Proceedings of the 2021 IEEE/SICE
International Symposium on System Integrations
Iwaki, Fukushima, Japan. January 11 - 14, 2021

2021 IEEE/SICE International Symposium on System Integration (SII) | 978-1-7281-7658-1/20/$31.00 ©2021 IEEE | DOI: 10.1109/IEEECONF49454.2021.9382607

Teaching System for Multimodal Object Categorization by Human-Robot Interaction in Mixed Reality

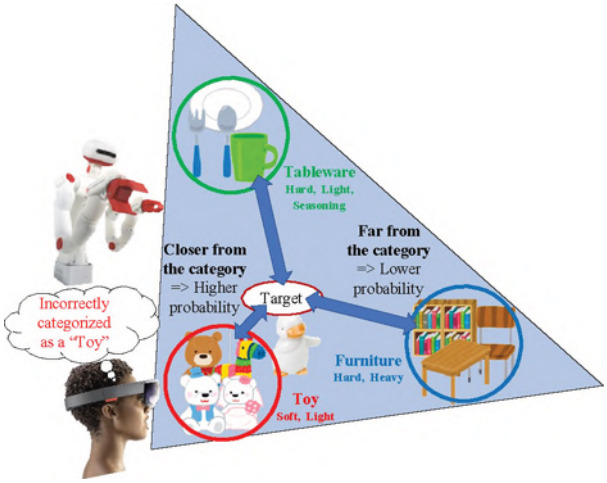
Lotfi El Hafi[1], Hitoshi Nakamura[1], Akira Taniguchi[1], Yoshinobu Hagiwara[1], and Tadahiro Taniguchi[1]

***Abstract*— As service robots are becoming essential to support aging societies, teaching them how to perform general service tasks is still a major challenge preventing their deployment in daily-life environments. In addition, developing an artificial intelligence for general service tasks requires bottom-up, unsupervised approaches to let the robots learn from their own observations and interactions with the users. However, compared to the top-down, supervised approaches such as deep learning where the extent of the learning is directly related to the amount and variety of the pre-existing data provided to the robots, and thus relatively easy to understand from a human perspective, the learning status in bottom-up approaches is by their nature much harder to appreciate and visualize. To address these issues, we propose a teaching system for multimodal object categorization by human-robot interaction through Mixed Reality (MR) visualization. In particular, our proposed system enables a user to monitor and intervene in the robot's object categorization process based on Multimodal Latent Dirichlet Allocation (MLDA) to solve unexpected results and accelerate the learning. Our contribution is twofold by 1) describing the integration of a service robot, MR interactions, and MLDA object categorization in a unified system, and 2) proposing an MR user interface to teach robots through intuitive visualization and interactions.**

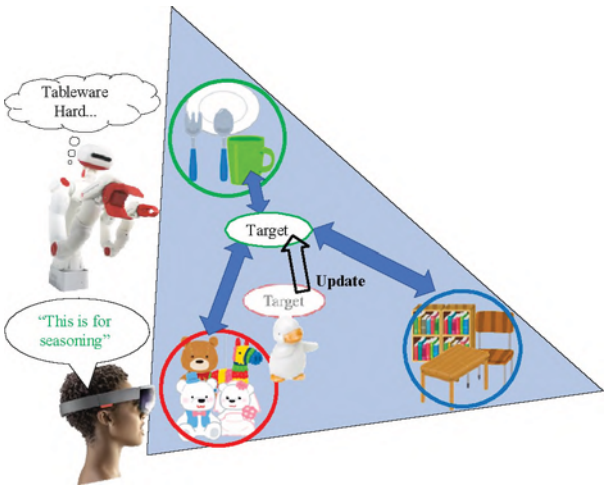
I. Introduction

As service robots are becoming essential to support aging societies, teaching them how to perform general service tasks is still a major challenge preventing their deployment in daily-life environments. In addition, developing an artificial intelligence for general service tasks requires bottom-up, unsupervised approaches to let the robots learn from their own observations and interactions with the users. However, compared to the top-down, supervised approaches such as deep learning where the extent of the learning is directly related to the amount and variety of the pre-existing data provided to the robots, and thus relatively easy to understand from a human perspective, the learning status in bottom-up approaches is by their nature much harder to appreciate and visualize.

Therefore, we propose in this paper a system to solve the problems that users are facing when teaching a robot additional object knowledge without knowing its current



(a) Visualization in MR of which category the target object is likely to be categorized into.



(b) Update of the categorization process with additional vocabulary information.

Fig. 1. In this example, the unknown object to categorize is an ambiguous duck-shaped saltcellar. How the robot will categorize this object from multimodal observations is a priori hard to guess by the user. However, the proposed system allows the user to visualize in MR which category the saltcellar is likely to be categorized into, by displaying them in a spatially organized manner around its position. In the case of Fig. 1a, the green and blue dotted circles are drawn at a far position from the saltcellar, which indicates that it has a lower probability to fall into these categories, respectively "tableware" and "furniture". On the contrary, the category displayed in the red dotted circle is closer, which means that the saltcellar is more likely to be incorrectly categorized as a "toy". Therefore, the user should update the multimodal categorization process with additional vocabulary information. In the example of Fig. 1b, saying the word "seasoning" to the robot will rectify the categorization process and allow the robot to correctly identify the saltcellar as a "tableware".

This study was supported by the Japan Ministry of Education, Culture, Sports, Science and Technology (MEXT) KAKENHI Grant-in-Aid for Scientific Research on Innovative Areas "Integrative Studies of Language Evolution for Co-Creative Human Communication" (Area no. 4903, Task no. 17H06379).

[1]Lotfi El Hafi, Hitoshi Nakamura, Akira Taniguchi, Yoshinobu Hagiwara, and Tadahiro Taniguchi are with Ritsumeikan University; 1-1-1 Noji-Higashi, Kusatsu, Shiga 525-8577, Japan. {lotfi.elhafi, nakamura.hitoshi, a.taniguchi, yhagiwara, taniguchi}@em.ci.ritsumei.ac.jp

learning status. Indeed, when a user wants to teach unknown objects through interactions with a robot, it is particularly difficult for the user to understand the robot's current level of knowledge, especially in the case of bottom-up learning where the extent of the object categorization is not directly related to a dataset with pre-defined categories but instead dependent on the robot's own observations and judgment.

In this regard, Nakamura et al. [1] proposed a model based on Multimodal Latent Dirichlet Allocation (MLDA) that allows a robot to achieve object categorization using multimodal observations obtained from its sensors. Subsequently, Araki et al. [2] enabled a user to teach the names of objects to a robot while it learns the object concepts by using Nakamura et al.'s MLDA model. Although these previous researches focused more on the object categorization itself, they found out that it was hard for the user to understand in which category the robot had categorized the unknown objects because of the lack of visual feedback. Consequently, it is also difficult to teach the robot how to categorize an unknown object into an expected category without knowing the categorization results first, as it is possible that the object falls into a category that the user is not expecting.

To address these issues, we propose a system for teaching robots more efficiently using Mixed Reality (MR)-based visualization and interactions. In particular, the proposed system enables a user to monitor and intervene in the robot's multimodal object categorization process based on MLDA to solve unexpected results and accelerate the learning. Our contribution is twofold:

1) We describe the integration of a service robot, MR interactions, and MLDA-based object categorization in a unified system.
2) We propose an MR-based user interface to teach robots through intuitive visualization and interactions.

The conceptual goal and interface of the proposed system is illustrated by an example in Fig. 1 of a situation where the object categorization from the robot's observations may produce an unexpected result.

The remainder of this paper is structured as follows. Section II introduces previous works related to interactive learning based on MR visualization. Section III describes the proposed system integration and user interface. Section IV details the experiment conducted to test the proposed system. Finally, Section V concludes this paper with avenues for future works.

II. Related Works

This section introduces previous works related to interactive learning and visualization in MR, and how the main contributions of our research differ from them.

A. Assisted Robot Decision-Making in MR

Liu et al. [3] proposed a system to teach a robot new knowledge while visualizing its internal representation in MR. In particular, their system revealed to the user the decision-making process of the robot during object manipulation using the Microsoft HoloLens [4] head-mounted MR display. Furthermore, their system allowed to interactively teach additional knowledge using hand gestures recognition in MR. They tested their system in a scenario where a robot tried to learn the complex movement sequence required to open a bottle cap. Although the robot could not initially open the cap by itself, the robot could eventually open the cap after the user directly edited the tree structure of its decision-making with hand gestures. As a result, they showed that their system could enhance the robot performances through interactions with the user in MR.

Compared to Liu et al.'s research where only the decision-making process of the robot was displayed, we propose a system that can directly influence the learning by visualizing multimodal object categorization results and teach additional knowledge according to them.

B. Visualization of Robot Perception in MR

El Hafi et al. [5] proposed a system to intuitively visualize a robot's sensor information and learning status in MR to realize natural human-robot interactions for service tasks in the context of a "future convenience store" [6]. Their system could display the robot's sensor observations and multimodal spatial categorization results on top of the real environment using a HoloLens.

Although El Hafi et al. could visualize the learning results of the robot, the final goal our proposed system is not to display information, but to teach additional knowledge through MR.

C. Interactive Learning between Users and Robots

Araki et al. [2] proposed a system to allow the robot to acquire vocabulary and object knowledge through online interactive learning with the user. In their study, the robot categorized objects using multimodal information, then the user taught the robot how the objects were called. In particular, the concepts of each object were acquired using Multimodal Latent Dirichlet Allocation (MLDA) [1]. However, the teaching efficiency was limited by the user's lack of understanding of the current knowledge status and learning process of the robot.

Therefore, we address the limitations of Araki et al.'s interactive learning by proposing a system that allows the user to visualize the learning process using MR in order to more efficiently teach the robot based on the information displayed.

III. Proposed System

This section describes the integration of the proposed system and its user interface that allows the user to teach the robot by visualizing in MR the likely results of the multimodal object categorization.

A. System Integration

Fig. 2 is the system diagram of our proposed solution. On the one hand, the user wears a Microsoft HoloLens [4] head-mounted MR display which runs a visualization application

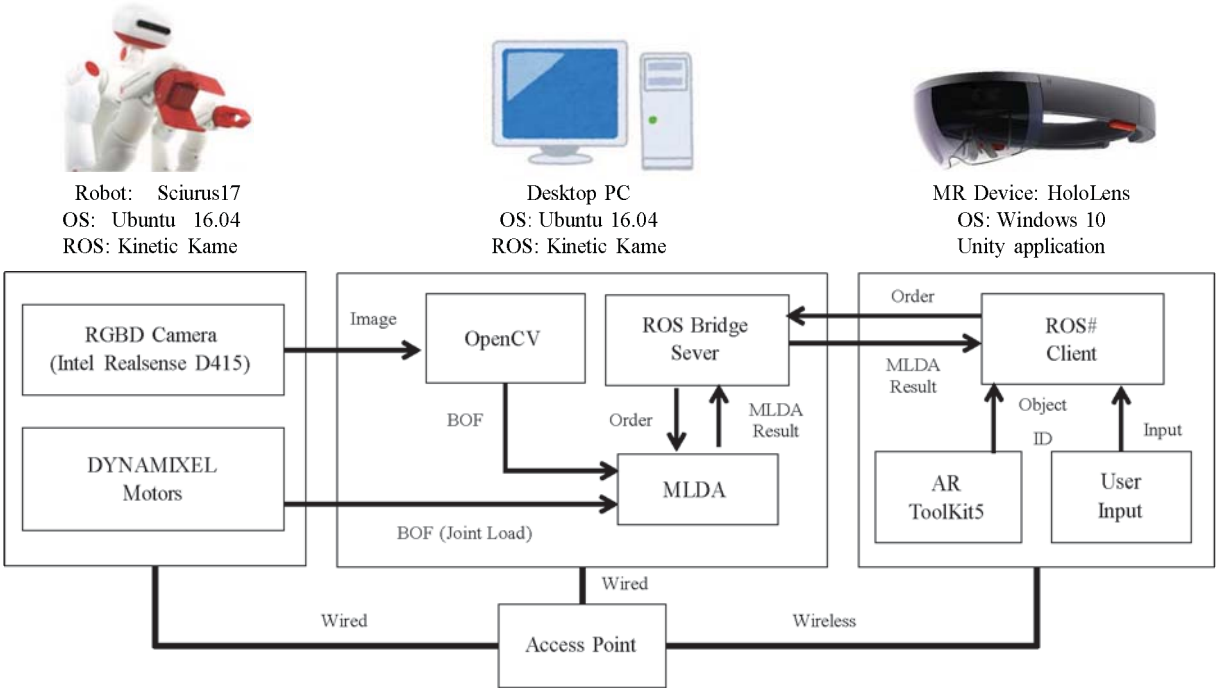


Fig. 2. System diagram of the proposed solution.

built on Unity [7]. On the other hand, the robot is a dual-arm RT Corporation Sciurus17 [8] which runs on the Robot Operating System (ROS) [9] middleware.

On the user's side, ARToolKit5 [10] is used by the visualization application to recognize the positions of the objects with AR markers attached to each of them. The information about the object selected by the user is sent to the robot using the ROS# [11] API which enables network communication between the ROS middleware and the Unity application.

On the robot's side, MLDA is used to learn the user-specified object. MLDA acquires image features as Bag-of-Features (BOF) using the OpenCV [12] implementation of AKAZE [13]. In addition, the robot acquires the maximum and minimum load values of each of its joints as BOF when grasping or lifting the specified object. This multimodal information is then used by the MLDA model to categorize the object, and the result is sent using ROS# to the HoloLens to be visualized by the user in MR.

B. User Interface

Fig. 3 is a conceptual diagram of the proposed MR user interface designed to intuitively display the object categorization results. The categorization results are displayed around a white sphere centered on top of the target object's position in the real world. The target object is also shown on a cube at the bottom of the sphere for visual confirmation of the selection by the user. The top results of the multimodal object categorization from MLDA are represented by purple links that extend from the target object, with the probability of each category displayed in percent. At the extremity of each link is a colored cube around which are displayed, in

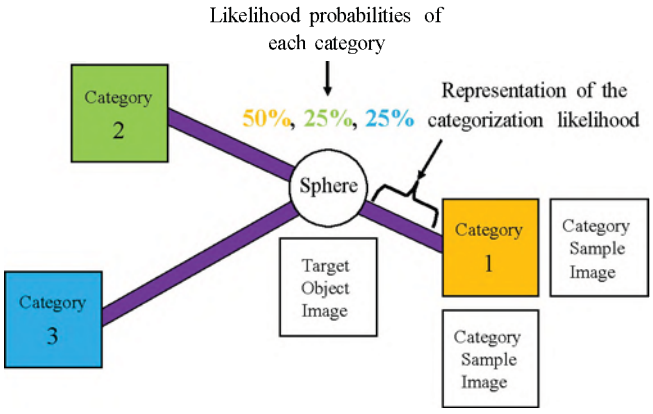


Fig. 3. Conceptual diagram of the proposed MR user interface.

separated white cubes, examples of objects that are likely belonging to that same category.

In addition, the lengths of the purple links represent the likelihood of categorization for each category. The higher the probability is, the shorter the length is. Therefore, the category that is the most likely to be selected by the robot is the closest to the target object, and the categories that unlikely to be selected are displayed farther away. The length of the purple link of a category k is calculated as follows:

$$d_k = \frac{\frac{1}{p_k}}{\sum_{i=1}^{K} \frac{1}{p_i}}, \quad (1)$$

where K is the number of categories, d_k is the length of the purple link of the category k, and p_k is the probability of being categorized into the category k.

(a) Empty bottle. (b) Filled spray. (c) Empty box.

(d) Empty bottle. (e) Filled spray. (f) Empty box.

Fig. 4. Dataset of the daily-life objects used in the experiment. They present various shapes and weights, but also similar visual features.

IV. Experiment

This section details the experiment conducted to test the proposed system and its early results.

A. Object Dataset

Fig. 4 shows the daily-life objects selected for the experiment. They present various shapes and weights, but also similar visual features. Some of them are empty, while others are filled with their original content.

B. Experiment Flow

The overall flow of the experiment is as described below:

Step (1): The user specifies the object to be taught to the robot.

Step (2): The robot observes and records features of the specified object.

Step (3): Step (1) and Step (2) are repeated until the robot has observed the features of all objects.

Step (4): The robot categorizes the objects with MLDA from the observed features.

Step (5): The user visualizes the multimodal object categorization results in MR.

Step (6): Return to Step (1) to restart the process.

C. Early Results

Fig. 5 shows the main menu of the user interface displayed in MR during Step (1). A holographic menu is displayed at each AR maker's position, allowing the user to select the object to be taught to the robot. The ID of the object is written with white characters, and the upper green button is for teaching that specific object. The lower green button is for displaying the categorization results. The user interacts with the buttons of the menus by hand gestures.

Fig. 6 shows how the user interface displays the multimodal object categorization results in MR during Step (5). It takes about 3 sec to render the results in MR after completing the MLDA categorization process. The top results from the multimodal object categorization are represented by purple

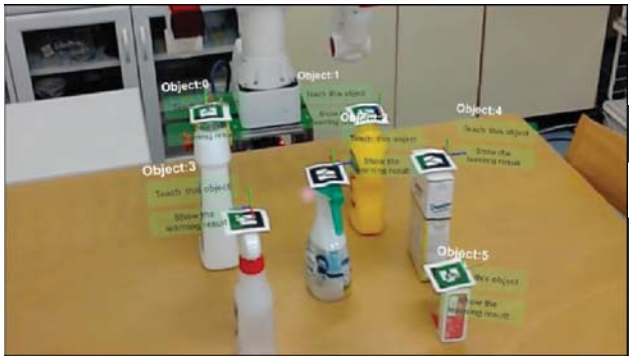


Fig. 5. Main menu of the user interface displayed in MR during the experiment with the robot.

TABLE I

Interpretation of the object categorization results.

Category	Probability	User Interpretation
Category 1 (closest)	62%	Category likely made of similar AKAZE image features (hard to name from a human perspective).
Category 2 (farthest)	18%	Category likely made of empty boxes.
Category 3	20%	Category likely made of filled sprays.

links that extend from the target object, with the probability of each category displayed in percent. The closer to the target object is displayed a category, the higher is the likelihood of the object to fall into that category, unless the user provide additional vocabulary information to update the results.

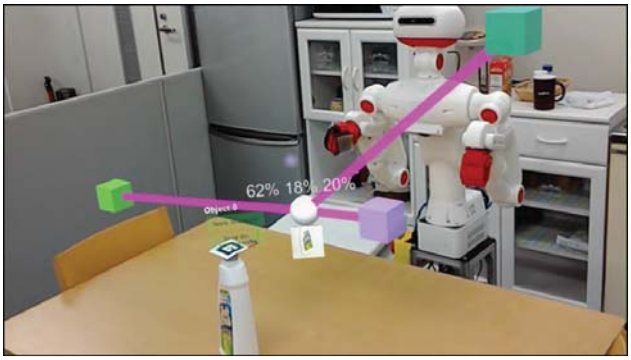
During the experiment shown in Fig. 6, the robot categorized the objects as described in Table I.

V. Conclusion

In this paper, we proposed an MR-based visualization system that allows a user to improve, correct, or accelerate the categorization process of a service robot by teaching additional knowledge when necessary. Compared to previous research on multimodal object categorization, our system showed that users could more efficiently teach an ambiguous object because they could better understand the current learning status of the robot from our proposed MR-based user interface.

However, we recognize that the proposed system is still at an early stage of development, and that further experiments in more realistic task scenarios are required for validation. Especially, we will quantitatively measure in a future work the increase of teaching efficiency when the results of the MLDA-based object categorization are visualized in MR.

In addition, the current system is limited by its reliance on AR markers to identify the object that the user is trying to teach to the robot. Not only the AR markers are unrealistic in daily-life environments, they also impact the speed at which the user interface can be rendered in MR when the number of objects to consider increases, and will ultimately prevent the real-time application of the proposed system. Therefore,



(a) Overview of the multimodal object categorization results.

(b) Sample objects of `Category 1` (closest).

(c) Sample objects of `Category 2` (farthest).

(d) Sample objects of `Category 3`.

Fig. 6. Example of multimodal object categorization results visualized in MR during the experiment with the robot.

future works will also focus on improving the estimation of the object to teach by using alternative interfaces such as the user's eye gaze and hand gestures. In this regard, we believe that head-mounted MR devices are particularly suited compared to mobile screen interfaces.

Finally, user's speech recognition will be implemented to provide the robot with vocabulary information in order to intuitively update the results of the multimodal object categorization with additional word features when necessary.

REFERENCES

[1] T. Nakamura, T. Nagai, and N. Iwahashi, "Grounding of Word Meanings in Multimodal Concepts using LDA," in *Proceedings of 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2009)*, St. Louis, United States, Oct. 2009, pp. 3943–3948.

[2] T. Araki, T. Nakamura, and T. Nagai, "Long-Term Learning of Concept and Word by Robots: Interactive Learning Framework and Preliminary Results," in *Proceedings of 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2013)*, Tokyo, Japan, Nov. 2013, pp. 2280–2287.

[3] H. Liu, Y. Zhang, W. Si, X. Xie, Y. Zhu, and S.-C. Zhu, "Interactive Robot Knowledge Patching using Augmented Reality," in *Proceedings of 2018 IEEE International Conference on Robotics and Automation (ICRA 2018)*, Brisbane, Australia, May 2018, pp. 1947–1954.

[4] Microsoft, "HoloLens," https://www.microsoft.com/en-us/hololens, 2016.

[5] L. El Hafi, S. Isobe, Y. Tabuchi, Y. Katsumata, H. Nakamura, T. Fukui, T. Matsuo, G. A. Garcia Ricardez, M. Yamamoto, A. Taniguchi, Y. Hagiwara, and T. Taniguchi, "System for Augmented Human-Robot Interaction through Mixed Reality and Robot Training by Non-Experts in Customer Service Environments," *RSJ Advanced Robotics*, vol. 34, no. 3-4, pp. 157–172, Feb. 2020.

[6] K. Wada, "New Robot Technology Challenge for Convenience Store," in *Proceedings of 2017 IEEE/SICE International Symposium on System Integration (SII 2017)*, Taipei, Taiwan, Dec. 2017, pp. 1086–1091.

[7] Unity Technologies, "Unity," https://unity3d.com/, 2005.

[8] RT Corporation, "Sciurus17," https://rt-net.jp/products/sciurus17/, 2018.

[9] M. Quigley, B. Gerkey, K. Conley, J. Faust, T. Foote, J. Leibs, E. Berger, R. Wheeler, and A. Ng, "ROS: An Open-Source Robot Operating System," in *Proceedings of 2009 IEEE Workshop on Open Source Software*, Kobe, Japan, May 2009.

[10] P. Lamb, "ARToolKit5," https://github.com/artoolkit/ARToolKit5, 2017.

[11] M. Bischoff, "ROS#," https://github.com/siemens/ros-sharp, 2017.

[12] Intel Corporation, "OpenCV: Open Source Computer Vision Library," https://github.com/opencv/opencv, 2000.

[13] P. F. Alcantarilla, J. Nuevo, and A. Bartoli, "Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces," in *Proceedings of 24th British Machine Vision Conference (BMVC 2013)*, vol. 13, Bristol, United Kingdoms, Sept. 2013, pp. 1–11.