology, The University of Tokyo 7-3-1, Hongo, Bunkyo-city, Tokyo, 113-8656, Japan `furushchev at jsk.imi.i.u-tokyo.ac.jp`

978-1-5090-4717-8/16/$31.00 ©2016 IEEE

3) Semantic map is also invariant to robot, therefore task can be described and performed not by specific robot.

An overview of our proposed framework is shown in Fig.1. First, we divide our entire framework into two major components, dataset generation from annotated world representation and sensor data during early task performance (Section III), and applying obtained dataset for object based navigation (Section IV). Then we evaluate our framework by conducting three experiments (Section V). Finally, we conclude by discussing the result of the experiments and future work (Section VI).

II. RELATED WORK

Integrated robot systems for performing robot tasks in home environment have been researched [1], [2], and we have also conducted researches about integrated robot system which can deal with daily assistive task sequences including complex procedure dependent on environment [3].

With these frameworks robots use geometric map where objects manipulated by robots are placed in a globally static map called "annotated world", and localization of robot pose with SLAM technology.

There are various methods that have been proposed to represent mapping of environment for navigation. Mapping with raw spartial data by Fox et al. [4] performed high precision of self localization, though this approach does not concern spartial relationship of objects and significantly depends on a characteristic of sensor devices.

Kuipers showed more abstracted representation of spartial mapping [5], where environment is represented as topological map derived from the theory on mapping and symbolization by human. [6] also showed that symbolic representation of map has high intuitiveness performing various type of tasks with "spot", which is meaningful location for manipulation. Mapping with symbolic representation can be compact and robust for a kind of dynamic change of environment, though the problem is the method of symbol extraction from raw sensor data is heuristic or fully hard-coded.

Our proposed approach integrates semantic maps with existing geometric object maps ("annotated world") by continuously updating object locations from data of early task performance.

This requires to segment and label objects to estimate object locations for continuous update of semantic map. Object segmentation has been researched for a long time in the field of both robotics and vision processing.

There are many methods about object segmentation which have been proposed before. Radhakrishna et al. proposed SLIC superpixels [7] to segment image regions and create object contours as clustered segmented superpixels by k-means. Then given segmented object images, we still need to classify each objects to give them labels. Or simply matching object images and camera images with SIFT [8] are also often used.

These methods need to design algorithms to extract features which are different for each objects. In contrast, our proposed method with deep learning object segmentation enables end-to-end learning in which feature of objects are automatically generated as weights of neural network.

Since deep learning has emerged, it has been showing powerful results than existing methods. In the field of both object classification [9], [10] and object detection [11], [12], deep learning based architectures have marked significantly good results. Long et al. proposed FCN (Fully Convolutional Networks) [13] where network is fully convolutional that enables end-to-end learning for object segmentation.

III. GENERATING DATASET FROM ANNOTATED WORLD

To enable deep learning based object annotating method in robot systems deployed in real world, it is necessary to create dataset suitable for each environment in which robots are deployed. In this section, we describe the way to autonomously generate dataset for deep learning based object segmentation in the perspective of handling robot sensor information.

A. Dataset for deep learning based object segmentation

In the field of image processing, there are well prepared large scale image dataset for learning. ImageNet [14] provides large scale image dataset including more than 1 million images with bounding box annotations and has contributed to success in many recent work of deep neural network [15], [9]. Though this dataset for deep learning is constructed by crowd sourcing and all images are annotated manually, it is very difficult to build original dataset suitable for learning in each environment where robots are deployed. To reduce the cost of labeling all data, some researches have been conducted [16], [17]. Images used in the most of these works are already almost segmented and labels are given as general purpose. In contrast, to perform daily assistive tasks in home environment robots need to recognize objects distinctive in each environment. Also image data obtained from robot sensor is not segmented at all.

In model-based robot system, robots have internal environmental model of real world to plan actions to achieve task goal. We use this internal environmental model to segment and label objects in camera image data.

B. Autonomous annotated image generation from annotated world

Fig.2 shows entire pipeline to obtain annotated object images for learning from sensor raw data and annotated world and robots have RGB-D sensor such as Microsoft Kinect. In this phase we assume robots can know positions of themselves from world coordinates. In an annotated world, locations, bounding boxes and labels of all objects from world coordinates are known. Goal of generating dataset is to generate annotated object image from sensor data. One possible solution is projection of object bounding boxes into camera perspective of robots. We can generate annotated images by simple computation, but when a part of target objects is occluded by other objects in the projected bounding

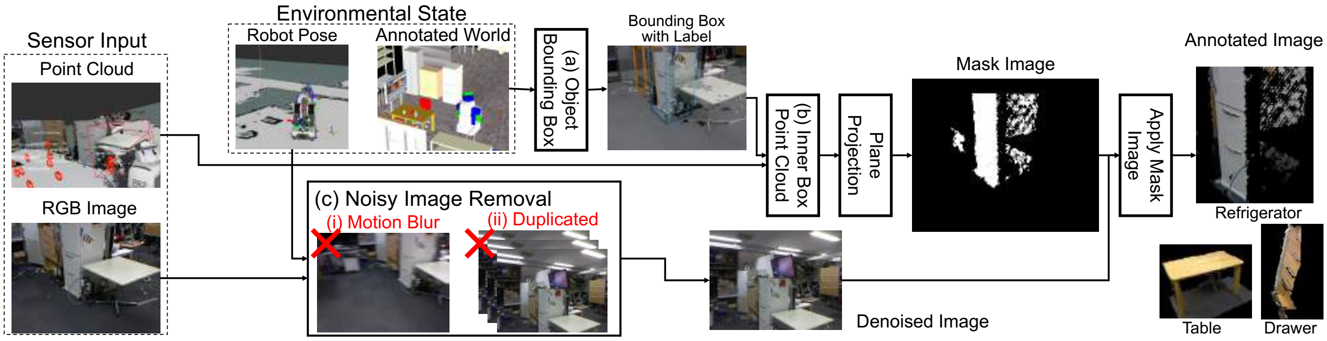


Fig. 2: The pipeline of generating dataset for learning from real robot sensor and annotated world

box (which we should count as noise), they cannot be removed.

In our proposed method, we first calculate bounding boxes of objects (Fig.2 (a)), then select pointclouds only inner bounding boxes, and project them into perspective of RGB camera (Fig.2 (b)). We thus generate labeled mask images from object bounding boxes and pointclouds from depth sensor. Finally annotated images are generated by applying labeled mask images to color images.

C. Noisy image removal using robot pose information

Candidate annotated images still contain noises derived from various causes: noises from sensor hardware devices, motion blur noises while camera device is moving and so on. There are also many duplicated images captured from the same position which causes imbalanced learning problem [18]. To address the former noises, we remove images while camera device is moving fast by filtering with the absolute velocity of the sensor device from world coordinates (Fig.2 (c)(i)). We use 0.01m/s as the threshold for both translational and rotational velocity. For the latter problem, we keep the first image at each pose of camera, and then remove images till the pose of the camera device is different from one at the previous time (Fig.2 (c)(ii)). We use parameter 0.01m/s for this filtering.

IV. OBJECT SEGMENTATION WITH DEEP LEARNING ENABLED ARCHITECTURE

In this section, we propose a new system to segment target object in 3D real world. Our proposed system consists of two components: the first is the 2D image segmentation by Deep Convolutional Network, and the second is 3D registration of the target object region with box fitting for localization of target object and size estimation.

A. Fully Convolutional Networks for 2D Object Segmentation

The early Researches on object segmentation task have been tackled with 2D image [19], and recent works with Fully Convolutional Networks (FCN) architecture report state-of-the-art results for this object segmentation task. [20] [21]

We also use this FCN architecture for our object segmentation system for our 5 object classes segmentation problem in home environment. The network architecture is shown in Fig.3, which consists of 16 convolutional layers and 6 max pooling layers. The numbers in the figure represents the channel size of each convolutional outputs, and (H, W) are the height and width of the input image respectively.

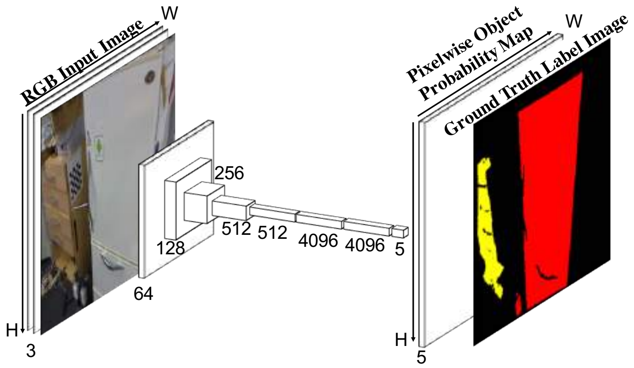


Fig. 3: Fully convolutional networks architecture

Our segmentation task contains only 5 object classes and environment where our system is used is focused on home. Compared to the previous works [19], in which the number of classes is over 20 and background of images does not depend on each other, it seems that our problem setup is just easy. However, there is the difference between our problem setup and previous works in that the dataset we use is not labeled by human but by robot. As shown in Fig.4, the dataset for our task includes annotated images whose label values are wrong. In this figure, the red represents the region labeled for refrigerator, blue for table, yellow for drawer, and no color for background. In failure examples, the labeled region is out of alignment because of the undetectable camera moving, mis-localization in SLAM, and the difference between the annotated world and the real world. In order to deal with these uncertainties in the dataset, we need to consider unreasonable training signals at the learning phase of the FCN, for avoiding wrongly labeled data is treated as a ground truth. We apply the gradient clipping strategy [22] to solve this problem, with suppressing the explosion of gradient and applying soft constraints. For this study, we set 5 as the L2 norm threshold for gradient clipping.

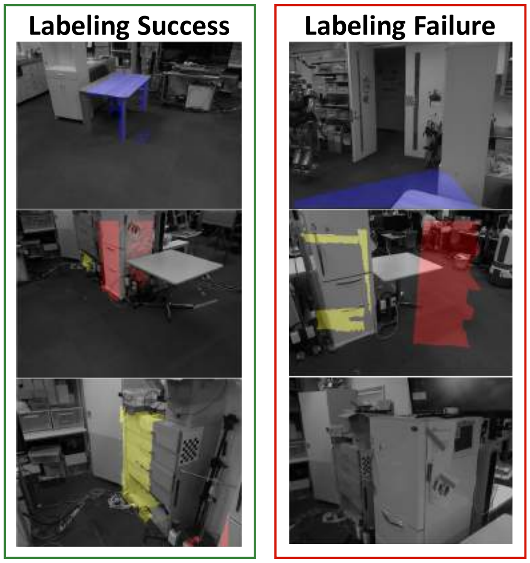


Fig. 4: Dataset generated with our method which is described in Section III. Left side of this figure shows the successfully labeled images, and right side shows the wrongly labeled images.

We split the dataset to 8:2 for training and validation, which contains 418 sets of RGB image and Label images. For the optimizer to minimize loss function, we used the Adam [23] with $\alpha = 1e - 5$ and $\beta = 0.9$ parameters. The learning curve is shown in Fig.5, and it shows the prediction accuracy arises in both training and validation dataset. The upper images in the Fig.5 represent the segmentation result at some states of training iterations, and it also shows the segmentation result enhances with training iterations.

B. 3D Object Segmentation System for Object-based Navigation

In previous section, we proposed a new method to segment object on 2D image with FCN Deep Learning architecture. The network outputs the label image which represents object label in each pixel. In this section, we describe our system to use this segmentation result on 3D world.

Fig.6 shows our 3D segmentation system which consists of three components: first is the registration of FCN segmentation result to the 3D point cloud, second is the denoising filtering for segmentation by clustering points, and third is the box fitting to estimate detected object size. The purposes of each component of this system are to use the object segmentation for object-based navigation task by detecting and localizing the target object to approach to it, and the box fitting process is to estimate object center and size for localization and threshold for detection.

V. EXPERIMENTS

In this section, we evaluated our proposed framework in three experiments: a) a PR2 robot performs generating annotated image dataset from annotated world by letting robot move about (Section V-A), b) a PR2 robot performs daily assistive task with modified position of furniture (Section V-B.1). From these two experiments, we confirmed that our framework has enough robustness to let robots perform task in dynamic environment. To show that our proposed framework is independent to a specific robot type, we also conducted another experiment where c) a HRP2 robot performs the same task with the same framework as a PR2 (Section V-B.2).

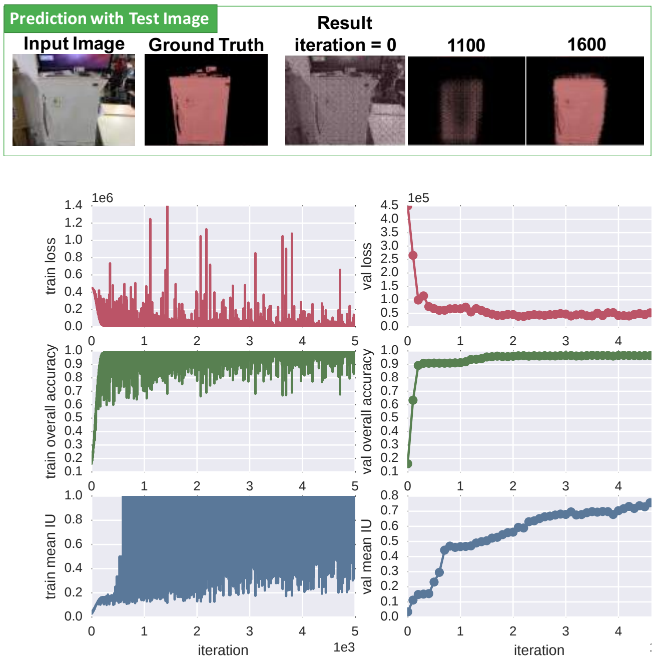


Fig. 5: Learning curve and segmentation result with test image on FCN. The upper images represent changes in segmentation result with iterations, and the lower figures show the learning curve of loss, accuracy and mean IU for both training (left side) and validation (right side) dataset.

A. Acquisition Phase

In this experiment, we used the Willow Garage's PR2 robot with head-mounted Microsoft Kinect RGB-D sensor. To create dataset from robot sensor data, we let a PR2 robot perform "moving around" task. This task consists of moving around a room with base and looking around rooms with moving head.

Transformations from robot to object bounding boxes are updated at about 60 Hz. We filtered not to use camera images while the robot is moving at the speed of camera device was faster than 0.01 m/s and also filtered duplicated images captured closer than 0.03m from previous camera position. After performing moving around task about 250 minutes, 450,973 color images and 445,611 depth images were captured by camera sensor and 417 filtered data were generated as annotated images (Table I).

B. Task Execution Phase

1) PR2 fetches can from refrigerator before and after room design is changed: After generating dataset, we con-

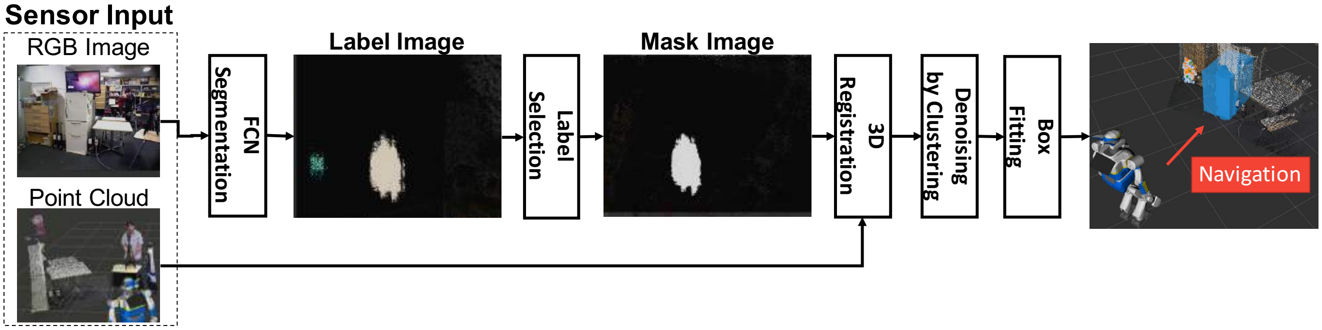
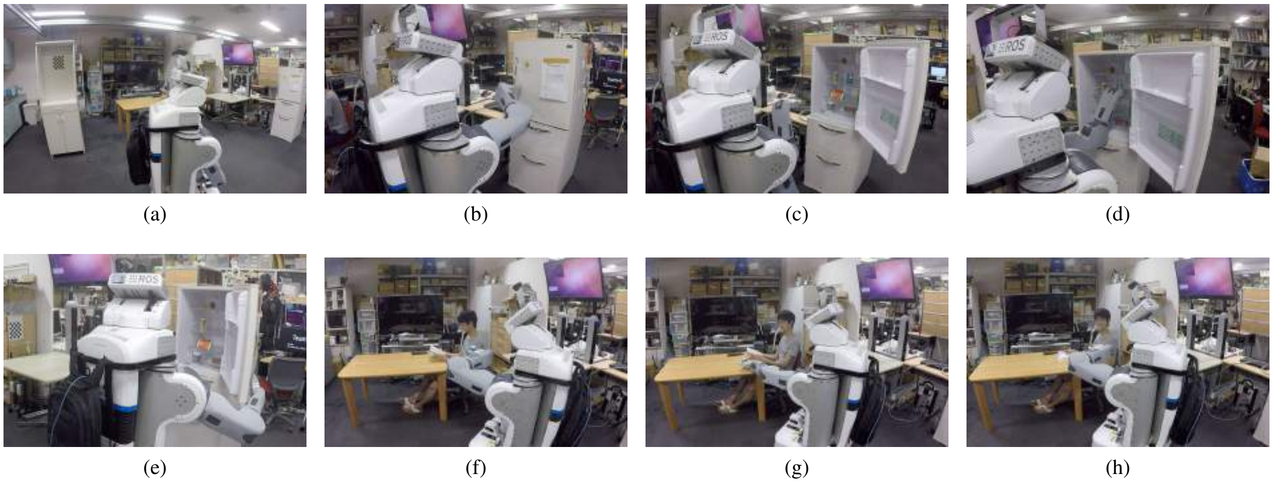


Fig. 6: Our proposed system for 3D object segmentation for object-based navigation



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 7: Robot fetches can from refrigerator after repositioning furniture and home electric appliances

Object Name	Number of Images	Rate
Background	417	100.0%
Refrigerator	162	38.85%
Drawer	119	28.54%
Table	81	19.42%
Door	63	15.10%
All class	417	100.0%

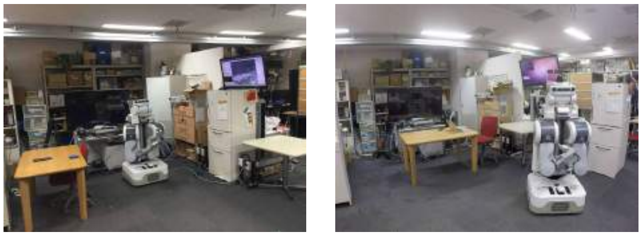
TABLE I: Annotated Object Class and Data

firmed effectiveness of our framework with "fetching can from refrigerator" task on home environment. The setup of experiment consists of a table, refrigerator and coffee can which is placed in the refrigerator.

This task is proceeded as below:

1) In Fig.7a, the robot approached to the refrigerator, target objects.
2) In Fig.7b, the robot opened door by detecting handle of refrigerator.
3) In Fig.7c-Fig.7d, the robot picked can up.
4) In Fig.7e-Fig.7f, the robot found the table and approaches there.
5) In Fig.7g-Fig.7h, the robot placed can to the table.

For planning this task and managing failure and recovery actions, we used "task compiler" [24]. In task compiler, given domain which represents environmental states and goal states of task, action transition graph to transit a current state to the goal is generated. The advantage of using task compiler is we can plan not only just action sequence but also failure states and recovery action from them. Fig.9 shows the entire action transition graph which represents task description for the task in this experiment.



(a) Environmental setup before rearranging the room

(b) Environmental setup after rearranging the room

Fig. 8: Experimental setup of home environment

At initial environment shown in Fig.8a where real environment is almost completely the same as annotated world as robot internal representation, we tested that the PR2 robot successfully performed the "fetching can from refrigerator" task with our proposed system. Then after relocating refrigerator, table randomly, we confirmed the

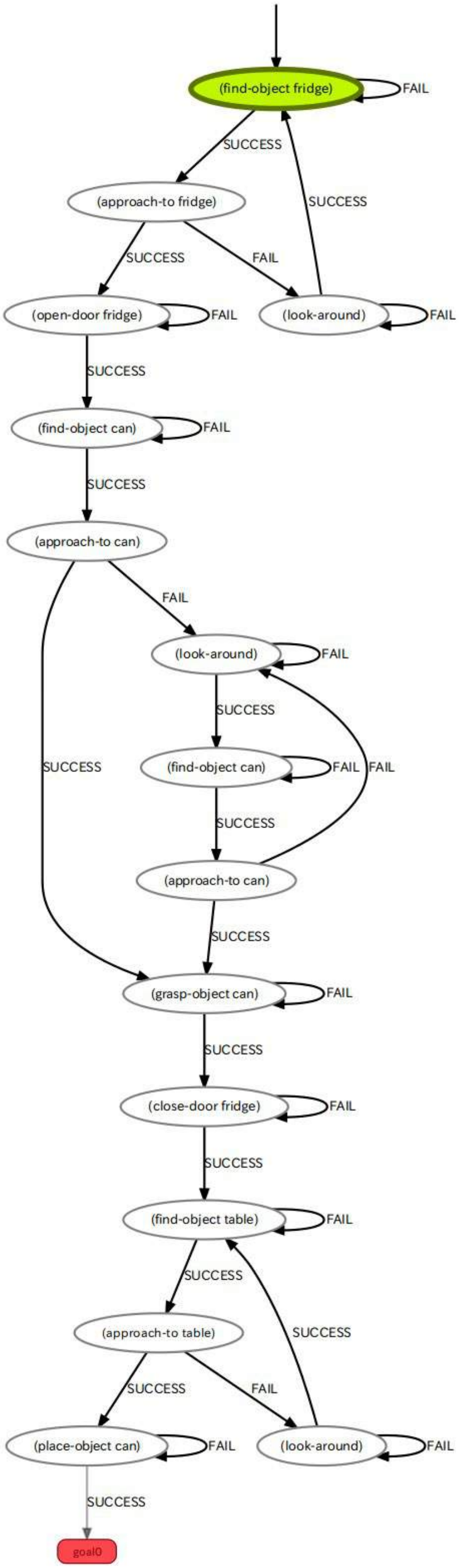
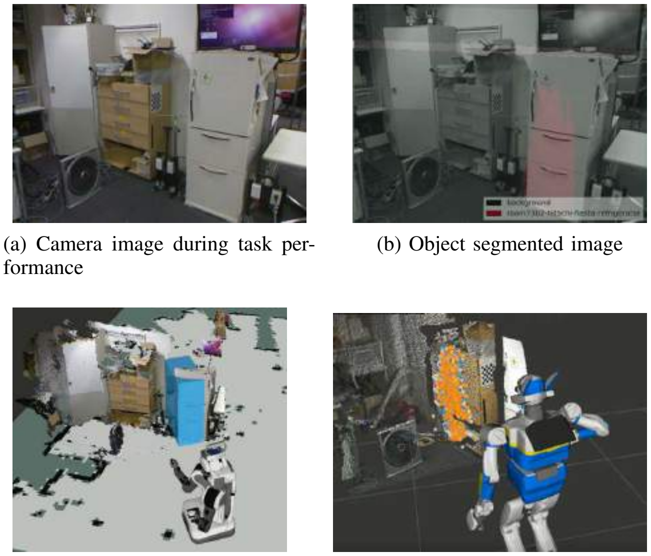


Fig. 9: Action transition graph for "fetching can from refrigerator" task



(a) Camera image during task performance

(b) Object segmented image

(c) Bounding box as object position estimation result for navigation

(d) Projection of object annotation result on pointcloud

Fig. 10: Object segmentation and object based navigation based on deep learning architecture

robot still successfully achieved the task. On rearranged room environment, the robot first tried to find refrigerator at `(find-object fridge)` in Fig.9, then after he found that the refrigerator does not exists where it was located in annotated world, action is transited to `(look-around)`, then again robot try to find the target object with deep learning based object segmentation. After the robot found the refrigerator, the robots approach it (`(approach-to fridge)`) and began to manipulate (`(open-door fridge)`). After the target objects are found with our proposed framework and robots successfully enough close approach them to register with the models they have, robot can manipulate them using the exitsing model based approaches.

2) HRP2 fetches can from refrigerator with same framework: We then let HRP2 robot perform the same task as PR2 did in the previous section, where HRP2 robot has no assumption that robots should know their location in environment. We confirmed that our approach using semantic map instead of geometrical map for navigation can extend feasibility of task performance with various types of robot through this experiment. In current 2D geometric map based SLAM, the performance of localization is sensitive to the sensor attached position. In contrast, with semantic map based navigation, robot is navigated by object segmentation, so the robot can approach to target objects even in case of badly localized.

VI. CONCLUSIONS

In this paper, we proposed a novel framework to generate semantic map for navigation for dealing with dynamic environment by automatically generating dataset for learning

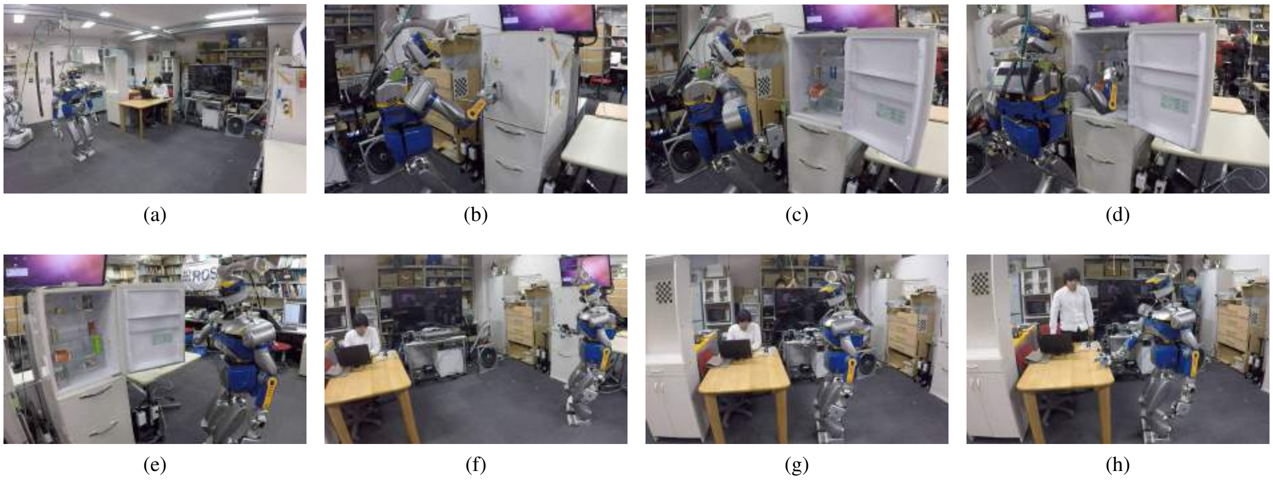

(a) (b) (c) (d)
(e) (f) (g) (h)

Fig. 11: Biped robot fetches can from refrigerator with the same framework as PR2

from annotated world representation, segment and label objects with deep learning technique. We successfully formed the robot system which can deal with daily assistive task even in dynamic environment where location of furniture and objects is changed. We confirmed the feasibility of our approach by performing task in dynamic home environment using a real robot. We also confirmed that our proposed framework enables robot task to be represented independent to robot types through experiment performed the same task by two different robots with the same architecture.

The algorithms, framework, and task descriptions for this paper are implemented and available in public source code repository `jsk_demos`[1].

References

[1] N. Sian, T. Sakaguchi, K. Yokoi, Y. Kawai, and K. Maruyama. Operating humanoid robots in human environments.

[2] K. Yamazaki, R. Ueda, S. Nozawa, M. Kojima, K. Okada, K. Matsumoto, M. Ishikawa, I. Shimoyama, and M. Inaba. Home-assistant robot for an aging society. *Proceedings of the IEEE*, Vol. 100, No. 8, pp. 2429 –2441, aug. 2012.

[3] Y. Furuta, Y. Inagaki, Y. Kakiuchi, K. Okada, and M. Inaba. Tidyup task sequence using pr2 by irt home assistant robot. In *The 31th Annual Conference on Robotics Society of Japan*, pp. 1I2–02, sep 2013.

[4] D. Fox, W. Burgard, and S. Thrun. Markov localization for mobile robots in dynamic environments. Vol. 11, pp. 391–427, 1999.

[5] B. Kuipers and Y. Byun. A robot exploration and mapping strategy based on a semantic hierarchy of spatial representations. *Robotics and autonomous systems*, Vol. 8, No. 1, pp. 47–63, 1991.

[6] K. Okada, M. Kojima, Y. Sagawa, T. Ichino, K. Sato, and M. Inaba. Vision based behavior verification system of humanoid robot for daily environment tasks. In *2006 6th IEEE-RAS International Conference on Humanoid Robots*, pp. 7–12. IEEE, 2006.

[7] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, Vol. 34, No. 11, pp. 2274–2282, 2012.

[8] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, Vol. 60, No. 2, pp. 91–110, 2004.

[9] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.

[11] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1440–1448, 2015.

[12] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pp. 91–99, 2015.

[13] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.

[14] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255. IEEE, 2009.

[15] A. Krizhevsky, I. Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[16] B. Collins, J. Deng, K. Li, and L. Fei-Fei. Towards scalable dataset construction: An active learning approach. In *European Conference on Computer Vision*, pp. 86–98. Springer, 2008.

[17] Y. Bai, K. Yang, W. Yu, C. Xu, W. Ma, and T. Zhao. Automatic image dataset construction from click-through logs using deep neural network. In *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 441–450. ACM, 2015.

[18] H. He and E.A. Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, Vol. 21, No. 9, pp. 1263–1284, 2009.

[19] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[20] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[21] V Badrinarayanan, A Kendall, and R Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *CoRR*, Vol. abs/1511.00561, , 2015.

[22] R Pascanu, T Mikolov, and Y Bengio. Understanding the exploding gradient problem. *CoRR*, Vol. abs/1211.5063, , 2012.

[23] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, Vol. abs/1412.6980, , 2014.

[24] K. Okada, Y. Kakiuchi, H. Azuma, H. Mikita, K. Murase, and M. Inaba. Task compiler: Transferring high-level task description to behavior state machine with failure recovery mechanism. In *ICRA Workshop on Combining Task and Motion Planning*, 2013.

[1] `https://github.com/jsk-ros-pkg/jsk_demos`