Team C^{2}M: Two Cooperative Robots for Picking and Stowing in Amazon Picking Challenge 2016

Hironobu Fujiyoshi[1], Takayoshi Yamashita[1], Yuji Yamauchi[1], Takahiro Hasegawa[1], Manabu Hashimoto[2], Shuichi Akizuki[2], Yukiyasu Domae[3], Ryosuke Kawanishi[3]

I. Intoroduction

At the Amazon.com logistics warehouse in the United States, the Kiva Pod robot made by Kiva Systems (Amazon Robotics as of 2016) automatically conveys products from the storage shelves to the people responsible for picking. Manual labor is still needed to pick these products from the shelves, but it is expected that this task will eventually be automated by introducing picking robots. In an e-commerce business where there are many types of products stored randomly on shelves, the key to the introduction of automation is being able to perform stable pick-and-place operations by recognizing diverse objects on the shelves and gripping them correctly. Against this background, Amazon set up the Amazon Picking Challenge (APC) as a competitive event for robots in the automation of logistics. The first such event was APC 2015, which focused on the problem of picking diverse items. Contestants were required to build picking robots that could extract 25 different items from 12 frames (called "bins") on a shelf [1]. At APC 2016 in the following year, the scope of the competition was made more realistic by having the contestants compete on two different tasks ("Pick" and "Stow"). This paper introduces the robot system of Team C^{2}M at Amazon Picking Challenge 2016, and its image recognition system.

II. Two Cooperative Industrial Robots

In this section, we discuss the Team C^{2}M robot system and its features.

A. Robot System

As shown in Fig. 1, the Team C^{2}M robot system consists of two MELFA industrial robots with load capacities of 7 kg (RV-7FL) and 4 kg (RV-4FL), both equipped with 3D vision sensors (MELFA-3D Vision), force sensors (4F-FS001) and multifunctional hands. The robot with a 7-kg load capacity was installed on a single-axis sliding platform. The robot that directly accesses the shelf needs to be able to pick up randomly placed items. We used the robot with a 7-kg load capacity for this purpose because it has a longer reach and can carry heavier loads. For the robot that extracts items from the tote, we used a robot with a 4-kg load capacity, which has sufficient reach and load capacity for this purpose.

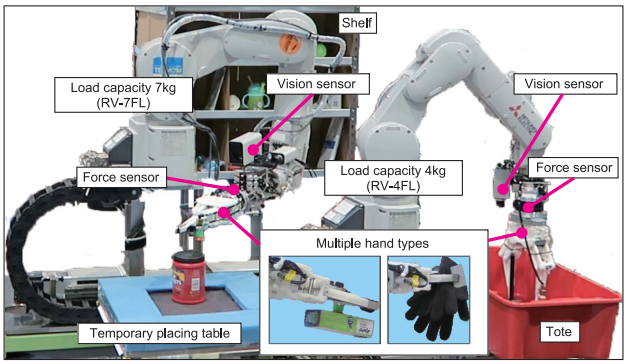

Fig. 1. Overall view of the Team C^{2}M robot system.

A force sensor is mounted at the end of the robot arm to prevent collisions and breakages by judging if unnecessary force is being applied to the robot, shelf or item when gripping products. Also, by touching the shelf with a hand equipped with a force sensor, we were able to judge the fine positional offset from the reference position, and implemented automatic calibration between the shelf and robot. The 3D vision sensors produced RGB image and depth image outputs by employing an active stereo method comprising a camera and projector. This sensor is used for three-dimensional measurement and recognition of items.

B. Features

When putting the items picked up by the robot on the shelf, it is necessary to estimate the item's 6D object pose and set it down with the optimal orientation. However, there are a wide variety of items to be picked including rigid objects, non-rigid objects and transparent objects, and it is difficult to estimate the 6D object orientation of all these types of object. Therefore, our team stored items on the shelf by having two robots work cooperatively via a temporary placing table. By placing items temporarily on the temporary placing table, the robots can hand over the items in such a way that they are easy to put away. By introducing a temporary placing table between the two robots, we simplified the complex task of

[1]H. Fujiyoshi, T. Yamashita, Y. Yamauchi, and T. Hasegawa are with the Machine Perception and Robotics Group, Chubu University, 1200 Matsumoto-cho, Kasugai-shi, Aichi 487-8501, JAPAN. {hf@cs., yamashita@cs., yuu@isc., tkhr@mprg.}chubu.ac.jp

[2]M. Hashimoto, and S. Akizuki are with the Intelligent Sensing Laboratory, Chukyo University, 101-2 Yagoto Honmachi, Showa-ku, Nagoya-shi, Aichi 466-8666, JAPAN. {mana@, akizuki@}isl.sist.chukyo-u.ac.jp

[3]Y. Domae, and R. Kawanishi are with the Advanced Technology R&D Center, MITSUBISHI Electric Corporation, 8-1-1, Tsukaguchi-honmachi, Amagasaki-shi, Hyogo 661-8661, JAPAN. {Domae.Yukiyasu@cb, Kawanishi.Ryosuke@bx}.MitsubishiElectric.co.jp

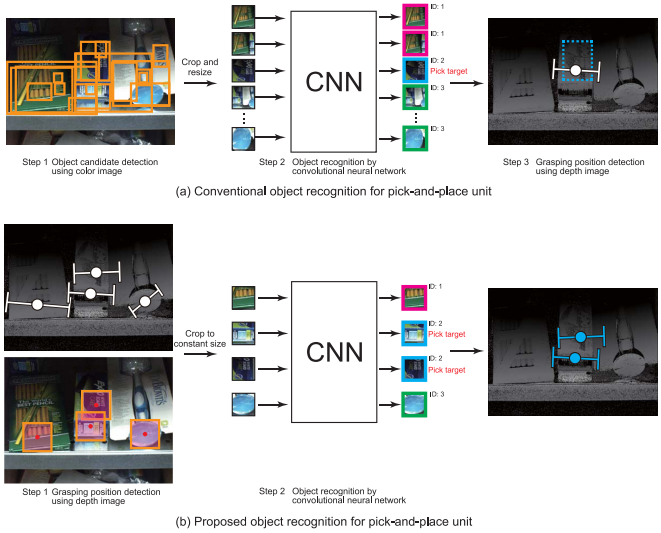

Fig. 2. Workflow of object recognition for pick-and-place unit.

putting away items on the shelf.

III. Vision Strategy

Here, we describe vision strategy of the Team C^{2}M.

A. Grasping Position Based Object Recognition by CNN

A Convolutional Neural Network (CNN) identifies items that appear in the input images. In general, the recognition of objects by a CNN is performed as shown in Fig. 2(a), where candidate object regions called "region proposals" are extracted and provided as input to the CNN for identification. To detect as many region proposals as possible to avoid missing out any picking items, this type of CNN must run many times, resulting in a large computational load.

To reduce this load, we used the approach shown in Fig. 2(b), where the image is first analyzed to detect grasping positions, and the CNN is only then used to identify these items based on the image regions surrounding these detected grasping positions. By detecting the grasping positions first, we can limit the recognition processing to these detected positions. This approach is much more efficient because the robot can then move directly on to picking actions. The grasping positions can be detected at high speed by using Fast Graspability Evaluation [2].

As shown in Fig. 3, the CNN in the proposed method consists of convolution layers and fully connected layers. The convolution layers use batch normalization in layers {1, 2}, and max pooling in layers {1, 2, 5}. We used the ReLU activation function for all layers. In class identification by a CNN, the class probabilities are generally calculated by the *softmax* function. Here, the class probabilities are calculated by the softmax function $P_r(\cdot)$ as shown in Eq. (2), where h_i represents the values of the CNN output units and C is the number of classes.

$$P_r(h_i) = \frac{\exp(h_i)}{\sum_{j=1}^{C} \exp(h_j)} \tag{1}$$

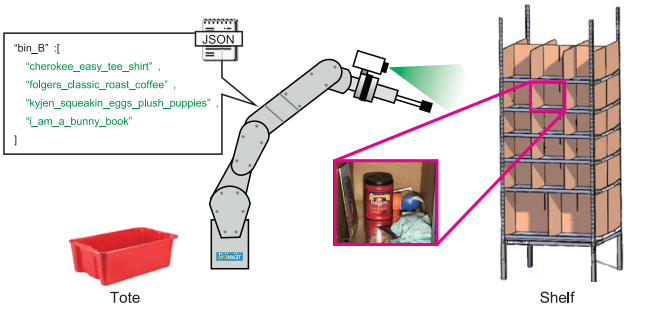

Fig. 4. Picking strategy.

In the task set by the Amazon Picking Challenge, the robot system received a JSON file containing the names of items stored in the bin, and the names of the items to be picked. The robot system thus knows what items are in the bin before it performs object recognition. We propose a *constrained softmax* function where the CNN output layer only calculates class probabilities for output units corresponding to items that are present in the bin. For example, when the bin contains items corresponding to the set of output units {1, 3, 4}, the constrained softmax function $P_c(\cdot)$ can be defined as shown in Eq. (2).

$$P_c(h_i) = \frac{\exp(h_i)}{\sum_{j=\{1,3,4\}} \exp(h_j)} \tag{2}$$

Using the constrained softmax function, it is possible to eliminate the possibility of mistakenly recognizing items that are not in the bin.

B. Picking Strategy

The Pick task involves moving each of the twelve items in a bin by picking them up one at a time and putting them in a box called a "tote". In this task, a JSON file containing the names of the items in the bin is read in and applied to the constrained softmax function. Fig. 4 shows the item recognition procedure. First, the target bin is photographed to acquire an image. Information about the items inside the bin is then read in from the JSON file. Next, the items to be picked are recognized by the constrained softmax function using only the output units that correspond to the items in the bin.

C. Stowing Strategy

The Stow task involves picking twelve randomly placed items from the tote and storing them in bins on the shelf. Fig. 5 shows the item recognition procedure in the Stow task. In the Stow task, we used two robot arms and force sensors either side of a temporary placing table. First, we pick a graspable item from the tote. The item is moved to the temporary placing table, and its weight is measured by the force sensor. Based on the item's weight, the list of candidate graspable items is narrowed down to four possibilities. The item placed on the temporary placing table is then recognized by the constrained softmax function using

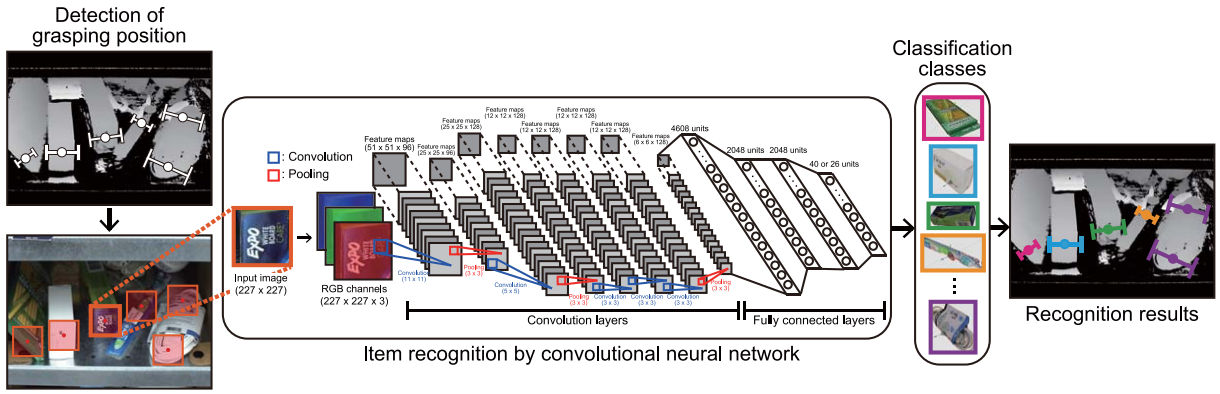

Fig. 3. CNN network structure.

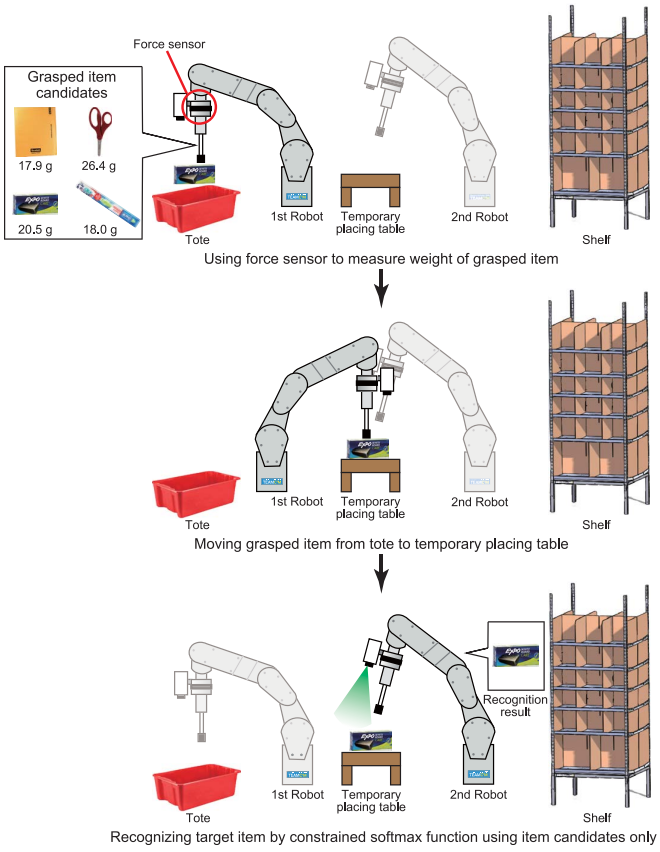

Fig. 5. Stowing strategy.

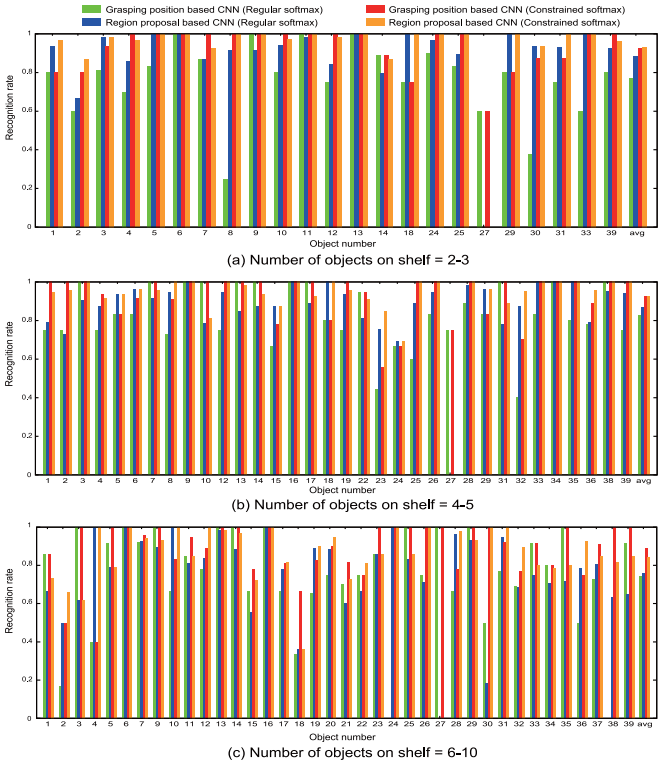

Fig. 6. Recognition rate of APC2016 dataset.

output units corresponding to these four possibilities based on the item's weight.

IV. Experiments

We performed an evaluation experiment to confirm the effectiveness of the proposed method. We compared the recognition accuracy and processing time of the conventional method (CNN based on region proposal) and the proposed method (CNN based on grasping position). We compared the recognition accuracy when using an ordinary softmax function and a constrained softmax function in the output layer of each method.

A. Recognition Results

We compared the recognition accuracy of CNN based on region proposal with the proposed method of CNN based on grasping position. Fig. 6 shows the recognition accuracy of each item in the APC2015 data set. We performed separate evaluations with 2–3 objects, 4–5 objects and 6–10 objects on the shelf. The horizontal axis of the graph represents the object number, and the final column shows the average recognition rate of all objects. Using constrained softmax, we confirmed that CNN based on grasping position can achieve an average recognition accuracy at least as high as that of CNN based on region proposal.

Fig. 7 shows the recognition results of the Pick test in Amazon Picking Challenge 2016. In the Pick task, the recognition process of the proposed method was applied

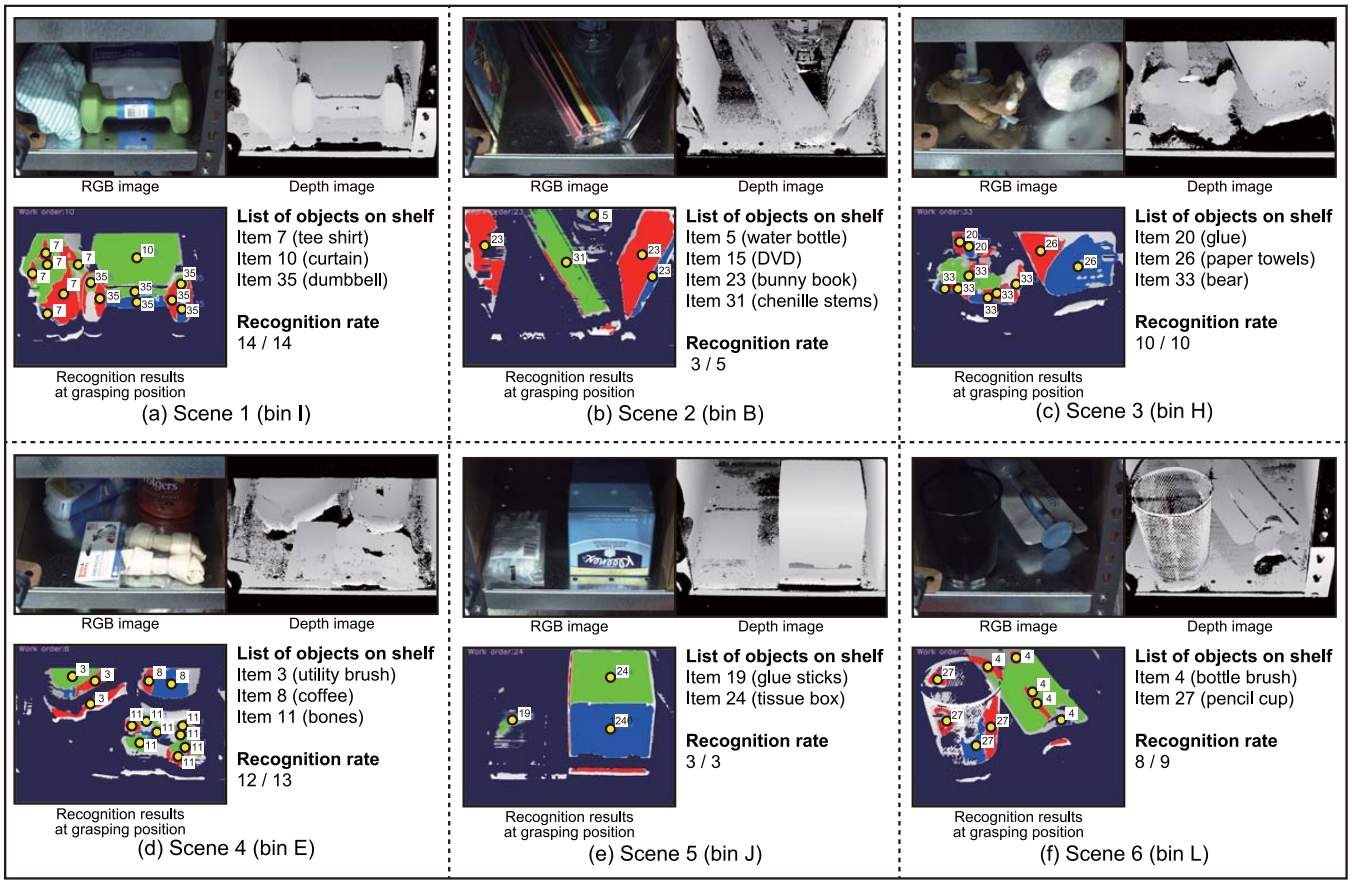

Fig. 7. Recognition results in APC 2016.

to six scenarios, and achieved a recognition accuracy of 92.6% (i.e., 50 correct recognition results out of 54 detected grasping positions).

B. Computational Time

The processing times of CNN based on region proposal and CNN based on grasping position are shown in Table I. We measured the processing times on a system with a 3.5 GHz Intel Core i7 CPU and an NVIDIA GeForce GTX 780M GPU. We detected an average of five grasping positions in CNN based on grasping position, and an average of 17 regions in CNN based on region proposals.

Compared with CNN based on region proposal, the processing speed of CNN based on grasping position was about 3.7 times faster when running on the CPU, and about 2.7 times faster when running on the GPU.

TABLE I

COMPARISON OF COMPUTATIONAL TIMES [MS].

	CPU processing	GPU processing
Grasping position based CNN	269.7	62.4
Region proposal based CNN	1008.7	173.3

V. CONCLUSION

Although APC 2016 featured harder problems than APC 2015, the problem scenarios were still somewhat artificial due to characteristics such as the low incidence of occlusion between items. Since real problems are more complex, it is expected that future events will have a higher level of difficulty in both the Pick task and the Stow task in order to approximate real-world problems more closely. We also felt the need to concentrate more on robot safety, following an accident where someone's arm was injured by colliding with a robot hand during the competition. Picking and storing items more flexibly is likely to be a key subject for future study. The purpose of the Amazon Picking Challenge is to analyze and discuss technical issues by building robot systems, analyzing their performance, and competing them against one another. Although there are still various issues that must be overcome before such systems become a practical reality, the open sharing of knowledge at this event aims to promote the resolution of key issues. Through the Amazon Picking Challenge, it is hoped that large advances will be made in picking robot technology.

REFERENCES

[1] N. Correll, K. E. Bekris, D. Berenson, O. Brock, A. Causo, K. Hauser, K. Okada, A. Rodriguez, J. M. Romano, and P. R. Wurman, "Lessons from the Amazon Picking Challenge," arXiv preprint arXiv:1601.05484, 2016.

[2] Y. Domae, H. Okuda, Y. Taguchi, K. Sumi, and T. Hirai, "Fast Graspability Evaluation on Single Depth Maps for Bin Picking with General Grippers," in *IEEE International Conference on Robotics and Automation*, 2014, pp. 1997–2004.