J Intell Robot Syst
DOI 10.1007/s10846-011-9605-1

Learning Novel Objects for Extended Mobile Manipulation

Tomoaki Nakamura · Komei Sugiura · Takayuki Nagai · Naoto Iwahashi · Tomoki Toda · Hiroyuki Okada · Takashi Omori

Received: 8 December 2010 / Accepted: 12 May 2011

Abstract We propose a method for learning novel objects from audio visual input. The proposed method is based on two techniques: out-of-vocabulary (OOV) word segmentation and foreground object detection in complex environments. A voice conversion technique is also involved in the proposed method so that the robot can pronounce the acquired OOV word intelligibly. We also implemented a robotic system that carries out interactive mobile manipulation tasks, which we call "extended mobile manipulation", using the proposed method. In order to evaluate the robot as a whole, we conducted a task "Supermarket" adopted from the RoboCup@Home league as a standard task for real-world applications. The results reveal that our integrated system works well in real-world applications.

Keywords Mobile manipulation · Object learning · Object recognition · Out-of-vocabulary · RoboCup@Home

T. Nakamura (✉) · T. Nagai
The University of Electro-Communications,
1-5-1 Chofugaoka, Chofu-shi, Tokyo, Japan
e-mail: naka_t@apple.ee.uec.ac.jp

K. Sugiura · N. Iwahashi
National Institute of Information
and Communications Technology, 3-5 Hikaridai,
Seika, Soraku, Kyoto, Japan

T. Toda
Nara Institute of Science and Technology,
8916-5 Takayama, Ikoma, Nara, Japan

H. Okada · T. Omori
Tamagawa University, 6-1-1 Tamagawagakuen,
Machida, Tokyo, Japan

1 Introduction

Mobile manipulation is a fundamental task required for domestic service robots. Therefore, many humanoid robots have been developed with the ability of mobile manipulation [1–5]. Recently, competitions have been proposed to evaluate such robots, such as RoboCup@Home [6], Mobile Manipulation Challenge [7], and Semantic Robot Vision Challenge [8].

We focus on object learning in mobile manipulation using natural speech instruction, such as "Bring me X" (X is an out-of-vocabulary (OOV) word), because we believe it important that users who do not have knowledge of robots can easily interact with one. We assume that a user uses only speech interaction with a robot when he/she teaches objects to it and asks it to bring something. Such mobile manipulation is defined as "extended mobile manipulation". This type of manipulation is desirable for domestic service robots because there are people who do not have extensive knowledge in robots in a domestic

Published online: 23 July 2011

environment. However, extended mobile manipulation is difficult because many features, such as navigation, manipulation, speech recognition, and image recognition, are required.

Image and speech recognition are difficult especially when novel objects are involved in the system. For example, there are objects specific to each home and new products can be brought into the home. It is impossible to register the name and image of all these objects with the robot in advance. We propose a method for learning novel objects with a simple procedure.

The robot, on which the proposed learning method is implemented, is intended to be used in a private domestic environment. Therefore, the procedure of teaching objects to the robot must be simple. For example, the user says, "This object is X" (X is the name of the object) and shows the object to the robot (Fig. 1, left). It is easy for a user to teach a robot many objects with this procedure. Then the user orders the robot to bring him/her something. For example, the user says, "Bring me X" (Fig. 1, right). As we mentioned earlier, such extended manipulation tasks are necessary for domestic service robots. However, there are three problems in teaching a robot novel objects. The first problem is speech recognition of an object's name. In usual methods, phonemes of names must be registered in an internal dictionary. However, it is impossible to register all objects in advance. The second problem is the speech synthesis. A robot must utter the name of the recognized object for interaction with humans such as "Is it X?". However, conventional robot utterance systems cannot utter a word which is not registered in the dictionary. Even if the phoneme sequence of an OOV word can be recognized, it cannot be used for speech synthesis since accuracy of phoneme recognition is less than 90%. The third problem is segmentation of the object region from a scene in the learning phase. When a robot learns an object, it must find where the object region is in the scene and segment it.

Methods for extracting OOV words in a speech have been proposed [9] for solving the first problem. Phonemes of OOV words can be obtained with these methods but they are not always correct. To solve the second problem, the user is required to restate the OOV word again and again so that correct phonemes are obtained [10]. The robot can utter the word correctly but it is best that the robot learns the word from one user's utterance. There are also methods for situations in which the correct phonemes are not obtained. With such methods, the user utters the spelling of the OOV word to correct the phonemes [11]. However, this requires a long time for the robot to learn OOV words by recognizing their spelling in Japanese or Chinese. The use of the keyboard to input the object name can be possible. However, the user of domestic service robots is not always familiar with the keyboard. Furthermore, we consider the speech is the most natural communication between robots and humans like it is a natural communication between humans.

We solve the first problem by extracting OOV words from a template sentence. The second problem may be solved by uttering phonemes of OOV words using a text-to-speech (TTS) system. However, it is difficult to recognize phonemes correctly. In the proposed method, the OOV part of the user's speech is converted to the robot's voice

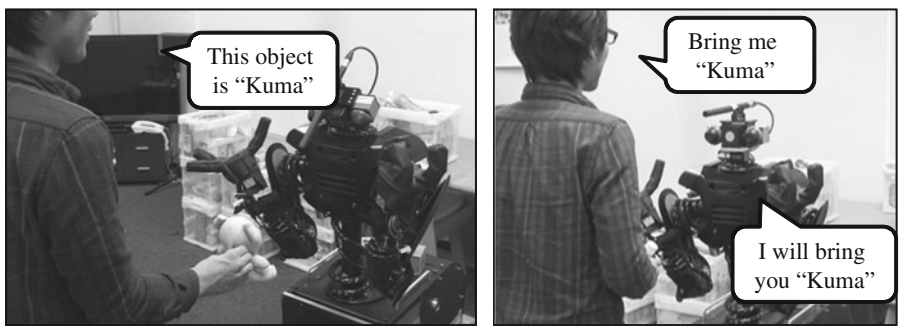


Fig. 1 *Left*: the user teaches the object to the robot. *Right*: the robot recognizes and utters the OOV word

by voice conversion using Eigenvoice Gaussian mixture models (EGMMs) [12].

There has been research on the segmentation of images [13–15] for solving the third problem. The method developed by Rother et al. [13] requires a rough hand-drawn region of an object. Shi and Malik's method [14] can segment images automatically, but it cannot determine which segment is the object region. Mishra and Aloimonos's method [15] can segment accurately using color, 3D information, and motion. However, an initial point that locates inside the object region must be specified.

On the other hand, an object, which a human moves, can be extracted from complicated scene because the proposed method is designed for a human to teach a robot an object. A color histogram and scale invariant feature transform (SIFT) are computed and registered in a database. This information is used for object recognition.

We implement the proposed method with a robot called "DiGORO". We believe it is important to evaluate the robot in a realistic domestic environment with a realistic task. When the robot moves to the object, it does not always arrive at an ideal position nor angle, and the illumination changes according to the position. A system is needed to work well in such an environment. In this paper, we used the "Supermarket" task of RoboCup@Home [6] as the extended mobile manipulation task. RoboCup@Home is a competition that tests the ability of robots in a domestic environment. Supermarket is a standardized task based on the fetch-and-carry operation. There are also other tasks which can be used for evaluation [7, 8]. The "Semantic Robot Vision Challenge" [8] evaluates the ability of a robot to find an object in a real environment; however, only three teams participated in the 2009 competition. Furthermore, Semantic Robot Vision Challenge is not for evaluating manipulation. The "Mobile Manipulation Challenge" was held at the 2010 International Conference on Robotics and Automation. This competition evaluates the mobile manipulation ability of robots; however, only four teams participated. t is difficult to determine what task should be used for evaluating robots, even though there are tasks [6–8] for it. We used one of the tasks of RoboCup@Home, which we believe is the most standard. RoboCup@Home has the largest number of participants[1] and has clearly-stated rules, which are open to the public. Besides, the rules are improved every year. From these reasons, such tasks are better than self-defined ones.

The contributions of this paper are as follows.

1. Segmentation of objects using motion attention and weights adaptation.
   When a user teaches a robot a novel object, the robot can segment the target object from the background by focusing on the motion cue and the object probability map based on the weighted color and depth information. These weights are adapted automatically. Details are described in Section 3.
2. Utterance of OOV words using voice conversion.
   A robot can utter an OOV word from listening to it once. Details are described in Section 4.
3. Evaluation of the robot using "Supermarket" task of RoboCup@Home.
   We evaluate the integrated robot in a realistic domestic environment. We used the "Supermarket" task of RoboCup@Home [6] as the extended mobile manipulation task.

2 Related Work

In this section we explain some related works.

There have been studies on online object learning [16–18]. Hasler et al. [16] and Kim et al. [17] have proposed object-learning methods in which the user shows objects to a robot. However, object segmentation methods proposed in them use only the depth cue, and they do not consider the names of the objects. Wersing et al. [18] has proposed a method in which the learning of objects and their names is done by the user showing the objects and uttering their names. Also, they have proposed a

[1] 24 teams participated in 2010 RoboCup@Home competition [6]. On the other hand, a few teams participated in Mobile Manipulation Challenge [7], and Semantic Robot Vision Challenge [8].

saliency map based on depth, color and motion cues. However, they have not considered out-of-vocabulary words, and the saliency map has been used only for gaze selection. Moreover, these methods have been evaluated using simple recognition tasks in which a user shows an object to the robot. We believe it is important for robots to be able to recognize objects in a realistic domestic environment with a realistic task.

Iwahashi studied language acquisition using a bottom up approach [19]. Roy [20] also used a bottom up approach. However, these studies used simple visual features, with which it is difficult to recognize everyday objects. On the other hand, our method can recognize various objects.

Fujita et al. [21] aimed at acquisition of OOV words. However, this research did not take into account segmentation of OOV words. In this research, the user utters only OOV words when the robot learns them and this research does not involve robot's utterance of OOV words. Nakano et al. also conducted research to obtain phonemes of OOV words by multiple utterances [10]. However, it is easy to learn OOV words with one utterance, and in that sense the proposed method is efficient in this task. Johnson-Roberson et al. [22] extracted novel objects through the user teaching a robot how many objects there are. However, the proposed method is easier to use because the user only moves objects.

Many domestic service robots [1–5] have recently been developed. They can carry out mobile manipulation because they have arms, cameras, and wheels. "HRP-2W" [1] can recognize pointing gestures and learn daily life behavior, such as handling of tableware or cleaning of furniture. "PR2", which is the successor to "PR1"[2], can carry out many tasks including mobile manipulation. "Care-O-Bot3" [3] has a touch panel on its abdomen and interacts with humans with this touch panel, not speech. These robots can carry out mobile manipulation; however, they were not evaluated on this feature. Also, the main difference between this study and the above studies is that DiGORO is able to learn novel objects and OOV words on site.

Many robots [4, 5] have been developed for RoboCup@Home. These robots are also designed to be used at home and can carry out extended mobile manipulation. They are evaluated in RoboCup@Home and obtain high scores. The ability of learning novel objects and OOV words on site is also the main difference between this paper (DiGORO) and the robots in [4, 5].

3 Finding Novel Objects in Cluttered Scene

3.1 3D Visual Sensor

Figure 2 shows the visual sensor used in this paper. The sensor can acquire color and accurate depth information in real time by calibrating a TOF and two CCD cameras.

The distance measurement capability of TOF camera is based on the TOF principle. In TOF systems, the time taken for light to travel from an active illumination source to the objects in the field of view and return to the sensor is measured. In this paper, an off-the-shelf TOF camera SwissRanger SR4000 [23] is used. It emits a modulated near-infrared (NIR) and the CMOS/CCD imaging sensor measures the phase delay of the returned modulated signal at each pixel. These measurements in the sensor results in a 176×144 pixel depth map.

In the geometric camera calibration, the parameters that express camera pose and properties can be classified into extrinsic parameters (i.e. rotation and translation) and intrinsic ones (i.e. focal

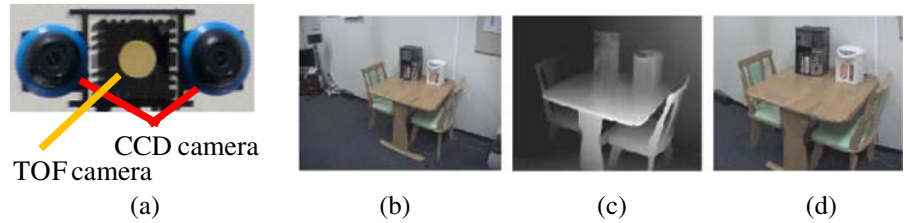


Fig. 2 **a** 3D visual sensor. **b** Color image (1024×768). **c** Depth image (176×144). **d** Mapped color image (176×144)

length, coefficient of lens distortion, optical center and pixel size). The extrinsic parameters represent camera position and pose in 3D space, while the intrinsic parameters are needed to project a 3D scene onto the 2D image plane. We use Zhang's calibration method in our proposed system, since the technique only requires the camera to observe a checkerboard pattern shown at a few different orientations. For the calibration of TOF camera, the reflected signal amplitude can be used to observe the checkerboard pattern. Therefore, it is straightforward to apply the same calibration method. Images captured from the calibrated sensor are shown in Fig. 2b, c and d.

3.2 Object Segmentation Based on Motion Attention

Since we assume that a user shows a target object to the robot, there may be people, objects, or furniture behind that object. The problem is object segmentation in such a complex background. Because the user has the object, the object can be segmented out by taking into account the motion cue. This fact motivates us to use object segmentation based on motion attention.

Figure 3 shows an overview of motion attention. A motion detector first extracts the initial object region $M(x, y)$. Then, object information, such as color (hue) image $H(x, y)$ and depth image $D(x, y)$, is taken from the region. In particular, a hue histogram $f_H(h)$ and depth histogram $f_D(d)$ are taken from the region and normalized. h and d represent the quantized value of hue and depth, respectively. Since these two histograms can be considered as probability density functions of the target object, the object probability map of each component ($P_D(x, y)$ and $P_H(x, y)$) at each pixel location can be easily obtained.

$$P_D(x, y) = f_D(D(x, y)), \tag{1}$$

$$P_H(x, y) = f_H(H(x, y)). \tag{2}$$

The weighted sum of these two object probability maps results in the object probability map $P_O(x, y)$.

$$P_O(x, y) = \mathrm{LPF}[w_d P_D(x, y) + w_h P_H(x, y)], \tag{3}$$

The weights w_d and w_h are automatically assigned inversely proportional to the variance of each histogram. If the variance of the histogram is larger, its information is considered as inaccurate and the weight decreases. LPF represents a low pass filter, and we use a simple 3×3 averaging filter as the low pass filter. The map is binarized, and then a final object mask is obtained using the connected component analysis. In the learning phase, object images are simply collected, then color histograms and SIFT features are extracted. These are used for object detection and recognition.

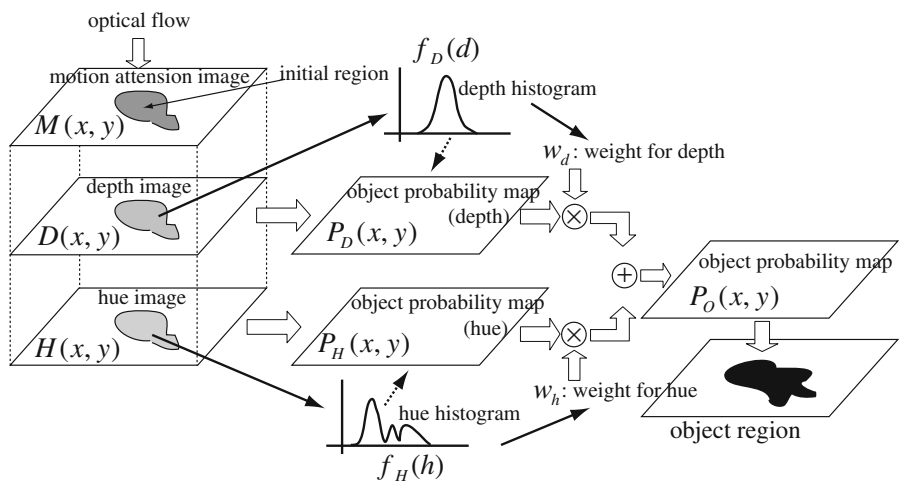


Fig. 3 Segmentation of object region using motion attention

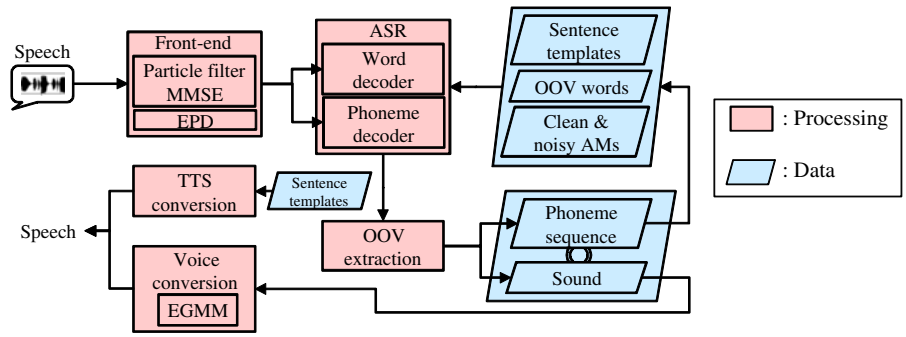


Fig. 4 Overview of the speech processing

3.3 Object Detection and Identification in Recognition Phase

When the robot recognizes an object, the target object should be extracted from the scene. However, the same method in the learning phase is not applicable because the object is placed somewhere and it is not held by the user. Therefore, if objects are on the table, the plane detection technique is beneficial for detecting the objects. The 3D randomized Hough transform [24] is used for fast and accurate plane detection. This plane detection method is summarized below.

1. 3D information is captured in the scene
2. Maximum plane is detected as the top of the table using randomized Hough transform [24]
3. The plane is removed from 3D information
4. The remaining point is projected on the plane
5. Connected components analysis is performed on the plane and each object is segmented out

We use SIFT descriptors for recognition. We first narrow down the candidates by using color information followed by the matching of SIFT descriptors, which are collected during the learning phase. It should be noted that the SIFT descriptors are extracted from multiple images taken from different viewpoints. Moreover, the number of object images is reduced for speeding up the SIFT matching process by matching among within-class object images and discarding similar ones. This process is also useful for deciding the threshold on the SIFT matching score.

4 Pronouncing Out-of-Vocabulary Words Using Voice Conversion

Figure 4 shows a schematic of the speech processing of the method, which uses an automatic speech recognition (ASR) system called ATRASR [25]. ATRASR is a hidden Markov model (HMM)-based speech recognition system, and it is used as a front-end and word/phoneme decoder. The phoneme decoder is used for obtaining the phoneme sequence of OOV words; therefore, word- and phoneme-level speech recognition is possible.

To suppress noise, a particle filter is first applied to the online estimation of non-stationary noise, and then minimum mean square error (MMSE) estimation is used for noise reduction [26]. Voice activity detection is conducted using endpoint detection (EPD) based on the frame's energy. This noise reduction part is of critical importance in RoboCup@Home tasks since the noise condition is severe.

Acoustic models (AMs) for the speech recognizer consist of "clean AMs" (male and female voices), which are trained using only clean voices, and "noisy AMs" (male and female voices), which are trained clean voices mixed with noise. This makes the speech recognition system robust in a noisy environment.

We use a template-based segmentation of words. To teach a robot an OOV word, the user is supposed to say template sentences such as "This is X." In terms of practical use, using a standard template sentence is reasonable since it is easy for users to understand how to teach a robot a word.

A set of segmented voice and phoneme sequences is registered in a database. The phoneme sequence is used for utterance recognition of an OOV word.

For generating an utterance with an OOV word, the proposed method first converts the segmented voice recorded when the OOV word is learnt. The other part of the utterance is synthesized using XIMERA [27], which is a TTS conversion system. The OOV word part is converted into the robot's voice since the original sound is the user's voice, which is not naturally concatenated with a synthesized voice. The voice conversion is based on Eigenvoice Gaussian mixture models (EGMMs) [12]. The recognized phoneme sequence of the OOV word is not used for synthesis since phoneme recognition accuracy is less than 90%, and the number of utterances for teaching an OOV word is virtually constrained to one owing to the time constraint of RoboCup@Home.

5 Procedure of Extended Mobile Manipulation Task

In this section, we describe the procedure of the extended mobile manipulation task called "Supermarket".

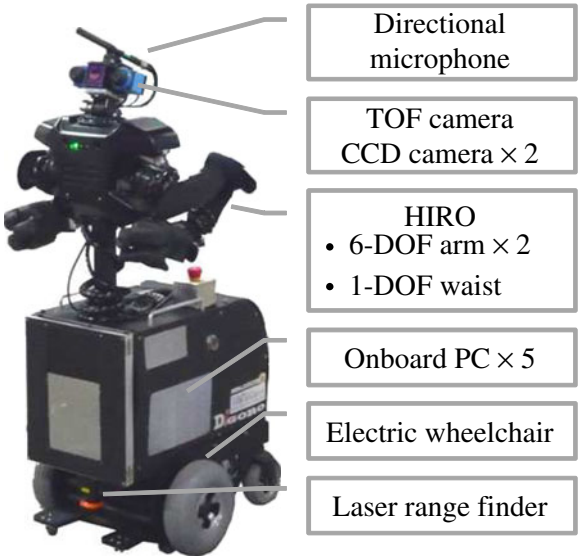


Fig. 5 The robot platform "DiGORO"

5.1 Robot Platform: DiGORO

Figure 5 shows the robot "DiGORO" we previously developed [28]. It is composed of the following hardware:

- Electric wheelchair
- HOKUYO laser range finder UTM-30LX
- KAWADA upper body humanoid HIRO
- Onboard PC (Intel Core2Duo processor) × 5
- Sanken directional microphone CS-3e
- YAMAHA loudspeaker NX-U10
- Mesa infrared TOF camera Swissranger
- Imaging Source CCD camera × 2

5.2 Learning of Novel Object

Before the task, we need to teach objects to the robot. The procedure of teaching objects is summarized below.

1. The user shows the object to the robot and say "DiGORO, this object is X (X is an OOV word.)" in Japanese.[2] The phoneme sequence and sound of the OOV word are extracted from the user's speech. (Fig. 6a)
2. The robot says "I start learning of X."
3. The user moves the object. Then 40 images of the object are segmented out and captured. (Fig. 6b)
4. Visual features (SIFT descriptors and color histogram) are calculated.
5. The phoneme sequence of the OOV word, sound of the OOV word and visual features of the object are registered in the object database. Then the robot says "I've memorized X."

If the user moves the object incorrectly, the motion attention cannot segment out the object. For example, if the user moves the object from outside the observed frame into it, the unintentional region is segmented out. To avoid such a situation,

[2] In this paper, the utterances are translated into English.

(a) (b)

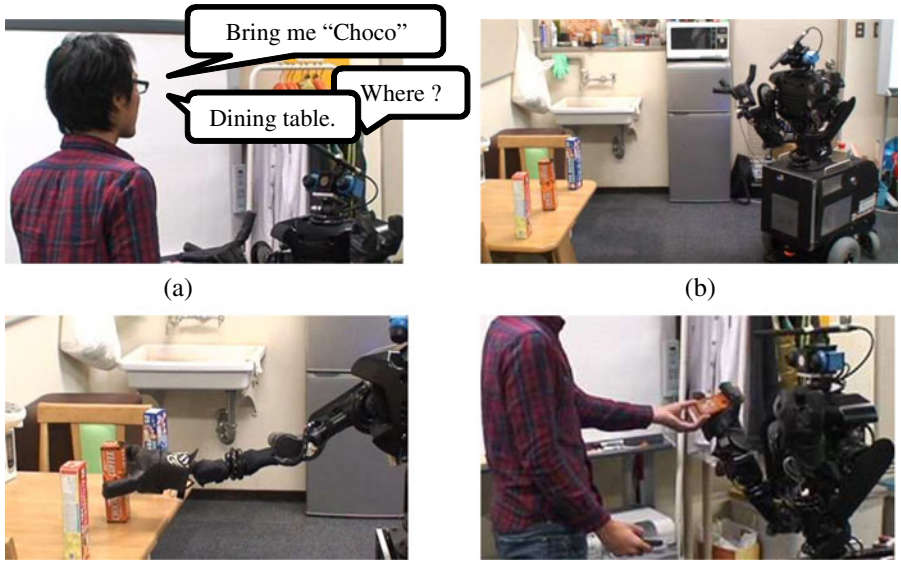
Fig. 6 Scenery of the learning phase. **a** The user gives a command for the robot to learn a new object. **b** The user shows the object from various directions

we instructed the user to say, “DiGORO, this is X.”, while showing the object to the robot and to keep on showing it until the robot says, “I’ve memorized X.”.

5.3 Supermarket Task

Supermarket is a task that the robot brings three specified objects. The detailed procedure of this task is summarized below.

1. The user says “DiGORO, bring me X. (X is an OOV word.)” (Fig. 7a)
2. The robot says “I’ll bring you X. Is that correct?”. Then if the user says “DiGORO, no”, go to 1. If the user says “DiGORO, yes” go to next.
3. The robot says “Where is X?”.
4. The user says “It is P. (P is name of a place.)”.
5. The robot says “I’ll go to P. Is that correct?”. Then if the user says “DiGORO, no”, go to 4. If the user says “DiGORO, yes” go to next.
6. The robot goes to P (Fig. 7b).
7. The robot looks around to find X using plane detection. If the robot can find X, the robot turns toward the object and says “I found X.”.
8. The robot grabs the object (Fig. 7c) and returns to the start position (Fig. 7d).
9. The task is completed when the robot brings three objects in total. Otherwise go to 1.



(a) (b)

(c) (d)

Fig. 7 Scenery of the supermarket task. **a** The user ordered the robot to bring “choco”, and the robot asked the user for needed information. **b** The robot navigated to the place where the user ordered. **c** The robot found the object and grabbed it. **d** The robot returned to the start position and handed over the object to the user

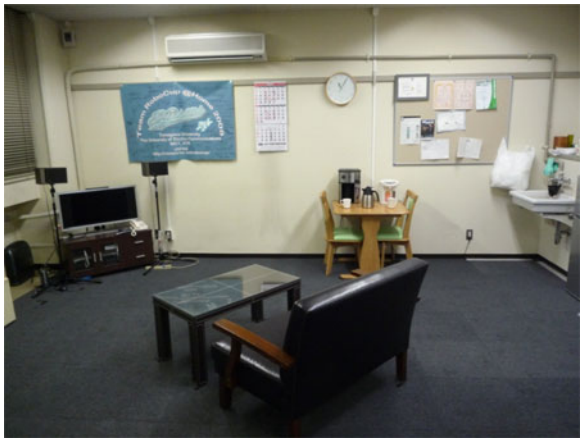


Fig. 8 Experimental environment

6 Experiment 1: Evaluation of Proposed Method

In this section, we evaluate performance of the robot's learning and recognition. Learning novel objects and recognition of image processing are evaluated first, and then that of speech processing is evaluated. Our integrated robot of these methods is evaluated in Section 7.

6.1 Segmentation Accuracy

In this experiment, we evaluated segmentation accuracy.

It is necessary to segment the object region from a complicated background. We used motion attention discussed in Section 3 because the object is held by a user in the learning phase.

The experiment was carried out in an ordinary living room, shown in Fig. 8. We used 120 ordinary objects, as shown in Fig. 9. A user taught the robot every object by showing and telling the names to the robot. The robot acquired 40 consecutive frames for each object and extracted the target object region from each image. Figure 10 shows examples of object segmentation.

We extracted 10 out of the 40 frames for evaluating the segmentation accuracy of each object. Detection accuracy was measured using recall and precision rates, which are generically used for evaluation of classification, as shown in Fig. 11a, because pixels are classified into two classes, object region and non-object region. "Object region" indicates the manually labeled object region. Figure 11b shows a 2D plot of recall vs. precision. Each point represents an average of a single object (ten frames). As a result, the averages of all objects were 76.2% for recall and 95.8% for precision. This result indicates that the inside of object regions is extracted correctly because precision is high.

Figure 12 shows examples of segmentation failure. The object in Fig. 12a was not segmented at all because it reflected near infrared rays and 3D information could not be obtained. A section of the object in Fig. 12b was segmented because the black region absorbed near infrared rays. Also, a section of the object in Fig. 12c was segmented because it reflected near infrared rays. We can

Fig. 9 120 objects used for experiments

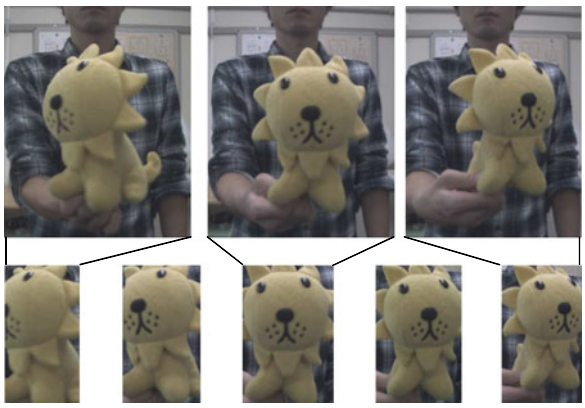


Fig. 10 Examples of object segmentation

see that black and metallic objects tend to cause low recall.

6.2 Object Recognition Accuracy

We used 120 common objects, which had been learnt by the robot, as mentioned in the previous subsection. Three different locations (different lighting conditions) in the living room were selected, and each object, which was segmented out using motion attention, was recognized. The results are listed in Table 1. The average recognition rate was about 90%. A major problem was false recognition between similar kinds of object such as cup noodles with different taste, because they have similar texture.

Next, we evaluated the proposed recognition method with the COIL100 database [29]. COIL100 consist of 100 objects and 72 images per object. 36 images of each object were used for learning and the other 36 images were used for recognition. The recognition rate was 97.6%.

6.3 Recognition Accuracy of Out-of-Vocabulary Words

We evaluated the recognition accuracy of OOV words. The experimental procedure is described as follows. The teacher taught the robot OOV words such as "This is X." In a domestic environment, the teacher may not be only one person but also his/her family or friends. Therefore, we conducted the experiment under the condition

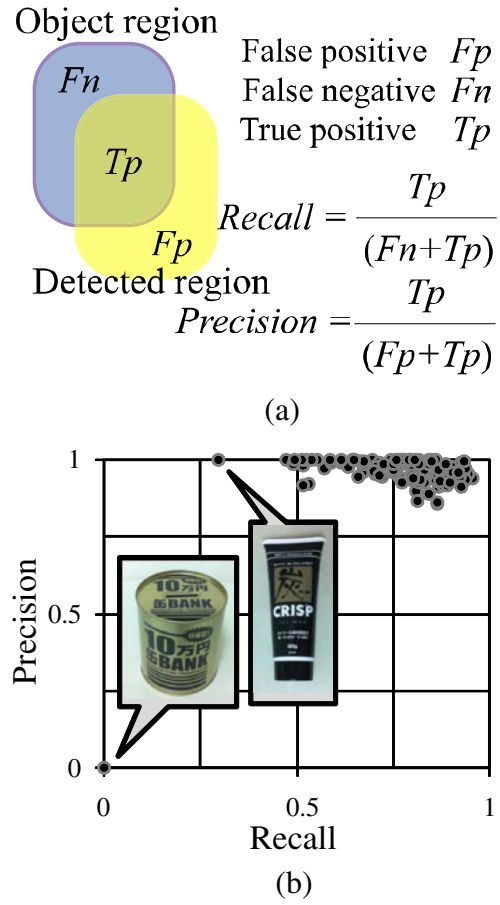


Fig. 11 **a** Definitions of recall and precision. **b** Results of object detection

that OOV words are taught by several teachers including the user who asked the robot to bring something. For comparison, we conducted the experiment under simpler conditions. In each condition, volunteers uttered sentences which included the 120 words "Bring me X." and the robot recognized X. There were eight volunteers and 960 utterances were recognized in total. The distance between the volunteer and the microphone was 50 cm. The ambient noise level in the experiment was set as 55 dBA, which simulated the standard noise level in the RoboCup@Home competition when there is no other noise sources such as announcement. If the speech recognition system can work in 55 dBA noise, it can also work in a domestic environment. The recognition rate was calculated from these utterances.

(a) (b) (c)

Fig. 12 Examples of object segmentation failure. **a** Object which could not be extracted. **b** *Left*: object *right*: results of segmentation. **c** *Left*: object *right*: results of segmentation

Figure 13a shows the recognition rate in each condition, and the details of each condition are as follows:

1. Recognition with correct phonemes: Correct phonemes of the 120 words were manually registered in the dictionary. Each volunteer uttered "Bring me X" (X is the object name) and the robot recognized the object name.
2. Teacher and user are same person: Each volunteer uttered 120 sentences "This is X." (X is the object name) and the robot learnt the 120 OOV words. The robot recognized the 120 OOV words spoken by the volunteer who was the same as the teacher.
3. Teachers taught OOV words: First, 120 words were randomly assigned to eight teachers and these words were taught to the robot by them. Then, the robot recognized the 120 OOV words spoken by the user who is one of teachers. Therefore, the words were not always taught by the user. 118 out of the 960 were spoken by the teacher, i.e. teacher was the same as the user, and 842 out of the 960 utterances were spoken by others, i.e. teacher was not the same as the user.

The recognition rate was 95.2% in Condition 1, as shown in Fig. 13a. On the other hand, the accuracy of phonemes was 69.3% and the recognition rate was 82.4% in Condition 2. This indicates that the recognition rate was over 80%, which is satisfactory in a practical situation. In Condition 3, the recognition rate was 75.2%, as shown in Fig. 13a. The recognition rate was 83.4% when the teacher was the same as the user and 74.1% when the teacher was not the same as the user. Note that the speech files used in the training and those used in the test were different, even if the trainer and the tester was the same person. We can see that the recognition rate was lower than that in Condition 2. However, this is not a problem if restating is allowed.

Table 1 Object recognition rates

	Place 1	Place 2	Place 3
Recognition rate	91%	89%	89%

6.4 Quality Evaluation of Robot's Utterances

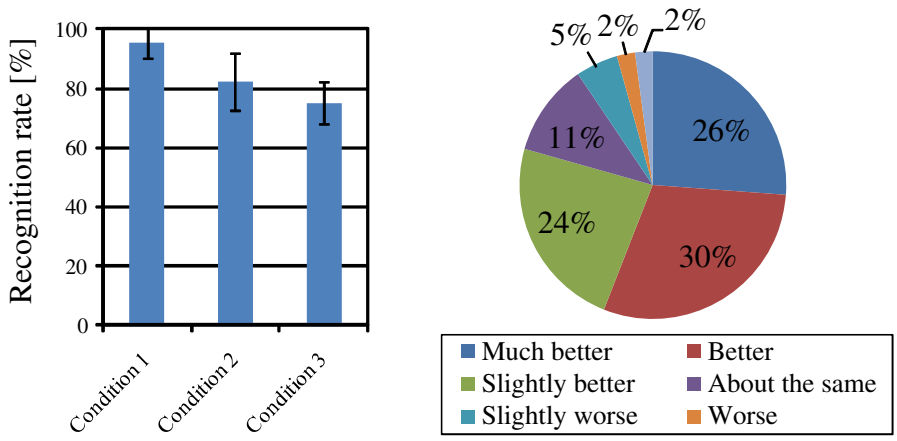
The objective of this experiment was to evaluate the quality of the robot's utterances. The experimental procedure is described below.

First, we made a database that included the 960 utterances. It had 120 unique words and each word was uttered by eight volunteers. The ambient noise level was 55 dBA and the distance between the volunteer and microphone was 50 cm. Next, robot's utterances were generated using the proposed method. Utterances were also generated using a baseline method for comparison. These two methods are summarized as follows:

Voice Conversion (VC) (proposed) The utterances in the database are converted to robot voice by using EGMMs [12] (details of the proposed method were explained in Section 4).

Text-To-Speech (TTS) (baseline) The phoneme sequences obtained by phoneme recognition were used for generating robot utterances.

We then formed another group of six volunteers to evaluate the quality of the generated utterances. Each volunteer listened to the utterances generated using TTS and VC. These



(a) (b)

Fig. 13 **a** Recognition results. (Condition 1: recognition with correct phonemes. Condition 2: teacher and user are same person. Condition 3: teachers taught the OOV words.) **b** Evaluation of voice conversion. The CMOS of VC was 1.45

utterances were composed of 120 unique words. The order of words was chosen at random. The order of TTS and VC samples was also chosen at random for each trial.

The comparison mean opinion score (CMOS) was used for evaluation. CMOS is specified by ITU-T recommendation P.800 [30]. In the field of speech synthesis, CMOS is used for comparing voices synthesized with two methods. Specifically, the evaluation was conducted using the following questionnaire.

> (Volunteer listens to two robot's utterances.) Do you think the former is more accurate than the latter in terms of pronunciation?

The evaluation and its scores are listed in Table 2.

The evaluation results are shown in Fig. 13b. The CMOS of VC was 1.45, which suggests that VC is preferred. We can see that the proposed method, which uses VC, is efficient if the word learnt is uttered once.

Table 2 CMOS evaluation and scores

Quality	Score
Much better	3
Better	2
Slightly better	1
About the same	0
Slightly worse	−1
Worse	−2
Much worse	−3

7 Experiment 2: Evaluation of Applied System in Extended Mobile Manipulation

We implemented an integrated audio-visual processing system on DiGORO and performed an experiment in a living room. The purpose of this experiment was to evaluate the robot to which the proposed method is applied in extended mobile manipulation. We chose a task called "Supermarket" in the RoboCup@Home league. The advantage of RoboCup@Home is it has the largest number of participants and has clearly-stated rules, which are open to the public. Besides, the rules are improved every year.

7.1 Experimental Setup

Figure 14 illustrates a map generated from DiGORO's own on board SLAM mapping module. The location of the tables/shelf is also shown.

We designed the task module according to the flow in Section 5.3. A volunteer first interacted with the robot at the start position. Then the robot navigated to a table/shelf, recognized the specified object, grasped it, and came back to the volunteer. This process was repeated for three objects.

We conducted the task under two conditions. One was similar to a real competition and the other was a more difficult condition. In each condition, we changed the dictionary of speech recognition because the teacher who teaches the object

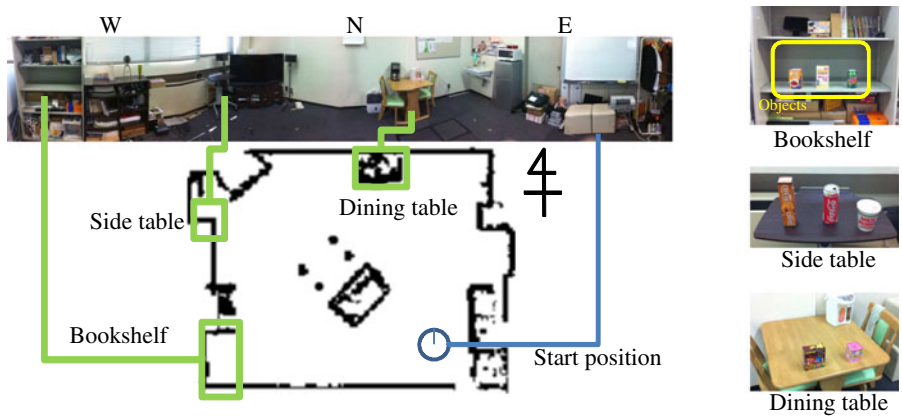


Fig. 14 The map and location of the tables/shelves

to the robot may not be the only person. The details of the conditions are as follows.

Condition 1: In the learning phase, each volunteer taught the robot the objects' names. The same volunteer asked the robot to bring the objects in the execution phase.

Condition 2: In the learning phase, 120 words were randomly assigned to eight volunteers and they taught these words to the robot. Each volunteer asked the robot to bring objects in the execution phase. Therefore, the names of the objects to bring were not always taught by the same volunteer who commanded the robot in the execution phase.

In the two different experimental setups, five volunteers who don't have prior knowledge of the robot conducted the task. Therefore, the robot was supposed to bring 30 objects throughout this experiment. In each task, 30 out of the 120 objects were randomly chosen. The training data for these objects were obtained in Experiment 1.

7.2 Experimental Results

We evaluated the results from three view points, success rate of each process, process elapsed time, and the score as total performance.

Figure 15 shows the success rate of each process. We can see that high success rates over 90% were obtained, except for speech recognition. The speech recognition rate was 93% in Condition 1. On the other hand, it was 80% in Condition 2. This is because the phoneme sequences in the lexicon were not accurate.

Figure 16 depicts the average elapsed time for each process (per object). The results suggest that the trial can be completed within 10 min (elapsed time should be tripled and added 60 sec for the robot's instruction). The phase of instruction to the robot took a long time. This was because of confirmation from the robot such as "I will bring X. Is this correct?" or the volunteer had to restate the instruction to the robot such as "Bring me X." when the robot could not recognize the object name. The instruction phase in Condition 1 was shorter than that in Condition 2 because false recognition in Condition 1 was less than that in Condition 2. This figure also shows that the time of the object recognition phase in Condition 2 was longer than that in Condition 1 because the object location was chosen randomly in both conditions.

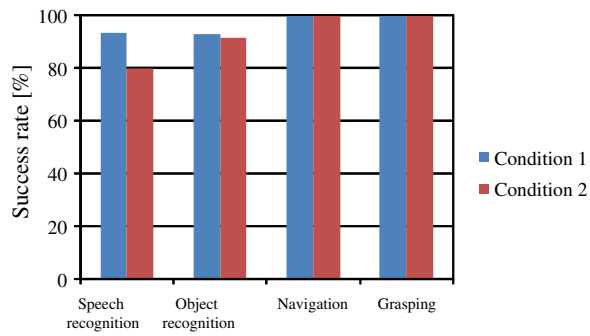


Fig. 15 Success rates. (Condition 1: words are taught by the same as requester. Condition 2: words are taught by different volunteers)

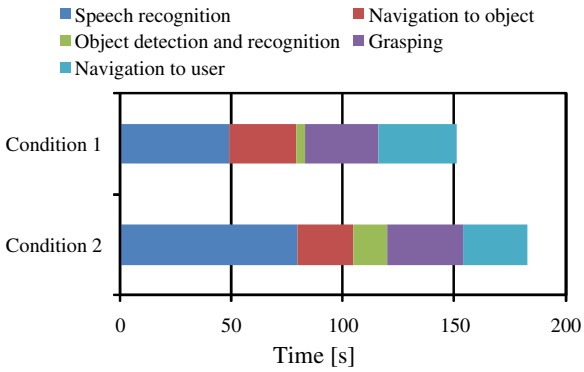


Fig. 16 Elapsed time of each process. (Condition 1: words are taught by the same as requester. Condition 2: words are taught by different volunteers)

It accidently took a long time to find objects in Condition 2, depending on their location.

Next, we evaluated the task scores as a reference. Note that the comparison of the scores may be unfair because there are differences between a laboratory and competition environments. However, we used the scores since they can be the only source for comparison among different robots through the same realistic task.

Figure 17 shows a comparison of scores among teams that participated in an actual competition in 2009. The average score in Condition 1 was 1560. From this score, DiGORO would outperform the best team in the competition. Furthermore, the average score in Condition 2 was 1320, which was comparable to Team A. It should be noted that we used the average scores; however, a team could performed a task only once in the competition. In that respect, this comparison may be unfair.

In an actual competition, three objects which the robot brings were selected from ten common objects whose names list was given to the teams. Therefore, it was possible to manually register the names of all the objects with the dictionary. On the other hand, objects which the robot brings were chosen from 120 objects in our experiment. Moreover, no manual process was included in the learning process. Considering these conditions, we can see that our robot obtained promising results even though the environment was different from the competition.

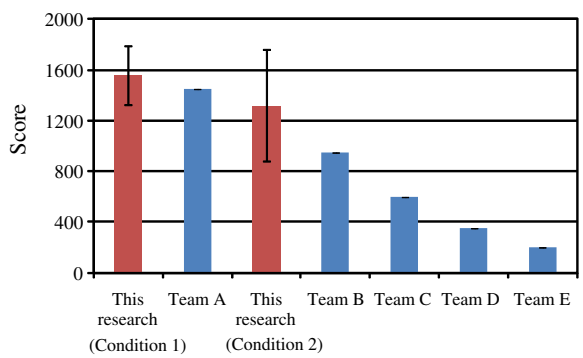


Fig. 17 Score comparison. (Condition 1: words are taught by the same as requester. Condition 2: words are taught by different volunteers)

8 Discussion

8.1 Image Processing

In this section, we discuss the results from the evaluation of segmentation accuracy. Precision was 95.8%, which indicates that the inside of the object region was extracted correctly. On the other hand, recall was 76.2%, which was less than precision. This indicates that sometimes only part of the object region was segmented. This is because the TOF camera could not capture 3D information due to the material of the object. For example, 3D information cannot be captured from black or metallic objects because these reflect or absorb near infrared rays. We believe this will be improved by using a stereo camera. DiGORO (Fig. 5) has two CCD cameras and can compute stereo disparity with them.

We now discuss the results of object recognition. The object recognition rate was about 90%. We used color and SIFT features for object recognition. Generally, it is difficult to recognize objects that have the same color and with no textures. For future work, we plan to use an object recognition method that integrates 3D shape information [31], which can significantly improve object recognition performance.

8.2 Learning and Recognition of OOV Words

For this research, the robot learnt OOV words from one user's utterance and it is possible for the

robot to recognize and utter them. The recognition rate was 82.4% and utterance was judged as better than the baseline method, which means a practical system is constructed. Failure in recognition was because false phonemes were learnt in the learning phase. The recognition rate can be improved by a user confirming which phonemes were learnt correctly or not after learning. For example, a user utters "This is X" and the robot learns the object. Then the user confirms which "X" can be recognized or not by asking "Did you memorize X?". If the robot utters "Yes, I memorized X'" ($X = X'$), the OOV word is registered correctly. Otherwise, the OOV word may not be registered correctly and the user can teach the object name again to the robot.

8.3 Evaluation in Domestic Environment

We evaluated the system in a domestic environment using the Supermarket task, which is one of the tasks in the RoboCup@Home league. Here, let us briefly discuss the evaluation task. As we mentioned earlier, it is difficult to determine what task should be used for evaluation, and there is no global standardized tasks for this. This situation makes it very difficult to evaluate robots, which were developed by different groups, through a same realistic task. We cannot compare our robot with others by using a self-defined task, since it is almost impossible to build their robots from scratch. Therefore, we think global standardized tasks are needed.

In this paper, we propose to utilize the format of the task of RoboCup@Home, since we strongly believe that the tasks are the most standard tasks for evaluating robots for the following reasons:

1. The rules are open in the public.
2. Many teams from around the world participate, i.e. the task has already been performed by many robots.
3. The rules have been improved by many robotics researchers.

Unfortunately, the comparison of the scores in the current form is unfair. Hence the score should be treated as a reference. Although the score is just for a reference, DiGORO outperforms the best team who participated in the competition, and it shows DiGORO can function in a domestic environment. Any deduction in points was a result of the robot not recognizing what a user wanted it to bring. This can be improved by user confirmation in the learning phase, as mentioned above.

The learning and recognition of OOV words can be applied to other tasks. For example the "Who is who?" task, which is one of the tasks in RoboCup@Home, involves the learning of human faces and names. In this task, a user utters "My name is X" and the robot learns "X" as his/her name. With this method, we can deal with a vast number of names.

Furthermore, DiGORO has many other abilities, and it can carry out eight other tasks. For example the robot can carry out the command "Follow Me", which is for following humans, and "Shopping Mall", which is for learning the location in an unknown place. These advanced features led our team to the 1st place at RoboCup@Home 2010. This suggests that DiGORO can stably work in a domestic environment. A video of the RoboCup@Home 2010 is available on the web site (http://apple.ee.uec.ac.jp/isyslab/digoro/press/video/).

9 Conclusion

We proposed a practical learning method of novel objects. With this method a robot can learn a word from one utterance. It is possible to utter an OOV word using the segmentation of the word from a template sentence and voice conversion. The object region is extracted from a complicated scene through a user moving the object. We implemented them all in a robot as a object learning system and evaluate it by conducting the Supermarket task. The experimental results show that our robot, DiGORO, can stably work in a real environment. A video of DiGORO operating with the proposed method is available on the web site (http://apple.ee.uec.ac.jp/isyslab/digoro/demo/video/demo001.html).

Acknowledgements This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (C), 20500186, 2010.

Appendix A: "Supermarket" Task

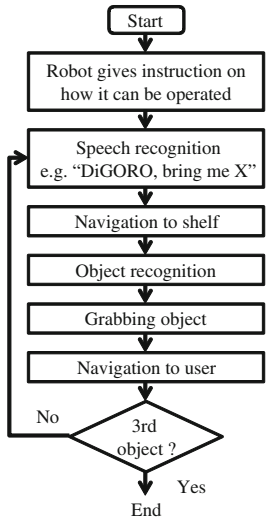


Fig. 18 Flowchart of Supermarket task

The Supermarket task is a competition in the RoboCup@Home league. This task has the following three advantages.

- The rules are open to the public.
- Many teams from around the world participate, that is, the task has been performed by many robots.[3]
- The rules are improved every year by many robotics researchers.

From these advantages, we can compare the scores of other teams in an actual competition.

The procedure of the task is described as follows. First, the robot learns the object and its name before the task. At this time, learning the OOV word and segmentation of the object are required. Then, the user instructs the robot to bring the specified object. At this time, recognition of the OOV word is required. This task is completed if the robot can move to the specified location, find the specified object, grab it, and deliver it to the user. Utterance is required in the task because the robot must interact with humans. The Supermarket task is suitable as a benchmark for domestic service robots because it includes object learning, object recognition, OOV word learning, OOV word recognition, navigation, and manipulation.

The official rules [32] are described as follows.

> Referees randomly select a person who does not have prior knowledge on how to use a robot. He/She gets the robot to deliver a maximum of three objects from one or more shelves within ten minutes. The robot is allowed to give instructions on how it can be operated. The three objects are taken from a set of common objects. The team which has the robot can choose one of the objects, and the other two objects are chosen by the referees (respecting the physical constraints of the robot). The referees put the three objects on one or more shelves. The team has to announce whether the robot would be able to reach the objects from different levels before the test starts.
>
> The score system is defined as follows:
>
> 1) Correctly understanding which object to get: For every correctly understood object, 50 points are awarded, i.e., by clearly indicating the object.
> 2) Recognition: For every correctly found object, 150 points are awarded.
> 3) Grabbing: For every correct object retrieved from the shelf, 100 points are awarded. If the object was lifted for at least five seconds another 100 points are awarded.
> 4) Delivery: For every object delivered to the person, 100 points are awarded.
> 5) Different levels: For getting objects from different levels, 200 points are awarded.
> 6) Multimodal input: For using gestures in a meaningful way, in addition to speech, to communicate with the robot, 300 points are awarded.
> 7) Onboard microphone: For recognition using an onboard microphone instead of a headset, 50 points are awarded.

We were not awarded 6) since our robot cannot recognize the gesture. Hence, the maximum score we can obtain was 1,750 points.

We scored using the following criteria. If the robot could say "I'll bring you X" correctly when

[3] 24 teams participated in 2010 RoboCup@Home competition [6]. On the other hand, a few teams participated in Mobile Manipulation Challenge [7], and Semantic Robot Vision Challenge [8].

the user asked to bring the object, we gave a score of (1). If the robot could turn toward the correct object when the robot said "I found X.", we gave a score of (2). If the robot returned within 1 meter of the start position, we gave a score of (4).

In our experiment, the tables/shelves were chosen based on whether the robot's arms could reach them, as it done in the competition. We actually prepared tables and a shelf with variable height and the heights were selected randomly within the reachable range of our robot's arm. We tried to execute the tasks in strict accordance with these rules.

References

1. Inamura, T., Okada, K., Tokutsu, S., Hatao, N., Inaba, M., Inoue, H.: HRP-2W: a humanoid platform for research on support behavior in daily life environments. Robot. Auton. Syst. **57**(2), 145–154 (2009)
2. Wyrobek, K., Berger, E., Van der Loos, H., Salisbury, J.: Towards a personal robotics development platform: rationale and design of an intrinsically safe personal robot. IEEE Int. Conf. Robot. Autom. 2165–2170 (2008)
3. Weisshardt, F., Reiser, U., Parlitz, C., Verl, A.: Making high-tech service robot platforms available. In: Proceedings-ISR/ROBOTIK 2010 (2010)
4. Stückler, J., Behnke, S.: Integrating indoor mobility, object manipulation, intuitive interaction for domestic service tasks. In: IEEE-RAS International Conference on Humanoid Robots (2009)
5. Holz, D., Paulus, J., Breuer, T., Giorgana, G., Reckhaus, M., Hegger, F., Müller, C., Jin, Z., Hartanto, R., Ploeger, P., et al.: The b-it-bots RoboCup@ home 2009 team description paper. RoboCup 2009@ Home League Team Descriptions, Graz, Austria (2009)
6. RoboCup@Home: (2010)
7. 2010 Mobile Manipulation Challenge: http://www.willowgarage.com/mmc10 (2010)
8. Semantic Robot Vision Challenge: http://www.semantic-robot-vision-challenge.org/ (2009)
9. Bazzi, I., Glass, J.: A multi-class approach for modelling out-of-vocabulary words. In: Seventh International Conference on Spoken Language Processing (2002)
10. Nakano, M., Iwahashi, N., Nagai, T., Sumii, T., Zuo, X., Taguchi, R., Nose, T., Mizutani, A., Nakamura, T., Attamim, M., et al.: Grounding new words on the physical world in multi-domain human-robot dialogues. In: 2010 AAAI Fall Symposium Series, pp. 74–79 (2010)
11. Holzapfel, H., Neubig, D., Waibel, A.: A dialogue approach to learning object descriptions and semantic categories. Robot. Auton. Syst. **56**(11):1004–1013 (2008)
12. Toda, T., Ohtani, Y., Shikano, K.: One-to-many and many-to-one voice conversion based on eigenvoices. In: IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 4, pp. 1249–1252 (2007)
13. Rother, C., Kolmogorov, V., Blake, A.: Grabcut: interactive foreground extraction using iterated graph cuts. ACM Trans. Graph. (TOG) **23**(3), 309–314 (2004)
14. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **22**(8), 888–905 (2002)
15. Mishra, A.K., Aloimonos, Y.: Active segmentation. Int. J. Human. Rob. **6**, 361–386 (2009)
16. Hasler, S., Wersing, H., Kirstein, S., Körner, E.: Large-scale real-time object identification based on analytic features. In: Artificial Neural Networks–ICANN 2009, pp. 663–672 (2009)
17. Kim, H., Murphy-Chutorian, E., Triesch, J.: Semi-autonomous learning of objects. In: Computer Vision and Pattern Recognition Workshop, p. 145 (2006)
18. Wersing, H., Kirstein, S., Gotting, M., Brandl, H., Dunn, M., Mikhailova, I., Goerick, C., Steil, J., Ritter, H., Korner, E.: Online learning of objects in a biologically motivated visual architecture. Int. J. Neural Syst. **17**(4), 219–230 (2007)
19. Iwahashi, N.: Robots that learn language: developmental approach to human-machine conversations. In: Symbol Grounding and Beyond, pp. 143–167 (2006)
20. Roy, D.: Grounding words in perception and action: computational insights. Trends Cogn. Sci. **9**(8), 389–396 (2005)
21. Fujita, M., Hasegawa, R., Takagi, T., Yokono, J., Shimomura, H.: An autonomous robot that eats information via interaction with humans and environments. In: IEEE International Workshop on Robot and Human Interactive Communication, pp. 383–389 (2002)
22. Johnson-Roberson, M., Skantze, G., Bohg, J., Gustafson, J., Carlson, R., Kragic, D.: Enhanced visual scene understanding through human-robot dialog. In: 2010 AAAI Fall Symposium on Dialog with Robots (2010)
23. Mesa imaging: http://www.mesa-imaging.ch/index.php
24. Okada, K., Kagami, S., Inaba, M., Inoue, H.: Plane segment finder: algorithm, implementation and applications. IEEE Int. Conf. Robot. Autom. **2**, 2120–2125 (2005)
25. Nakamura, S., Markov, K., Nakaiwa, H., Kikui, G., Kawai, H., Jitsuhiro, T., Zhang, J., Yamamoto, H., Sumita, E., Yamamoto, S.: The ATR multilingual speech-to-speech translation system. IEEE Trans. Audio, Speech, Lang. Process. **14**(2), 365–376 (2006)
26. Fujimoto, M., Nakamura, S.: Sequential non-stationary noise tracking using particle filtering with switching dynamical system. In: IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1 (2006)
27. Kawai, H., Toda, T., Ni, J., Tsuzaki, M., Tokuda, K.: XIMERA: a new TTS from ATR based on corpus-based technologies. In: Fifth ISCA Workshop on Speech Synthesis, pp. 179–184 (2004)

28. Okada, H., Omori, T., Iwahashi, N., Sugiura, K., Nagai, T., Watanabe, N., Mizutani, A., Nakamura, T., Attamimi, M.: Team eR@sers 2009 in the @home league team description paper (2009)
29. Nene, S.A., Nayar, S.K., Murase, H.: Columbia Object Image Library (COIL-100). Technical report (1996)
30. International Telecommunication Union: ITU-T P.800. http://www.itu.int/rec/T-REC-P.800/en
31. Attamimi, M., Mizutani, A., Nakamura, T., Nagai, T., Funakoshi, K., Nakano, M.: Real-time 3D visual sensor for robust object recognition. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 4560–4565 (2010)
32. RoboCup@Home league committee: RoboCup@Home rules & regulations. http://www.ai.rug.nl/robocupathome/documents/rulebook2009_FINAL.pdf (2009)