https://doi.org/10.20965/jrm.2020.p0245

Paper:

Semi-Automatic Dataset Generation for Object Detection and Recognition and its Evaluation on Domestic Service Robots

Yutaro Ishida and Hakaru Tamukoh

Kyushu Institute of Technology
2-4 Hibikino, Wakamatsu-ku, Kitakyushu, Fukuoka 808-0196, Japan
E-mail: ishida.yutaro954@mail.kyutech.jp, tamukoh@brain.kyutech.ac.jp
[Received September 28, 2019; accepted November 18, 2019]

This paper proposes a method for the semi-automatic generation of a dataset for deep neural networks to perform end-to-end object detection and classification from images, which is expected to be applied to domestic service robots. In the proposed method, the background image of the floor or furniture is first captured. Subsequently, objects are captured from various viewpoints. Then, the background image and the object images are composited by the system (software) to generate images of the virtual scenes expected to be encountered by the robot. At this point, the annotation files, which will be used as teaching signals by the deep neural network, are automatically generated, as the region and category of the object composited with the background image are known. This reduces the human workload for dataset generation. Experiment results showed that the proposed method reduced the time taken to generate a data unit from 167 s, when performed manually, to 0.58 s, i.e., by a factor of approximately 1/287. The dataset generated using the proposed method was used to train a deep neural network, which was then applied to a domestic service robot for evaluation. The robot was entered into the World Robot Challenge, in which, out of ten trials, it succeeded in touching the target object eight times and grasping it four times.

Keywords: domestic service robot, object detection and classification, dataset generation, RoboCup@Home, World Robot Challenge

1. Introduction

In recent years, due to social issues such as the declining birthrate and growing proportion of elderly people, the realization of domestic service robots has been anticipated worldwide. Domestic service robots coexist with humans in homes and public spaces and assist them. For instance, it is anticipated that they can tidy up rooms or work as waiters in restaurants [1–3]. Domestic service robots must have the functions of recognition, judgment, and control to fulfill such roles. The recognition function involves receiving spoken instructions given over a microphone, or detecting and identifying objects from camera images [4–6]. Judgment involves determination of the proper action, and control involves operating actuators.

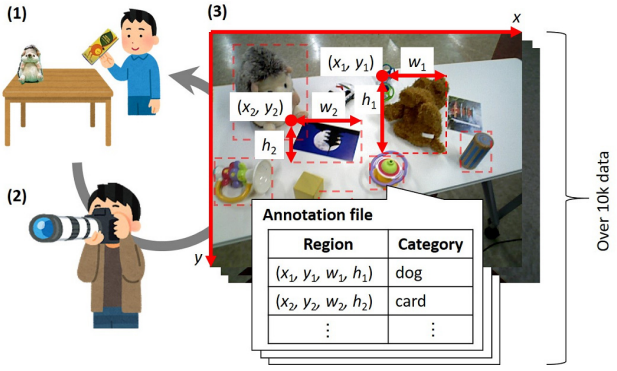


Fig. 1. Procedure of manual dataset generation.

Of these functions, recognition has been achieved with high accuracy by using deep neural networks [7]. In particular, there are high expectations that deep neural networks for general object recognition, as represented by You Only Look Once (YOLO) [8] and Single-Shot MultiBox Detector (SSD) [9], which perform end-to-end object detection and classification from images, can be applied to domestic service robots. In this context, object detection involves displaying a bounding box (BB) to indicate the region where an object exists in an image, and classification involves determining the category of the object in the BB, or in other words, identifying the object. However, the categories recognized by domestic service robots, such as daily commodities, miscellaneous goods, and household implements, differ from those used in general object recognition [10, 11] and depend on the environment in which the robot works, such as homes or public spaces. Therefore, it is necessary to retrain the deep neural network using transfer learning according to the specific environment, which requires the generation of new datasets. The dataset is generated manually according to the procedure shown in **Fig. 1**. The procedure is described using the numbers given in **Fig. 1**. (1) A setting such as the floor or furniture within an environment is randomly selected, a random number of objects are selected, and then they are placed at random positions

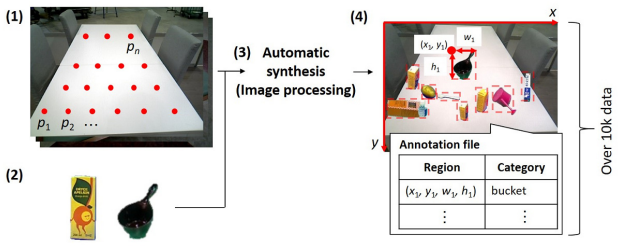


Fig. 2. Overview of proposed system.

in random poses. (2) The setting including the objects is photographed from randomly selected camera angles. (3) For the captured image, annotation files consisting of the regions (reference coordinate (x,y) and horizontal and vertical sizes (w,h)) and categories of the objects are generated to train the deep neural network. Then, the procedure from (1) to (3) is repeated several tens of thousands of times to generate big data [12]. As this operation requires many steps and the selection of many parameters (location and type, number, position, pose, and capture angle of objects), considering that the number and type of objects vary from day to day in daily life, it is unrealistic to perform it manually for specific environments on a continual basis.

Therefore, this paper proposes a method for the semi-automatic generation of a dataset for object detection and classification by domestic service robots. The proposed method is shown in **Fig. 2**, where the data flow is indicated by arrows. The procedure to generate the dataset is described using the numbers in **Fig. 2**. (1) Various locations in the environment are manually photographed as background images. The points p where objects are to be placed are selected. (2) Each object is manually photographed from various angles. (3) The system randomly selects a background image and several object images. Through image processing, the object images are composited with the background image by positioning them at different points p. (4) Annotation files indicating the regions (reference coordinate (x,y) and horizontal and vertical sizes (w,h)) and categories of the objects are generated for the composite image of the virtual scene. These files can be automatically generated, as the regions and categories of the composited objects are known by the system.

Steps (3) and (4) are then repeated several tens of thousands of times by the system to produce big data. Thus, the procedures (1)–(3) in **Fig. 1**, which previously had to be repeated manually, are instead performed by the system in steps (3) and (4) in **Fig. 2**; consequently, the human workload is reduced.

This study compares the generation of a dataset through manual means and the proposed method. The results show that the time required to produce a data unit was reduced from 167 s to 0.58 s, i.e., by a factor of approximately $1/287$. The dataset generated using the proposed method was used to train a deep neural network, which was then applied to a domestic service robot for evaluation. The robot was entered into the World Robot Challenge (Partner Robot Challenge, Service Robotics Category, WRC), where, in ten trials, it could successfully make contact with the target object eight times and grasp and lift the object four times. The contributions of the present study are as follows:

Fig. 3. Domestic service robot, "Toyota HSR."

- The proposed method makes it possible to reduce the human workload required to generate the dataset used for object detection and classification by the domestic service robot.
- The dataset generated using the proposed method was used to train a deep neural network, which was then applied to a domestic service robot for evaluation.

2. Related Studies

2.1. Domestic Service Robots

2.1.1. Configuration and Characteristics of Robot

Figure 3 shows the configuration of the domestic service robot Toyota HSR [1] used in this study. A domestic service robot is required to work in the home environment or public spaces alongside humans. In other words, it must be able to manipulate daily commodities, miscellaneous goods, and household implements in an environment inhabited by humans. Therefore, the robot configuration simulates the five senses and motor functions of humans. Specifically, it carries a microphone, which corresponds to ears, an RGB-D camera and laser range finder, which correspond to eyes, a manipulator, which corresponds to the arm and hand, and a mobile platform, which corresponds to legs. The domestic service robot is characterized by its ability to assume various view angles to execute tasks according to the environment. The end effector of the robot has a working range of 0.0–1.375 m from the floor to enable the robot to manipulate objects on the floor or furniture. The camera is capable of pan/tilt and vertical movement with a range of 1.0–1.3 m from the

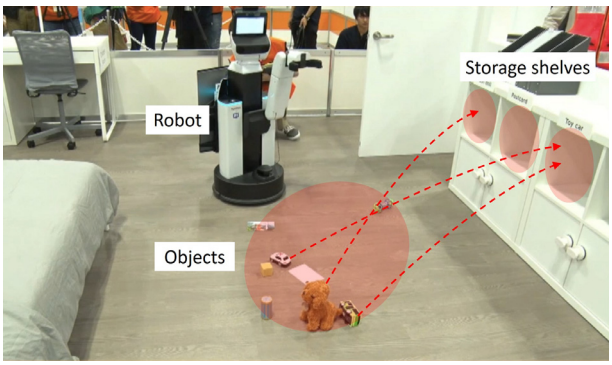


Fig. 4. Arena of WRC.

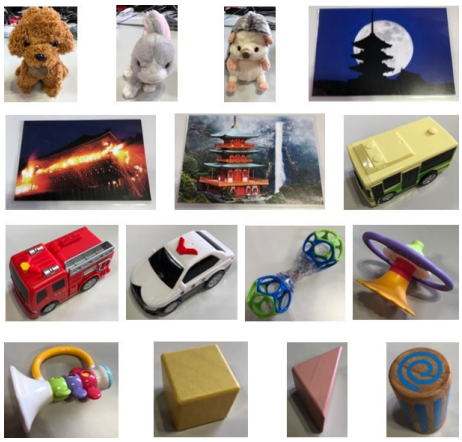


Fig. 5. Fifteen objects in WRC.

floor to secure a wide range of vision for accommodating the working range of the end effector, as shown in **Fig. 3**. These capabilities allow the robot, for instance, to inspect the floor or a shelf of height 1.0 m or more to perform a task. In addition, the robot is equipped with an RGB-D camera capable of distance and three-dimensional measurements in addition to the capture of color images, as an object's accurate three-dimensional coordinates are required to manipulate it. Other robots with similar hardware configurations include "Exi@" [13], developed by the present authors, Fetch Mobile Manipulator [14], and PAL Robotics TIAGo [15].

2.1.2. Benchmark Test

RoboCup@Home [16, 17] and WRC [18] serve as benchmark tests of domestic service robots. In this section, we describe the WRC competition, Tidy Up (Stage 1), which was used for evaluation in this study. WRC is an international robotic competition aimed at the realization of domestic service robots that can provide various assist functions in homes. According to the rules [a] of Tidy Up (Stage 1), the robot first autonomously travels to and enters the competition arena (children's room) shown in **Fig. 4**. Subsequently, the robot must find ten objects, from the 15 toys shown in **Fig. 5**, scattered around the floor, grasp each one with the manipulator, and carry it to a storage shelf. As shown in **Fig. 4**, there are multiple storage shelves, each of which is designated for a different toy category, as indicated by the arrows in the figure. The shelves are assigned their respective categories before the competition, and the position and pose in which the object is placed in the shelf do not matter as long as the object is deposited. Five points are given if the toy is placed in the correct storage shelf, and three points if placed in the wrong shelf. No points are given if the robot fails to place the toy in any storage shelf. Each robot is given 12 min, and the robots are ranked according to the points scored.

Under these rules, the robot must first capture the position of the toy to determine the control target of the manipulator. Subsequently, it must identify the category of the toy to determine the correct storage location. In other words, the robot must be capable of object detection and classification to carry the toy to the storage shelf so that the score is determined largely by the performance of these functions.

There is no standard environment in Tidy Up (Stage 1). The only stipulation is that the competition arena simulates a certain setting in a home environment, and the competitors (and their robots) are not notified of the specific environment until the date of competition. Moreover, they are not notified regarding the specific objects (toys) used. Therefore, participants must employ a deep neural network that can be trained quickly so that the robot can detect and classify objects in an unknown environment, similar to the environments in which future domestic service robots must operate.

2.2. Generation of Dataset

In this section, we provide an outline of the methods for generating datasets reported in previous studies and summarize the related issues. The first method generates datasets manually, as described in Section 1. A single cycle, consisting of (1) to (3) in **Fig. 1**, requires several minutes so that its repetition for several tens of thousands of times to produce a dataset will require a time period ranging from a week to a month.

In the method proposed by Georgakis et al. [19], image synthesis technology is used by the system to generate datasets, thus reducing the human workload. Datasets are produced using several image superimposition strategies and then used to train the deep neural network; the respective datasets are evaluated in terms of the mean average precision (mAP), which is a performance index of object detection and classification. The processing flow of the strategy that attained the highest score (selective positioning-blending-selective scale, SP-BL-SS) is described. (1) Using the GMU Kitchen Scenes dataset [20] and Washington RGB-D Scenes v2 dataset [21], both of which provide RGB-depth images, as the background, the floor on which objects are placed and its parallel planes are extracted using segmentation [22] and plane detection [23]. (2) Cropped object images are created using GraphCut [24] from the BigBIRD [25] and Washington

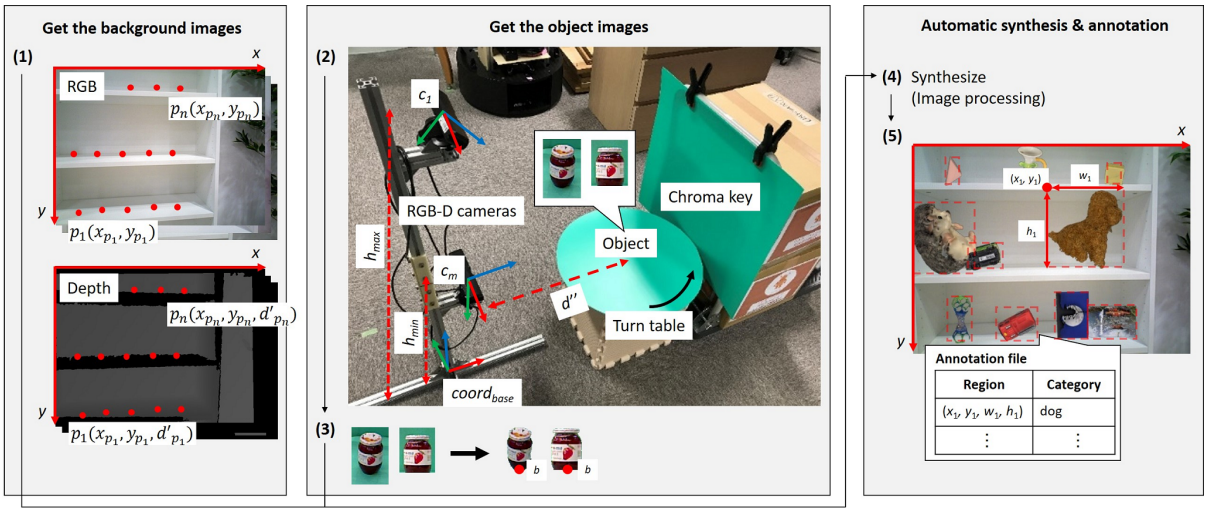


Fig. 6. Proposed system; semi-automatic dataset generation for object detection/recognition on domestic service robots.

RGB-D v1 datasets [26]. (3) The points on the extracted planes for compositing the object are determined, the distance information of these points is acquired from the background depth image, and the objects are scaled to their correct sizes according to their distances. (4) The object is composited with the background using fast seamless cloning [27]. (5) The annotation files are automatically generated simultaneously, as the region and category of the composited objects are known.

When the deep neural network (SSD) was trained using the dataset generated through this method (synthetic data), the mAP of 33.5 was obtained. When a manually generated dataset (real data) was used for training, the mAP was 65.6. When real data amounting to 10% of the synthetic data were added for training, the mAP was 71.6, which was higher than that obtained using only real data. A limitation of this method is that images provided by existing datasets are used for both the background and objects so that the paper contains no discussion regarding how the images are captured. Therefore, the viewpoint of the robot is not considered, and it is assumed that the robot can visualize the surface (and objects) in the background scene. Furthermore, the deep neural network trained with the synthetic data is not applied to a robot for evaluation, and there is no discussion of its application to domestic service robots.

In the present study, we systematized dataset generation in the same manner as the method proposed by Georgakis et al. to reduce the human workload. We modify the method proposed by Georgakis et al., which describes how to composite objects with the background, to take advantage of the various viewpoints of the domestic service robot, as mentioned in Section 2.1.1. We thus propose a method for dataset generation considering its application to domestic service robots and evaluate its application to an actual robot.

3. Proposed Method

3.1. Semi-Automatic Dataset Generation for Object Detection and Classification

In this study, we propose a method for the semi-automatic generation of a dataset for object detection and classification by a domestic service robot. In this method, it is necessary to determine in advance the parameters related to the pose of the robot when it detects and identifies objects, namely, the RGB-D camera's distance d from the object, its height h from the floor, tilt angle θ, and their respective ranges $d_{\min}$–$d_{\max}$, $h_{\min}$–$h_{\max}$, and $\theta_{\min}$–$\theta_{\max}$. The proposed method is outlined in **Fig. 6**. The procedure of dataset generation is described using the numbers in **Fig. 6**.

(1) First, the RGB-D camera is used to capture manually the background images (RGB and depth images) of various settings in the environment. The RGB-D camera is set up so that its distance from the objects lies in the range from $d_{\min}$ to $d_{\max}$, its height from the floor is in the range from $h_{\min}$ to $h_{\max}$, and its tilt angle is in the range from $\theta_{\min}$ to $\theta_{\max}$. Subsequently, the points p at which objects are to be placed are selected in the background images. This is performed by clicking positions in the RGB image to set up the point coordinates $p(x_p, y_p)$, using the graphical user interface (GUI) implemented by the authors using OpenCV. The distance d'_p of point $p(x_p, y_p)$ is determined by referring to that point in the depth image. When the distance cannot be determined in the depth image, it is directly entered via the keyboard. The RGB image is corrected based on the "gray world hypothesis" [b], which is a color constancy hypothesis.

(2) The objects are photographed using a setup consisting of RGB-D cameras, a turntable, and uniform color background (chroma key). Furthermore, m cameras, denoted c_1–c_m, are used. Camera c_1 is

set at the height $h_{\max}$ from the floor at a tilt angle of $\theta_{\max}$. Camera c_m is set at the height $h_{\min}$ from the floor at a tilt angle of $\theta_{\min}$. The other cameras are set at heights between $h_{\min}$ and $h_{\max}$, and at tilt angles between $\theta_{\min}$ and $\theta_{\max}$. The distance between camera c_m and the object is defined as d'', and all cameras are aligned vertically with camera c_m. The turntable is used to rotate the object so that it can be captured from various directions. The uniform color background is used as a chroma key, making it easy to separate the object from its surrounding in the subsequent image processing. The objects are captured as point clouds, and their camera coordinates are transformed to the $coord_{base}$ coordinate system to allow processing in the same coordinate system. Using the point cloud library (PCL), the plane of the turntable is extracted [23] from the transformed point clouds. The point cloud of the object is obtained by removing areas outside of the turntable and the turntable itself. The RGB image of the cropped object is obtained by using PCL to transform the object region in the point cloud to the object region in the RGB image. The RGB image of the cropped object is corrected based on the gray world hypothesis.

(3) The RGB image of the cropped object is transformed to HSV and then subjected to thresholding to extract the uniform color background, which is filtered out. The center of the base of the object image is denoted b and used in subsequent processing.

(4) The background image and object image are composited using image processing. First, a single background image and l object images are randomly selected. The object images are rotated by affine transformation within the range $\theta'_{\min}$–$\theta'_{\max}$. Then, each object image is randomly assigned to a point p in the background image. The object image is scaled according to the distance from its assigned point p by determining the coefficient k in Eq. (1). Finally, the images are composited by matching points b in the object images with the corresponding points p in the background image.

$$k = \frac{d''}{d'_p} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

(5) Against the obtained composite image, annotation files consisting of the regions (reference coordinates (x,y) and horizontal and vertical sizes (w,h)) and categories of the objects are generated. This is performed automatically, as the system already possesses the composited regions of the objects and their categories.

The system repeats steps (4) and (5) several tens of thousands of times to generate big data.

3.2. Comparison with Related Studies

The proposed method is compared with the related studies mentioned in Section 2.2 to verify its validity. The methods proposed by Georgakis et al. and the present authors, in both of which the system employs the image synthesis technique used in **Fig. 6** (3) and (4) to generate the datasets automatically, considerably reduce the human workload as compared with the manual generation of datasets, shown in **Fig. 1** (1) to (3).

The proposed method is similar to the strategy (SP-BL-SS) employed by Georgakis et al., which attained the highest mAP of the methods in their study. They differ in whether image correction is based on the gray world hypothesis, as in the former case, or fast seamless cloning, as in the latter case. However, we use the gray world hypothesis for image correction because it is used by the software of the domestic service robot in this study for object detection and classification, and the difference between the two methods is insignificant.

A major difference lies in the selection of the points p for object placement, which is performed manually via the GUI in the proposed method; thus, it is possible to composite objects at points on an unseen surface. In other words, it is possible to generate data for a horizontal line-of-sight of the robot, in addition to inspecting viewpoints. Therefore, the deep neural network trained using the dataset generated using the proposed method is more suitable for application to the domestic service robot.

4. Experiment and Discussion

4.1. Evaluation of Performance for Object Detection and Classification

Datasets generated manually and using the proposed method were used to train the deep neural network separately, and then their performances for object detection and classification were evaluated and compared. First, following the procedure described in Section 3, the proposed method was used to generate the dataset (synthetic data) as follows. The parameters used in the proposed method are presented in **Table 1**. (1) The floor and furniture (desk, shelf, chair, etc.) were captured by an RGB-D camera (ASUS Xtion Pro Live) to obtain 306 background images. Points p for object placement were selected using the GUI in each background image. (2) The 15 toys used in WRC, shown in **Fig. 5**, were photographed with the setup shown in **Fig. 6** (2) to obtain object images. (3) The background image and object images were composited, and annotation files were generated simultaneously. This resulted in 15,600 composite images, which were used as the synthetic training data; no synthetic data were used as the test data. **Fig. 7** shows examples of the synthetic data. It took 2.5 h to generate the 15,600 synthetic training images.

Subsequently, a dataset (real data) was generated manually by following the method described in Section 1 for a comparison with the proposed method, as follows. (1) From the 15 toys shown in **Fig. 5**, a few were placed on the floor and furniture (desk, shelf, chair, etc.) at random locations and poses and then photographed. (2) Annota-

Table 1. Parameters for experiments.

Items	Values
$d_{\min}$	1.0 m
$d_{\max}$	2.5 m
$h_{\min}$	1.0 m
$h_{\max}$	1.3 m
$\theta_{\min}$	0.0°
$\theta_{\max}$	45.0°
m	2 cameras
d''	1.0 m
l	10 images
$\theta'_{\min}$	−10.0°
$\theta'_{\max}$	10.0°

Fig. 7. Samples of synthesized images.

tion files were generated for each of the captured images. This resulted in 108 real data images, of which 54 were used as real training data, and the remaining 54 as real test data. It took 2.5 h to obtain the 54 real training images, which is the same time as that required to obtain the synthetic training data.

The synthetic and real training data were separately used to train YOLO v2 [8], which is a deep neural network that performs end-to-end object detection and classification from images. YOLO v2 employs model parameters trained on ImageNet [28] and COCO datasets [10] up to the 23rd layer, whereas the remaining layers are obtained via transfer learning. Ten thousand epochs were executed for training on a PC (Intel Core i7-8700K, DDR4 32 GB, nVIDIA GTX 1080) using Darknet. The object detection/classification performance in each case was evaluated by using the real test data as the ground truth.

The experiment results of object detection and classification obtained using synthetic data are shown in **Fig. 8**. The results obtained using synthetic data and real data are compared in **Table 2**. The synthetic data yielded an mAP lower than that obtained using real data by approximately two points. The difference in mAP between synthetic and real data is lower than that obtained using the method proposed by Georgakis et al., although different conditions were used for their evaluation. Furthermore, the time required to generate a single unit of synthetic data was lower than that required for real data by a factor of approximately 287, i.e., 0.58 s against 167 s. The proposed method, which reduces the human workload for dataset generation, is thus valid, particularly when considering its application to domestic service robots.

Fig. 8. Results of object detection/recognition.

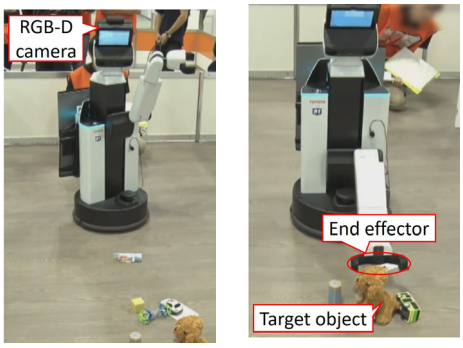
4.2. Application to Domestic Service Robot and Evaluation

The synthetic training data, described in Section 4.1, were used to train the deep neural network, which was then implemented on the Toyota HSR, whose performance in the WRC Tidy Up (Stage 1), described in Section 2.1.2, was evaluated as follows. First, the robot detects and classifies objects on the floor using the RGB-D camera tilted at 45°, as shown in **Fig. 9(a)**. Subsequently, the robot announces which object it will attempt to manipulate (grasp) and controls the end effector so that it reaches the target object, as shown in **Fig. 9(b)**. Videos [c, d] of the motion of the robot were used to determine whether the end effector had come into contact with the object, as shown in **Fig. 9(b)**. Subsequently, the robot controls its end effector to lift the target object. The aforementioned videos were also used to determine whether the target object was completed lifted off the floor. The target coordinates used by the robot to grasp the object were obtained by averaging the three-dimensional coordinates of pixels in the object detection region, from the values in the depth image that corresponds to the object detection image (RGB image) and the camera parameters, as shown in **Fig. 10**.

Figure 11 shows the judgment results regarding whether the end effector came into contact with the target object. The "target object" is the object announced by the robot, whereas the "touched object" is the one touched by the robot. Of the ten trials, the robot succeeded in touch-

Table 2. Comparison of dataset generating/making performance.

Evaluation index		Proposed method	Manpower
Number of data		15600	54
Work time	per 1 object [min/object]	4 to 12	N/A
	per 1 data [s/data]	0.58	166.67
	per 1 train dataset [h/dataset]	2.5	2.5
mAP		64.77	66.69



(a) (b)

Fig. 9. Robot's behavior of grasping. (a) Tilt the camera to 45° to recognize objects. (b) Extend the end effector to the target object.

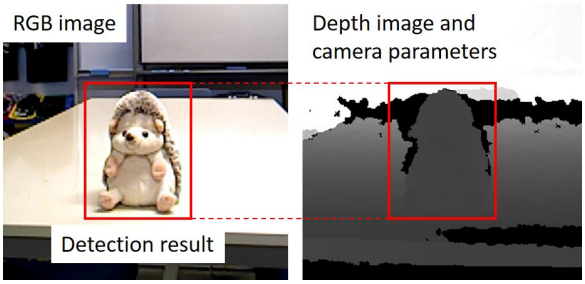


Fig. 10. Estimate the object three dimensional position in the camera coordinate using detection result, depth image and camera parameters.

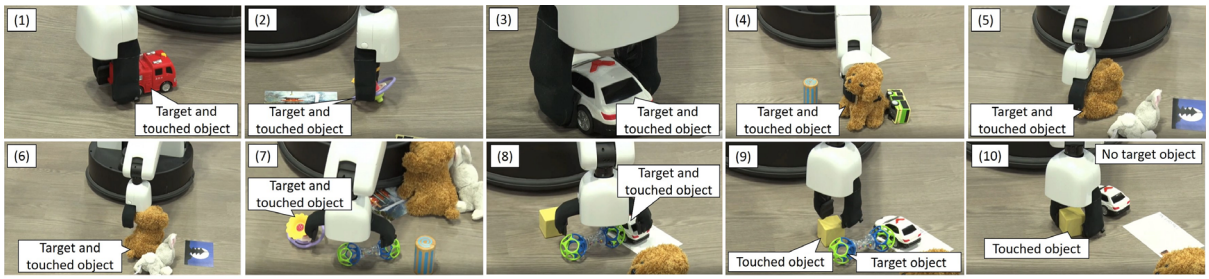


Fig. 11. Images of robot's movement when touching an object.

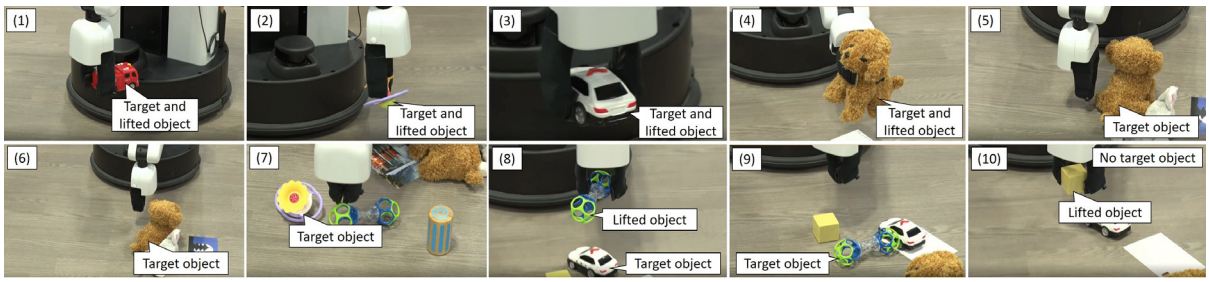


Fig. 12. Images of robot's movement when lifting an object.

ing the target object in eight trials, shown in **Fig. 11** (1)–(8). In **Fig. 11** (9), the robot touched an object adjacent to the target object, and thus failed. In **Fig. 11** (10), the robot touched an object in an area where the target object did not exist and thus failed. **Fig. 12** shows the judgment results regarding whether the target object was lifted off the floor. The "target object" is the same as that described above, whereas the "lifted object" is the object lifted off the floor by the robot. In the same ten trials described above, the robot succeeded in lifting the target object off the floor in four trials, shown in **Fig. 12** (1)–(4). In **Fig. 12** (5)–(8), the robot touched the target object but

failed to lift it. In **Fig. 12** (9) and (10), the robot failed to lift the target object because it failed to touch it.

In summary, the robot could touch the target object eight times and lift it four times during the ten trials. Consequently, it scored 15 points in the WRC event, which was the highest score among all participating teams.

5. Conclusions

Herein, we proposed a method for the semi-automatic generation of a dataset for object detection and classification by a domestic service robot. The experiment results showed that the proposed method could reduce the time required to generate a data unit from 167 s, when performed manually, to 0.58 s, i.e., reduced by approximately $1/287$th. The dataset generated using the proposed method was used to train a deep neural network applied to a domestic service robot. The robot succeeded in touching the target object eight times and grasping it four times out of ten trials.

An issue to be addressed in the future is that the success rate for manipulating objects must be improved when applying the proposed method to domestic service robots. To this end, it is necessary to investigate whether the object region was correctly estimated in cases when the robot partially touched the object (**Fig. 11** (7) and (8)) and when it failed to touch the object (**Fig. 11** (9) and (10)). Furthermore, with regard to the case where the robot touched the object but failed to lift it (**Fig. 12** (5) and (6)), it will be necessary to generate a dataset obtained by training with the points to be grasped by the robot in addition to only the object region.

Acknowledgements

This study received funding from the New Energy and Industrial Technology Development Organization (NEDO) and a Grant-in-Aid for Scientific Research (Project No.17H01798) from the Japan Society for the Promotion of Science (JSPS).

References:

[1] T. Yamamoto, K. Terada, A. Ochiai, F. Saito, Y. Asahara, and K. Murase, "Development of Human Support Robot as the research platform of a domestic mobile manipulator," ROBOMECH J., Vol.6, 4, 2019.

[2] Y. Nakagawa and N. Nakagawa, "Relationship Between Human and Robot in Nonverbal Communication," J. Adv. Comput. Intell. Intell. Inform., Vol.21, No.1, pp. 20-24, 2017.

[3] J. Cai and T. Matsumaru, "Human Detecting and Following Mobile Robot Using a Laser Range Sensor," J. Robot. Mechatron., Vol.26, No.6, pp. 718-734, 2014.

[4] M. Tanaka, H. Matsubara, and T. Morie, "Human Detection and Face Recognition Using 3D Structure of Head and Face Surfaces Detected by RGB-D Sensor," J. Robot. Mechatron., Vol.27, No.6, pp. 691-697, 2015.

[5] M. Hashimoto, Y. Domae, and S. Kaneko, "Current Status and Future Trends on Robot Vision Technology," J. Robot. Mechatron., Vol.29, No.2, pp. 275-286, 2017.

[6] Z. Chai and T. Matsumaru, "ORB-SHOT SLAM: Trajectory Correction by 3D Loop Closing Based on Bag-of-Visual-Words (BoVW) Model for RGB-D Visual SLAM," J. Robot. Mechatron., Vol.29, No.2, pp. 365-380, 2017.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Proc. of the 25th Int. Conf. on Neural Information Processing Systems, Vol.1, pp. 1097-1105, 2012.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 779-788, 2016.

[9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector," European Conf. on Computer Vision, pp. 21-37, 2016.

[10] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," Proc. of European Conf. on Computer Vision, pp. 740-755, 2014.

[11] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," Int. J. of Computer Vision, Vol.88, pp. 303-338, 2010.

[12] J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. M. A. Patwary, Y. Yang, and Y. Zhou, "Deep Learning Scaling is Predictable, Empirically," arXiv:1712.00409, 2017.

[13] S. Hori, I. Yutaro, Y. Kiyama, Y. Tanaka, Y. Kuroda, M. Hisano, Y. Imamura, T. Himaki, Y. Yoshimoto, Y. Aratani, K. Hashimoto, G. Iwamoto, H. Fujita, T. Morie, and H. Tamukoh, "Hibikino-Musashi@Home 2017 Team Description Paper," arXiv:1711.05457, 2017.

[14] M. Wise, M. Ferguson, D. King, E. Diehr, and D. Dymesich, "Fetch and freight: Standard platforms for service robot applications," Proc. of Workshop on Autonomous Mobile Service Robots, 2016.

[15] P. Jordi, M. Luca, and F. Francesco, "TIAGo: the modular robot that adapts to different research needs," Proc. of Int. Workshop on Robot Modularity, 2016.

[16] T. Wisspeintner, T. van der Zant, L. Iocchi, and S. Schiffer, "RoboCup Home: Scientific Competition and Benchmarking for Domestic Service Robots," Interaction Studies, pp. 392-426, 2009.

[17] L. Iocchi, D. Holz, J. Ruiz-del-Solar, K. Sugiura, and T. van der Zant, "Analysis and results of evolving competitions for domestic and service robots," Artificial Intelligence, pp. 258-281, 2015.

[18] H. Okada, T. Inamura, and K. Wada, "What competitions were conducted in the service categories of the World Robot Summit?," Advanced Robotics, Vol.33, No.17, pp. 900-910, 2019.

[19] G. Georgakis, A. Mousavian, A. C. Berg, and J. Kosecka, "Synthesizing Training Data for Object Detection in Indoor Scenes," arXiv:1702.07836, 2017.

[20] G. Georgakis, M. A. Reza, A. Mousavian, P. Le, and J. Košecká, "Multiview RGB-D dataset for object instance detection," Proc. of 2016 4th Int. Conf. on 3D Vision (3DV), pp. 426-434, 2016.

[21] K. Lai, L. Bo, and D. Fox, "Unsupervised feature learning for 3D scene labeling," Proc. of IEEE Int. Conf. on Robotics and Automation, pp. 3050-3057, 2014.

[22] A. Mousavian, H. Pirsiavash, and J. Kosecka, "Joint semantic segmentation and depth estimation with deep convolutional networks," Proc. of 2016 4th Int. Conf. on 3D Vision (3DV), pp. 611-619, 2016.

[23] C. Taylor and A. Cowley, "Parsing indoor scenes using RGB-D imagery," Robotics: Science and Systems, 2012.

[24] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.23, Issue 11, pp. 1222-1239, 2001.

[25] A. Singh, J. Sha, K. Narayan, T. Achim, and P. Abbeel, "A large-scale 3D database of object instances," Proc. of IEEE Int. Conf. on Robotics and Automation, pp. 509-516, 2014.

[26] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view RGB-D object dataset," Proc. of IEEE Int. Conf. on Robotics and Automation, pp. 1817-1824, 2011.

[27] M. Tanaka, R. Kamio, and M. Okutomi, "Seamless image cloning by a closed form solution of a modified Poisson problem," Proc. of Special Interest Group on Computer GRAPHics Asia, 15, 2012.

[28] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," Proc. of IEEE Conf. Computer Vision and Pattern Recognition, pp. 248-255, 2009.

Supporting Online Materials:

[a] World Robot Challenge Partner Robot Challenge Real Space Rules & Regulations. https://worldrobotsummit.org/download/rulebook-en/rulebook-Partner_Robot_Challenge.pdf [Accessed October 25, 2019]

[b] colorcorrect. https://github.com/shunsukeaihara/colorcorrect [Accessed October 25, 2019]

[c] World Robot Challenge Partner Robot Challenge Real Space Day 1 Video. https://youtu.be/a8dQi-NLtfE?t=25886 [Accessed October 25, 2019]

[d] World Robot Challenge Partner Robot Challenge Real Space Day 2 Video. https://youtu.be/5qpdPbMePXE?t=25595 [Accessed October 25, 2019]

Name:
Yutaro Ishida

Affiliation:
Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology

Address:
2-4 Hibikino, Wakamatsu-ku, Kitakyushu, Fukuoka 808-0196, Japan

Brief Biographical History:
2015 Received B.E. from Kyushu Institute of Technology
2017 Received M.E. from Kyushu Institute of Technology

Main Works:
- Service robots
- Hardware/software complex system

Membership in Academic Societies:
- The Institute of Electrical and Electronics Engineers (IEEE)
- The Robotics Society of Japan (RSJ)
- The Institute of Systems, Control and Information Engineers (ISCIE)
- The Institute of Electronics, Information and Communication Engineers (IEICE)

Name:
Hakaru Tamukoh

Affiliation:
Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology

Address:
2-4 Hibikino, Wakamatsu-ku, Kitakyushu, Fukuoka 808-0196, Japan

Brief Biographical History:
2001 Received B.E. degree from Miyazaki University
2003 Received M.E. degree from Kyushu Institute of Technology
2006 Received Ph.D. from Kyushu Institute of Technology
2006- Postdoctoral Research Fellow of 21st Century Center of Excellent Program, Kyushu Institute of Technology
2007- Assistant Professor, Tokyo University of Agriculture and Technology
2013- Associate Professor, Kyushu Institute of Technology

Main Works:
- Hardware/software complex system
- Digital hardware design
- Neural networks
- Soft-computing
- Home service robots

Membership in Academic Societies:
- The Institute of Electronics, Information and Communication Engineers (IEICE)
- Japan Society for Fuzzy Theory and Intelligent Informatics (SOFT)
- Japanese Neural Network Society (JNNS)
- The Institute of Electrical and Electronics Engineers (IEEE)
- The Japanese Society for Artificial Intelligence (JSAI)
- The Robotics Society of Japan (RSJ)