Proceedings of the 2016 IEEE/SICE International Symposium on System Integration, Sapporo Convention Center, Sapporo, Japan, December 13-15, 2016

Automatic Houseware Registration System for Informationally-Structured Environment

Kazuto Nakashima[1], Julien Girard[2], Yumi Iwashita[3] and Ryo Kurazume[4]

Abstract— **To provide daily-life assistance appropriately by a service robot, the management of houseware's information in a room or a house is an indispensable function. Especially, the information about what and where objects are in the environment are fundamental and critical knowledge. We can track housewares with high reliability by attaching markers such as RFID tags to them, however, markerless housewares management system is still useful since it is easy-to-use and low cost. In this work, we present an object management system using an egocentric vision and a region-based convolutional neural network (R-CNN) to automatically detect and register housewares. The proposed system consists of smart glasses equipped with a wearable camera, a cloud database which manages object information, and a processing server for detecting and registering housewares to the cloud database. We perform two experiments. First, we train the R-CNN on a newly-constructed dataset to detect various housewares and configure a houseware-specific detector. All systems are composed of ROS packages. Second, we conduct experiments for automatic housewares registration using the proposed system. We demonstrate that the proposed system can detect, recognize, and register housewares approximately in real time.**

I. Introduction

In recent years, the rapid aging of the population has caused serious problems such as a labor shortage in hospitals or care facilities. To mitigate this problem, the development of a service robot coexisting with a human in a daily-life environment is an urgent challenge. We have been focusing on the concept of a context-aware intelligent space, which is so-called Informationally-Structured Environment (ISE), and developing its architecture named ROS-TMS[1][2]. The ISE is an intelligent space where various sensors are distributed and combined, organizing a sensor network within the environment. According to the main idea of the ISE, the ROS-TMS not only utilizes the sensors mounted on the robot but also embeds the various sensors such as laser range finder or cameras in the environment. Thus, the service robot can obtain information of an extensive scope and carry out various tasks without being limited by own performance, by using all the sensors and resources the ISE provides. Furthermore, it models an integrated system modularizing all of functions and processes. Executing the modularized processes and interprocess communication is based on the Robot Operating System (ROS)[3] framework.

In the field of developing service robots, the most prospective service is a fetch-and-carry task, which is comparatively easy to design. To achieve it, a service robot needs to know where the target object is or which state it is. It is also necessary to measure position and status of objects within the symbiosis space on a regular basis. Current ROS-TMS manages houseware information by attaching some identifiable markers such as an RFID tag on them, as we describe in Section II-A. This approach is taken in the research fields of ubiquitous computing and the Internet of Things (IoT). Although these systems are able to reliably track the objects, they may not completely recognize the status of the environment if the sensors can only measure particular objects, if its measurement range is limited, or if a new object is added. One solution to make the measurement of scattered housewares more flexible is to use the technique of markerless object tracking. However, it still requires high-resolution images and high-precision object detection technique. Nakayama et al.[4] developed a smart goggles system which recognizes objects appearing in the daily-life scene in real-time and enables to improve a visual memory. Owing to the goggles equipped with a camera, rich information of the object that the user is gazing and has interest can be obtained from close position. Their object recognition technique is based on a hand-designed image feature. In recent years, an effectiveness of the feature automatically learned from massive volumes of data has been shown in the various field of real-world perception, which is based on the well-known machine learning technique Deep Learning. Furthermore, that can not only process in real-time on the parallel computing environment, but also can be useful to extend the scale of object recognition and improve the runtime performance.

In this paper, we propose an automatic houseware registration system using smart glasses and the latest object detection technique based on the Deep Learning, which is introduced in Section III.

Our contributions can be summarized as follows:

- We train the houseware-specific detector based on the latest object detection technique Faster R-CNN[5] and modularize it with ROS messaging framework.
- We develop an automatic houseware registration system using smart glasses, our cloud database, and the houseware-specific object detection technique.

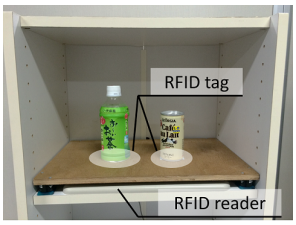
[1]Kazuto Nakashima is with the Graduate School of Information Science and Electrical Engineering, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka, Japan. k_nakashima@irvs.ait.kyushu-u.ac.jp

[2]Julien Girard is with the ENSTA ParisTech, 1024, Boulevard des Marechaux, 91762 Palaiseau Cedex, France. julien.girard@ensta-paristech.fr

[3]Yumi Iwashita is with the Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, USA. Yumi.Iwashita@jpl.nasa.gov

[4]Ryo Kurazume is with the Faculty of Information Science and Electrical Engineering, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka, Japan. kurazume@ait.kyushu-u.ac.jp

978-1-5090-3328-7/16/$31.00 ©2016 IEEE



(a) Intelligent Cabinet System

(b) Motion Capture System

Fig. 1. Conventional systems to manage housewares in ROS-TMS

- We perform an empirical evaluation and demonstrate that our system can detect housewares and register their information in real time even without attaching any identifiable markers on them.

II. Informationally-Structured Environment

A. ROS-TMS

As mentioned in the previous section, we have been developing an integrated service robot system for Informationally-Structured Environment, named ROS-TMS. Currently, it has over 150 modules useful for a service robot to provide daily-life assistance. Those modules include user interface, task scheduler, robot controller, movement path planner, and a database which manages ambient information. In this section, we firstly explain about existing systems to measure housewares in the environment and a cloud database to manage their information. Secondly, we address the current issues in terms of managing housewares.

1) Existing houseware measurement systems: In current ROS-TMS, there are several systems to measure housewares in a room. Figure 1 (a) shows an exterior of one of the systems, Intelligent Cabinet System (ICS). In the system, an RFID reader and load cells are installed on its storage area and measure the RFID tags attached to stored objects to recognize individually and estimate the positions by analyzing the weight distribution on multiple load cells. Besides, our refrigerator, named Intelligent Refrigerator System (IRS), also has a similar type of this mechanism. Figure 1 (b) shows motion capture camera, Vicon Bonita. In our experiment space, over 20 cameras are embedded to track position and posture of objects by measuring reflective markers attached on them. To recognize an individual object, it needs to have at least three markers on it and register those structure model in advance.

2) Cloud database: We manage the housewares in the environment with the cloud database. Our cloud database is composed of a database (MongoDB) and an interface to read/write data through the wireless network. As well as the ambient information that the embedded sensors capture, it manages every information relating to the symbiosis space including static/dynamic maps and the task information that the robot needs for operation. As for the houseware, its position and name are provided in the database.

B. Problems in managing housewares

In current systems, there are some limitations to manage housewares. First of all, although an RFID-based measurement system such as the ICS can reliably track a stored item, its measurement has several constraints. Housewares should be located within the bounded measurement area of the RFID reader and tags should be attached to them manually. Meanwhile, a motion capture has a wide range of measurement, however, it requires some laborious processes to attach identifiable features on the object to be recognized as well.

For these reasons, we focus on a generic object detection technique to enable markerless tracking that uses the appearance of the item itself. In current ROS-TMS, various cameras are distributed to the human, the robot, and the environment, thus it is possible to detect housewares from various images of the daily life scene and manage them in a wider scope. However, among these cameras, embedded camera has difficulty to detect small-sized items such as housewares, because of the low resolution and the measurement distance. In this project, we take particular note of an egocentric vision from the wearable camera where it is possible to obtain the houseware image with comparably high resolution and comprehend the interests of the inhabitant. In the following section, we introduce a notable system that manages housewares automatically by applying generic object detection to the egocentric vision.

III. Automatic Houseware Registration System

A. System overview

The proposed system consists of three parts as illustrated in Figure 2:

- *Smart glasses* to stream the inhabitant's egocentric vision and enable the interactive services.
- *Processing server* to detect housewares from the streamed image and register their information to the cloud database.
- *Cloud database* to manage the information of an inhabitant, a robot, and housewares in the ROS-TMS.

The procedure of automatic registration is as follows. At first, the smart glasses the inhabitant wears capture an image by its embedded camera and stream them continuously to the network as egocentric vision. The processing server receives them to detect housewares. The detection function is based on the neural network-based object detection technique as we describe in Section III-D. The detected housewares are registered to the existing ROS-TMS database with their category name. Thus the system can automatically manage scattered housewares in the environment.

B. Houseware detection from the egocentric vision

Egocentric vision, also known as first-person vision, can be employed as a clue to recognize human actions or activities, estimate their intentions, and understand the life patterns since the continuously-acquirable vision contains the rich information about interactions with objects in a daily life.

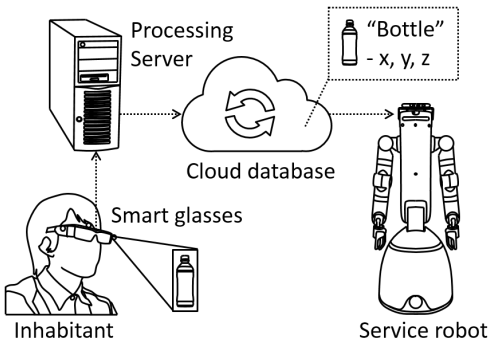


Fig. 2. System overview

Fig. 3. Smartglasses: EPSON Moverio BT-200. The left pane is the exterior. We attached some reflective markers on it to track the position and the posture. The right pane is a manually-reconstructed scene for illustrative purpose.

Egocentric vision can indeed capture contextual information and head motions related to the use of a particular object. There are several works to solve such tasks by using object recognition/detection techniques on egocentric vision. In the activity recognition task, for instance, Pirsiavash et al.[6] focused on observable objects to encode an egocentric video into a representative feature. They employed Deformable Part Model[7] approach for detecting objects. Faith et al.[8] used regions of hands and objects which can be observed in the vision as well, and are recently improved with convolutional neural network (CNN) frameworks[9]. Meanwhile, as for an application, Huang et al.[10] proposed an egocentric interaction system based on the hand/fingertip detection and they utilized the latest object detection model, Faster R-CNN[5]. Owing to the model, they achieved real-time accurate hand detection.

As described before, we focus on the fetch-and-carry task as the most promising service for daily life assistance. Regarding the task, the measurement of housewares which have a possibility to be required by the inhabitant is important to let the robot know their positions promptly. Thus we focus on egocentric vision to tackle the problem. Owing to the fact that the housewares which the inhabitant has interests tend to appear in egocentric vision, high immediacy of the service provision can be expected by detecting objects from the vision and updating our database. Furthermore, a wearable camera can obtain higher-resolution images from the position close to the object than an embedded camera fixed in walls or ceilings.

In recent years, several high-performance wearable cameras are provided in the market and that enables us to capture egocentric vision easily just only by attaching it to the head. Especially, smart glasses are very popular as a human-computer interaction device since it has not only onboard cameras but microphones, speakers, and see-through displays. Besides, it operates on a lightweight OS, so that we can use it as a portable computing device. Thus we use the smart glasses as a egocentric camera, supposing the applications of a user interface in the future.

C. Object detection techniques based on Deep Learning

Since the ImageNet Large Scale Visual Recognition Competition (ILSVRC) 2012 that Krizhevsky et al.[11] won with deep neural networks named AlexNet, convolutional neural networks (CNNs) have gained a great deal of attention as a powerful method to achieve a high accuracy of the image understanding. As the latest result, CNNs have surpassed human's ability to identify images, with achieving an extremely low error rate[12][13].

In addition to image-based classification, the acquired knowledge can be applied to a generic object detection task, taking an approach to input regions of object proposal in an image into a CNN. In the architecture called region-based convolutional neural network (R-CNN)[14], it detects multiple object proposals from an image in advance by applying the segmentation method called Selective Search[15] and predicts a semantic label of the region-of-interest (RoI) by propagating it through a CNN, so that achieves high accuracy object detection with a multiclass classification. More recently, there are some works to speed up the detection time. Fast R-CNN[16] introduced RoI pooling layer which can pool any region in the final convolutional layer which corresponds to object proposal into the fixed-size feature map, so that enables to detect multiple objects by a one-shot propagation. Moreover, Faster R-CNN[5] replaced Selective Search that was a bottleneck at runtime with a neural network called Region Proposal Networks (RPN). Now thus, with a hardware acceleration by a graphics processing unit (GPU), it has become available to detect multiple objects with high accuracy approximately in real time. In the proposed system, we employ the latest architecture Faster R-CNN as a powerful function to understand what object humans are seeing in the daily life scene.

D. Houseware-specific detection network

In the original implementation of the Faster R-CNN[17], some models which are pre-trained on a large-scale dataset are opened to the public and available to utilize while their categories to detect differ from our situation; some of the default categories such as an animal or a vehicle in the PASCAL VOC dataset[18] are unlikely to appear in the room. We train the network with a newly created dataset which contains appropriate classes such as a bottle, a book, and other housewares. Subsequently, we explain about our settings of the network.

1) Network models: Faster R-CNN is composed of two neural network-based modules. One is Region Proposal Network (RPN) that are based on a CNN and produces a set of

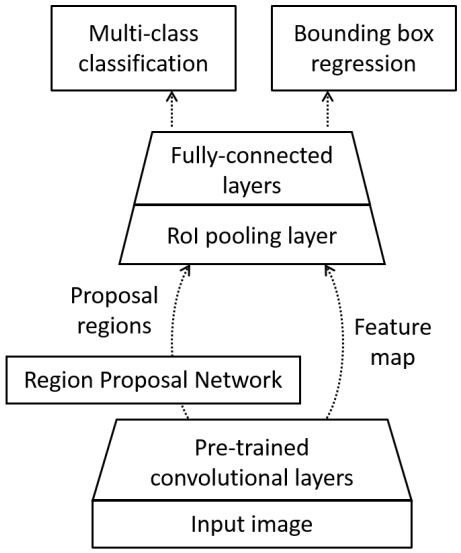


Fig. 4. The architecture of Faster R-CNN[5]

object proposals. The other is the Fast R-CNN detector[16] that are also based on a CNN. It classifies an object in the region proposed by the RPN and simultaneously regresses the region of the detected object (bounding box). Both modules have homogeneous convolutional layers to extract a full-image feature and share their parameters, thus the Faster R-CNN models a unified network as shown in Figure 4. As the sharable convolutional features, we use the ZF model[19] which has 5 convolutional layers and the VGG-16[20] model which has 13 convolutional layers, which are investigated in the paper[5] as the "fast" model and the "deep" model respectively. We compare these models in terms of accuracy and detection speed in runtime.

2) Dataset: We constructed a new dataset consisting of 11 housewares categories, which is a subset of the well-known large-scale image dataset ImageNet[21] providing bounding boxes for thousands of object categories as ground truth. The detail of our dataset is indicated in the Table I. Among the massive available images, we chose the categories with the following two criteria.

- It already exists in our experiment environment.
- The bounding boxes are available as ground truth.

3) Training strategy: Three methods are discussed in the paper[5] to train both RPN and Fast R-CNN, sharing their convolutional layers. We employ the approximate joint training method which makes a time-consuming evaluation easier. This method is not rigorous for ignoring a part of the network response preferred to be treated, however, it works effectively with reducing the training time and keeping the accuracy. Firstly, the sharable convolutional layers are initialized with a model pre-trained on the ImageNet classification set and the entire networks are simultaneously trained on the ImageNet detection set as one network, minimizing the combined loss from both RPN and Fast R-CNN.

TABLE I

THE HOUSEWARE-SPECIFIC DATASET CONTAINING 2659 IMAGES OF 11 HOUSEWARE CATEGORIES. WORDNET ID IS IDENTIFIABLE NUMBER MANAGED IN THE IMAGENET. # REPRESENTS "NUMBER OF".

Category	WordNet ID	# regions	# images
Book	n02870526	133	116
Coffee can	n03062985	213	179
Controller	n03096960	200	179
Cup	n03147509	169	153
Dish	n03206908	128	110
Glass	n03438257	179	149
Kettle	n03612814	205	183
Teapot	n04398044	599	579
Water bottle	n04557648	478	437
Watering pot	n04560292	176	153
Wine bottle	n04591713	556	421
Total		3036	2659

E. Implementation details

1) Modularization by ROS framework: The ROS-TMS employs Robot Operating System (ROS)[3] as an interprocess communication framework. ROS is efficient to configure a large distributed computing system. Furthermore, it is critical for CNNs to utilize GPU acceleration for speeding up the runtime, hence the process possibly occupies much of the available computing resources. In the future of ROS-TMS, launching the multiple object detection processes for each service is inefficient and can be a high load. For these reasons, we modularized the Faster R-CNN operation with ROS messaging functions and made the function sharable to improve a functional reusability and decentralize the load on the large-scale processing. Object detection is running on a decentralized server and can be used by other applications, or in any other generic research purpose. Basically, we extend the original implementation[17] by Python launguage.

2) Android application on the smart glasses: Including ours, most of the smart glasses are an Android-based device. In order to stream the egocentric images into the local network, we build an Android application running on them and even here employ the ROS messaging for packets of the image.

IV. EXPERIMENTS

A. Evaluation of the houseware-specific Faster R-CNN

As described, we evaluate both the ZF model and the VGG-16 model on our houseware dataset. In general, CNNs require much of image resources to optimize their parameters. In this experiment, due to the shortage of our dataset, we evaluate the performance of the houseware-specific Faster R-CNN by 4-folds cross validation. At first, we divide the entire dataset into four subsets and validation is repeated four times with different combinations of a training set and a test set. For each fold, one of the subsets is used as a test set and the other three subsets form a training set. When training on this set we use 70k mini-batch iterations. The evaluation metric is mean Average Precision (mAP) in accordance with

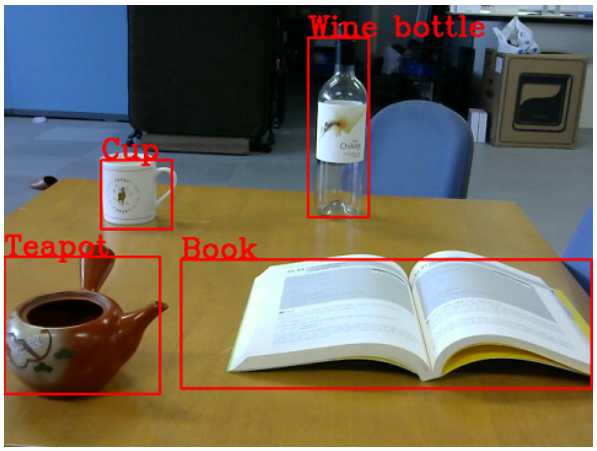


Fig. 5. Example of houseware detection on a web camera. Four housewares are bounded with red boxes and given the classified labels

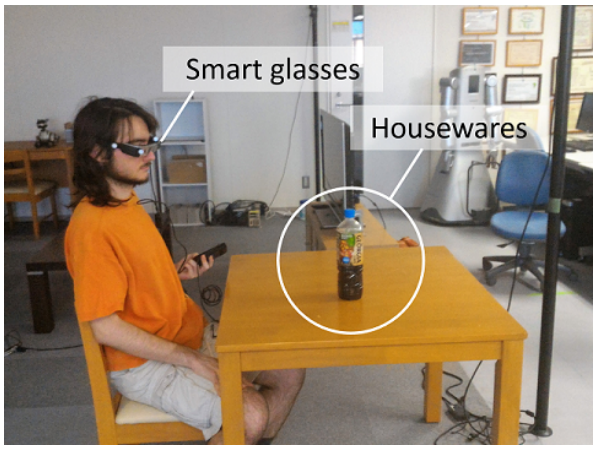


Fig. 6. The setting of Experiment B described in Section IV-B

the PASCAL VOC benchmark[18]. Predicted bounding box is counted as a correct detection when the intersection area with its ground truth exceed 50%. Thus the average accuracy across all folds is computed as a final result. The evaluations are all performed on the NVIDIA GTX Titan X GPU.

1) Performance of the detection: In Table II we show the evaluation results for both VGG16 model and ZF model. The VGG-16 model achieves the higher result for all categories with an mAP of 72.4% and the ZF model has an mAP of 66.3%. Although the way of evaluation is different from the original metric, we can say that the training properly works, by taking into account the fact that the public result on PASCAL VOC 2012 is 70.4% mAP[5]. The "book" and "glass" categories show the low accuracy comparing with the others. One of the reasons is the remarkable changes in object appearance according to the status (e.g., an open or closed book) and the transparency of materials. Therefore we think the accuracy can be improved by increasing the number of training samples. In training for each fold, the VGG-16 and the ZF took 9.6 hours and 4.5 hours respectively. Figure 5 shows an example of a detection result.

B. Automatic registration experiments

Next, we conduct an experiment to automatically register the detected housewares to the ROS-TMS database, with the two models trained in the Experiment A. In this experiment, we confirm the following functions.

- Our system captures the egocentric vision from smart glasses and stream them to the processing server.
- Our system detects multiple housewares within the given image and register the predicted result to the database.

We observe a success rate of registration and its processing speed on the setting of that the smart glasses detect housewares over consecutive frames. As for the details, we set the smart glasses and multiple housewares in advance as the glass can capture all of housewares. Subsequently, our system starts to detect and register housewares from the images captured by the smart glasses. Figure 6 shows the settings of this experiment.

1) Results: As a result, our proposed system could register all of the detected housewares. However, image quality is deteriorated by the movement of the smart glasses and the accuracy of object detection gets worse, for example, if the user changes the face direction quickly. With the VGG-16 model, the proposed system achieved the detection in 0.131 sec and its frame rate is 7.5 fps. Meanwhile, with the ZF model, the system achieved the detection in 0.064 sec and its frame rate is 8.8 fps. The entire processing time of the VGG-16 model is slower than the ZF model, however, the process is operated in a sufficiently high speed for real-time registration. Figure 7 shows some examples that the proposed system detected on the houseware dataset.

V. Conclusions

We developed the automatic houseware registration system using the latest object detection architecture and smart glasses. We conducted the two experiments. In the first, using the public large-scale image set, we trained houseware-specific Faster R-CNN and modularized it as a sharable houseware detection server. In the second, we carried out the registration experiment with the entire system. From those results, we confirmed that our system can detect housewares within the smart glasses' vision and automatically register their information to the database in real time.

The current system has not adopted the houseware identification on the instance level. Furthermore, it manages only the information about a category. Our future works include implementations of two functions. One is to estimate a position of a detected houseware using the posture taken by the position tracker in Figure 3 (a). The other is to identify instances of housewares using the position information or their appearance, that enables the service robot to find an intended houseware from among scattered objects promptly.

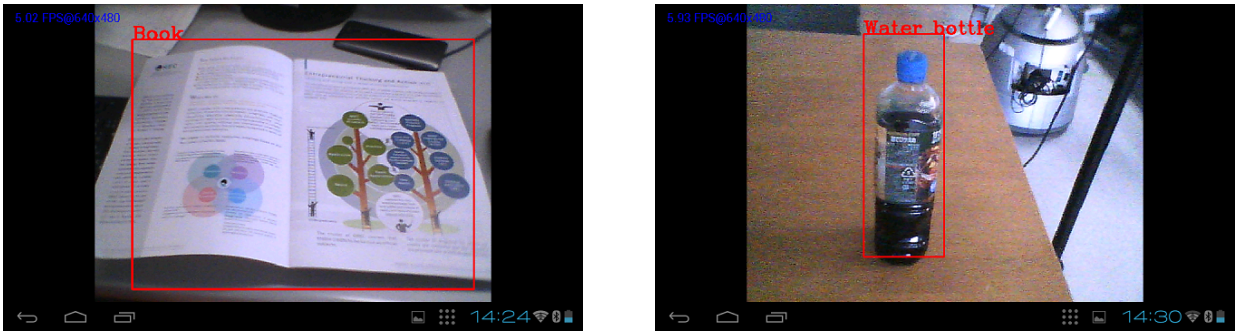
Acknowledgement

This research is supported by The Japan Science and Technology Agency (JST) through its"Center of Innovation

TABLE II

RESULTS IN 4-FOLDS CROSS VALIDATION. MAP DENOTES MEAN AVERAGE PRECISION.

Model	mAP	Book	Coffee can	Controller	Cup	Dish	Glass	Kettle	Teapot	Water bottle	Watering pot	Wine bottle
VGG-16	71.7	50.7	89.9	80.5	64.4	67.4	53.4	67.8	88.8	74.8	67.4	83.3
ZF	66.3	45.2	89.6	77.6	58.8	58.7	46.0	61.9	85.6	67.6	59.8	78.9



(a) Detection of the book

(a) Detection of the water bottle

Fig. 7. Example of screens on the smart glasses which render the result of the housewares detection.

Science and Technology based Radical Innovation and Entrepreneurship Program (COI Program)."

REFERENCES

[1] Y. Pyo, K. Nakashima, S. Kuwahata, R. Kurazume, T. Tsuji, K. Morooka, and T. Hasegawa, "Service robot system with an informationally structured environment," *Robotics and Autonomous Systems*, vol. 74, pp. 148–165, 2015.

[2] https://github.com/irvs/ros_tms.

[3] M. Quigley, K. Conley, B. P. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, "ROS: an open-source robot operating system," in *ICRA Workshop on Open Source Software*, 2009.

[4] H. Nakayama, T. Harada, and Y. Kuniyoshi, "AI goggles: Real-time description and retrieval in the real world with online learning," in *Computer and Robot Vision, 2009. CRV '09. Canadian Conference on*, May 2009, pp. 184–191.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015. [Online]. Available: http://arxiv.org/abs/1506.01497

[6] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2847–2854.

[7] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[8] A. Fathi, A. Farhadi, and J. M. Rehg, "Understanding egocentric activities," in *International Conference on Computer Vision*. IEEE, 2011, pp. 407–414.

[9] Y. Zhou, B. Ni, R. Hong, X. Yang, and Q. Tian, "Cascaded interactional targeting network for egocentric video analysis," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[10] Y. Huang, X. Liu, X. Zhang, and L. Jin, "A pointing gesture based egocentric interaction system: Dataset, approach and application," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 16–23.

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

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1026–1034.

[13] ——, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[15] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[16] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[17] https://github.com/rbgirshick/py-faster-rcnn.

[18] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015.

[19] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*. Springer, 2014, pp. 818–833.

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

[21] J. Deng, W. Dong, R. Socher, L.-J. 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*. IEEE, 2009, pp. 248–255.