Common Reality: An Interface of Human-Robot Communication and Mutual Understanding

Fujian Yan, Vinod Namboodiri, and Hongsheng He(✉)

School of Computing, Wichita State University, Wichita, KS 67260, USA
hongsheng.he@wichita.edu

Abstract. An interface that can share effective and comprehensive mutual understanding is critical for human-robot interaction. This paper designs a novel human-robot interaction interface that enables humans and robots to interact by their shared mutual understanding of the context. The interface superimposes robot-centered reality and human-centered reality on the working space to construct a mutual understanding environment. The common-reality interface enables humans to communicate with robots through speech and immersive touching. The mutual understanding is constructed by the user's commands, localization of objects, recognition of objects, object semantics, and augmented trajectories. The user's vocal commands are interpreted to formal logic, and finger touching is detected and represented by coordinates. Real-world experiments have been done to show the effectiveness of the proposed interface.

Keywords: Human-robot collaboration · Speech recognition · Discourse representation structure · Interactive display

1 Introduction

The demands for robotic applications in unstructured environments are increasing. Robots are designed to work in different fields, such as assisting in medication [7], treating Autism [23], and supporting people's daily lives [17]. Efficient human-robot interaction (HRI) plays a vital role in helping robots and humans collaborate [2]. Compared with conventional robots that are pre-programmed in structured environments, social robots are expected to face a wider variety of tasks [21].

In the past decades, researchers were endeavoring to design human-robot interaction interfaces [10]. Previous work [5,9,16,19,22] has developed context-depended frameworks that enable robots to interact with humans by facial expressions with visual devices. These context-dependent interfaces are not suitable for a dynamic environment [12]. Several augmented reality (AR) based

This work is partially supported by NSF CMMI 2129113.

H. Li et al. (Eds.): ICSR 2021, LNAI 13086, pp. 319–328, 2021.
https://doi.org/10.1007/978-3-030-90525-5_27

interfaces have been proposed [3,8]. Those AR-based methods used markers to recognize objects in the environment. As the number of objects increases the required computation power increased as well [6]. Traditional AR applications deployed in robotics are focused on enhancing the reality of humans by superimposing user's goals on additional devices. Other information such as what robots have learned from user's commands and object semantics is missing.

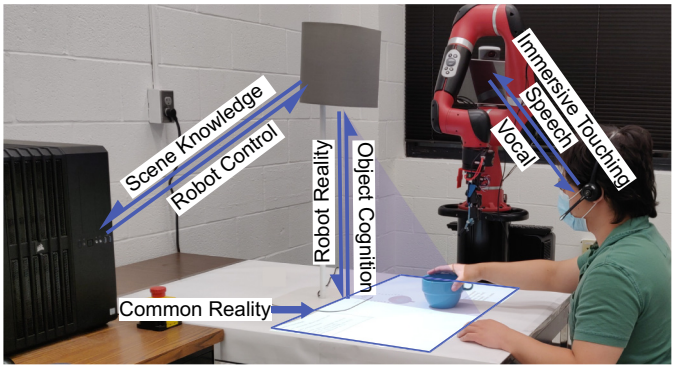

Fig. 1. Common reality for human-robot collaboration. The designed interface can take both speech and immersive touching from users. The shared knowledge, communication process, and the mutual understanding of the context are projected on the working space.

Thus, an HRI interface that enables robots and humans to share a mutual understanding of unstructured environments is urgently needed. In order to share a mutual understanding between robots and humans, the HRI interface needs to have the ability to dynamically recognize objects and understand the semantics of objects in the unstructured environment. It should provide a way that enables robots and humans to intuitively, effectively, and efficiently interact with each other.

In this paper, we propose a novel human-robot interface that can superimpose common reality for robots and humans. The designed common reality interface focuses on sharing a mutual understanding of the context to create an immersively interactive environment. The architecture of the common-reality interface is shown in Fig. 1. The common-reality interface supports immersive touching and speech while interacting with robots, which will avoid massive prior training for users. A mutual understanding of the working context is required for humans and robots to finish tasks collaboratively. The designed interface can detect and recognize objects in the working space by a deep learning model. By parsing the dictionary definitions of recognized objects, important attributes are extracted to construct a knowledge base by the language model. We choose to visualize the mutual understanding of the context instead of using a traditional question-answer manner because the ambiguity born with a natural language can hinder communicating in human-robot collaboration [4]. Also, information that

is visually presented is more intuitive than auditory [15]. The major contribution of this paper is designing an intuitive, effective, and efficient interface. The common-reality interface can visualize the common reality of a scene, thereby bridging the gap between human knowledge and the perception of robots.

2 Human-Robot Common Reality

As the number of robots deployed to a human-centered environment increases, the design of the HRI interface should not only understand the user's goals, but also demonstrate what the robot has understood from the user's goal [20]. In this paper, a novel HRI interface has been designed that combines multi-modal human-robot communication and augmented robot-human communication to ensure an intuitive, effective, and efficient interaction between robots and humans. The superimposed common reality contains user's commands and interpreted formal logic from user's commands. It also contains object localization and identification, augmented action, and object semantics. An illustration of the components of the common reality interface has been shown in Fig. 2. Humans can communicate with robots by integrating speech and immersive touching. It converts human's understanding of the context into a digital representation, which robots can understand. Robots perceive the objects and learn the semantics of the objects in the context. Then, robots communicate with humans by visualizing their understanding of context.

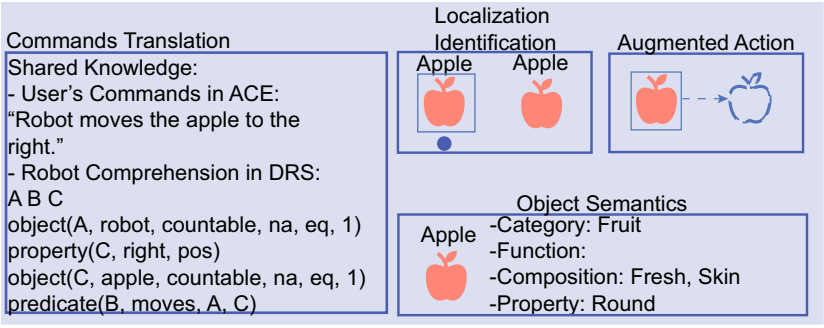

Fig. 2. Major components of the common reality interface.

2.1 Human-Robot Communication

To enable humans to communicate with robots naturally, the common-reality interface integrates multi-modal methods that include speech and immersive touching. Humans give commands through natural languages or touching the surface of the working environment.

DRS Translation. Robots need to have unambiguous, deterministic, and expressive instructions for executing actions. For robot understanding, we interpreted natural language into Discourse Representation Structures (DRS). We extended the lexicon to make it suitable for robotic applications. The parsed DRS results consist of referents and conditions. The referents are used to define semantic units that are embedded in the commands. These semantic units include subjective objects, objective objects, executable actions, and the condition of actions. The conditions are used to describe the relations of each parsed referent. According to part-of-text, dependency, and syntactical rules [11], there are four general declarations for covering three major HRI scenarios. The four general declarations are object declaration, predicate declaration, query declaration, and property declaration. The object declaration is used to describe the *NOUN* in user's commands. It refers to *subject* or *object* in the translated commands. The predicate declaration describes the executable actions, and the property declaration describes the condition of these actions. We design several robotic actions such as pick, move, and lift to ground the predicate to actual robotic actions. These actions can be added as the complexity of the tasks increase. An example of pared results for each scenario is shown in Fig. 3.

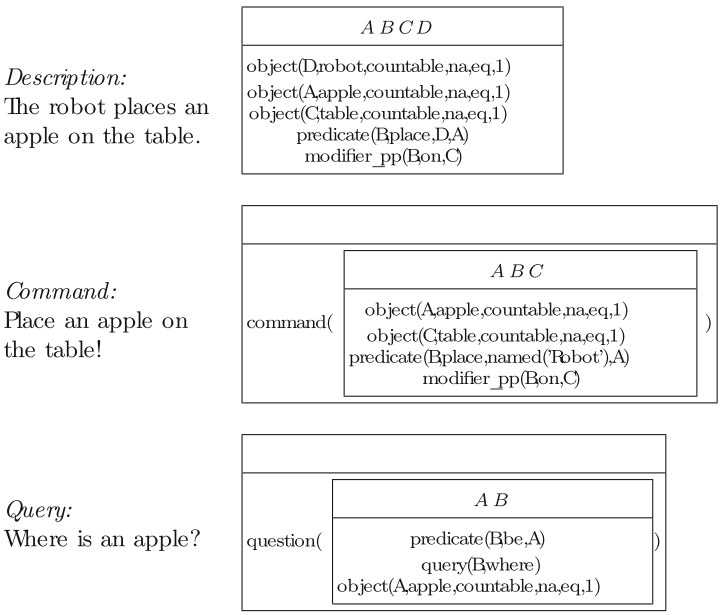

Fig. 3. Example of DRS Translation. Three fundamental HRI scenarios with parsed DRS results are shown.

Identification and Localization. In some scenarios, robots can not understand the user's purposes by only using vocal commands. Some objects may have the same texture, names, or characteristics in the same scene. For example, there are two apples in the scene. The command from the user is "Pick up an apple", the robot can not differentiate which apple it should pick up. In this case, additional instructions from users are needed to assist robots in localizing the target object. Robots need to percept the context by themselves to work with objects. To enable robots to percept the context, we used the Faster-R-CNN [18] model to detect and recognize objects. The inputs of the model are images of the context, and outputs are object labels. By referring to the labels, objects can be identified. To assist localization, the user can press an interactive button that associates with detected objects. The button is projected on the context. Users touch the buttons with their fingertips to interact with robots. To detect the fingertips of a hand, we used the model proposed in [1]. It is a convolutional neural network (CNN) that can take an image of a hand, and the outputs are coordinates of each recognized fingertips. A depth sensor is used to detect whether the user has touched the button or not.

2.2 Mutual Understanding

Humans have doubts about interacting with intelligent robots because they do not understand what robots will execute. The survey [13] has shown that 19% of 22% of participants fear intelligent robots because they do not understand. An interface that can share mutual understanding is needed. The common-reality interface can share human-robot communication, including the DRS translation, object identification, and object localization. It can also share action visualization and object semantics.

Visualizing the trajectory before the actions have been executed can help human users understand the robot's intention. The common-reality interface projects the trajectories of actions to improve mutual understanding. We used the MoveIt toolkit to plan the trajectory by giving the coordinate of the initial position and the final position. To project the trajectory in 3D space onto a 2D surface, we transform the robot's end effector in the world frame to the projector frame. To project the points in the projector frame to the $X - Y$ surface, we make the value on the $Z - axis$ equal to zero.

In order for robots to understand the semantics of the context, we used a language model [24] that can parse important attributes of objects from their dictionary definitions. These attributes include category, function, composition, and property. We used those attributes and the objects parsed from the input commands to form a dynamic knowledge base. To construct the knowledge base, we first translate the objects that are parsed from commands and learned attributes of objects into ACE sentences by a pre-defined template. We parsed these ACE sentences into DRS.

We used a Lampix projector to project these three components on a surface to construct an immersive interface contains a projector, a depth camera, and Raspberry Pi. Three major components can be visualized: the command that

robots have learned, the characteristics that are held by each object in the context, and the hypothetical trajectories that robots will execute. A depth camera embedded in the projector can detect the depth difference of objects and surfaces. Objects can be detected by using the depth difference between objects and the working surface. The movement-based segmenter can be achieved with this difference as well. By visualizing the context, redundant speech can be eliminated. Objects in the working space are detected by the depth sensor that is embedded in the projector. The input commands, parsed logic representations, and characteristics of objects are projected based on a pre-defined template. The animation of trajectory was shown based on the planned path of the robot motion at the pixel level.

3 Experiment

We evaluated the effectiveness of the designed common-reality interface by using four different cases in the real-world scenario. We evaluated the satisfaction level of the common-reality interface based on the questionnaire for different people.

3.1 Real-World Scenario

There were four trials of the sample results that were used to illustrate the working process of the interface. The results were illustrated in Fig. 4. The left column is the image taken from the workspace, and the right column is the sequence of images that were able to illustrate the movement. Commands were given to the robots as inputs. Both the characteristics of objects and the shared language in logic representation were projected.

3.2 User Satisfaction Evaluation

The robustness of the common-reality interface was evaluated based on the user's satisfaction regarding the demonstration of the system. The evaluation of the user's sanctification was measured based on questionnaires that were generated based on Lewis' After-Scenario Questionnaire (ASQ) [14]. The degree of satisfaction was evaluated with five different levels: very unsatisfied, unsatisfied, neutral, satisfied, and very satisfied. The satisfaction was evaluated based on three standards, which were readability, correctness, and intuitiveness.

There were six people taking part in this experiment. The experiment participants were from different levels of education. The age range of the experiment participants was from 20 to 50. There were ten interactive scenarios generated. The generated scenarios included the input commands from users and the shared language projected on the working space. Each experiment participant was asked to write the command of the projected knowledge, which was understood by the robot. Different participants evaluated the written commands, and those experiment participants were asked to fill the questionnaire based on the written commands.

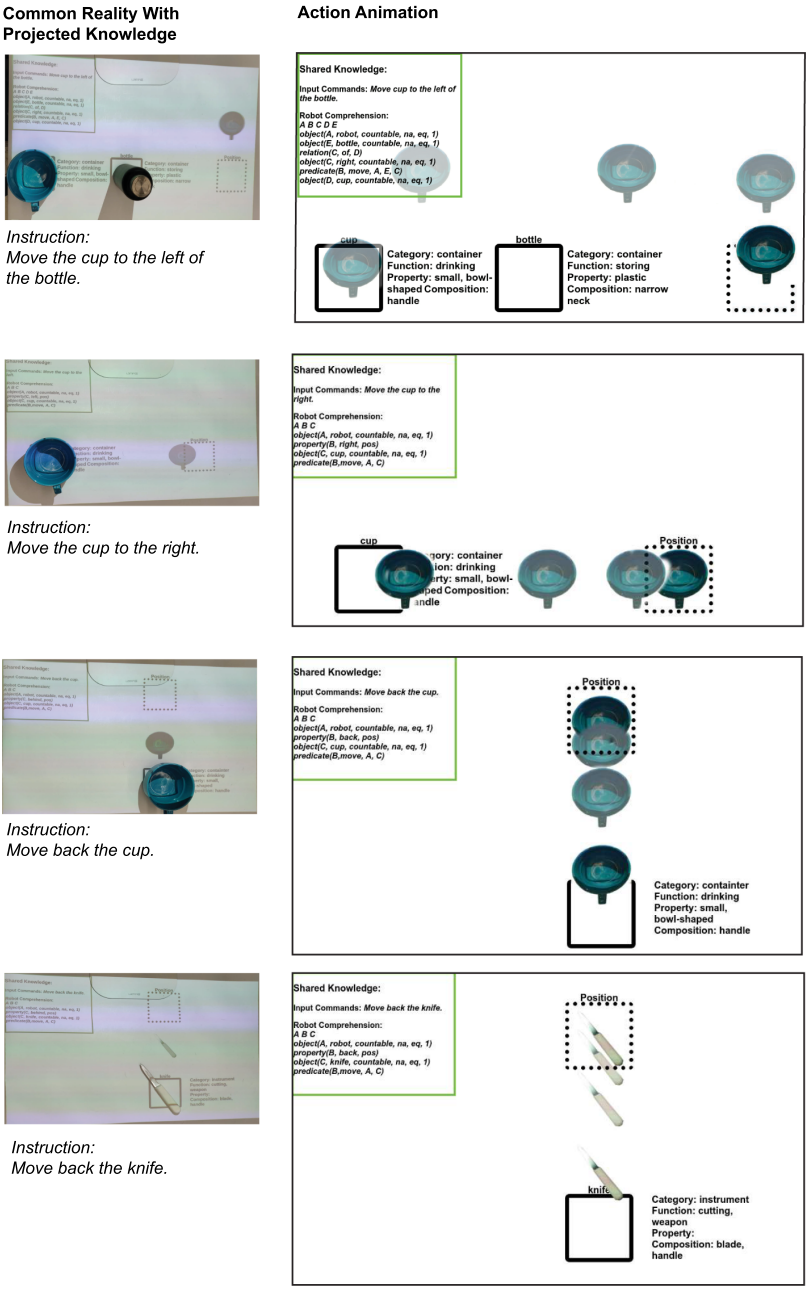

Fig. 4. Real-world action planning by using the common-reality interface. There were a total of four different trials shown. The left column is the common reality with projected knowledge, and the right column is the animation of the action.

Based on the questionnaire, no participant response "very unsatisfied" or "unsatisfied" to the proposed interface. We evaluated the average user's response to the proposed interface based on the questionnaire. We calculated the mean and standard deviation of each participant in each evaluation category. The results were shown in Fig. 5. The common-reality interface was satisfied based on the feedback of the questionnaire. Overall, most users who have taken the questionnaires are very satisfied with the proposed interface.

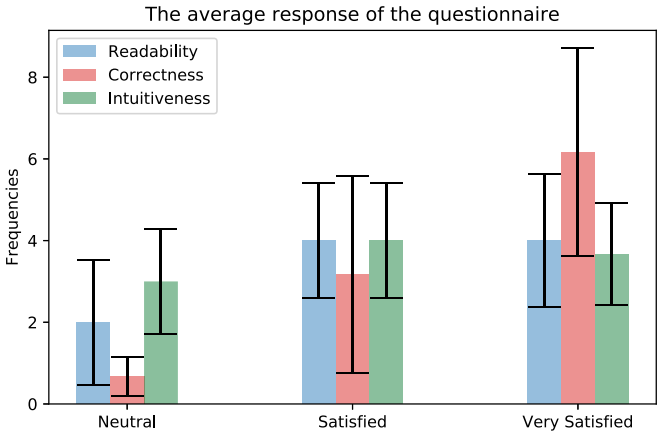

Fig. 5. Average response to the interface.

4 Conclusion

This paper presented a novel human-robot interaction interface, which enhances the HRI by superimposing the common reality including shared language, augmented semantics, and mutual understanding of the context. The mutual understanding of the context included the commands, the characteristics of objects, and the planned trajectory that were understood by robots. The results of the questionnaires have demonstrated the general acceptance of the common reality interface by users.

References

1. Alam, M.M., Islam, M.T., Rahman, S.: A unified learning approach for hand gesture recognition and fingertip detection. UMBC Student Collection (2021)
2. Bauer, A., Wollherr, D., Buss, M.: Human-robot collaboration: a survey. Int. J. Humanoid Rob. **5**(01), 47–66 (2008)
3. Bischoff, R., Kazi, A., Seyfarth, M.: The morpha style guide for icon-based programming. In: Proceedings. 11th IEEE International Workshop on Robot and Human Interactive Communication, pp. 482–487. IEEE (2002)
4. Branavan, S.R.K., Hackman, J.E., Heckel, F.W.P., Isaksen, A.: Updating natural language interfaces by processing usage data. US Patent 10,210,244, 19 Feb 2019

5. Breazeal, C.: Toward sociable robots. Rob. Auton. Syst. **42**(3–4), 167–175 (2003)
6. Chacko, S.M., Kapila, V.: An augmented reality interface for human-robot interaction in unconstrained environments. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 3222–3228. IEEE (2019)
7. Datteri, E.: Predicting the long-term effects of human-robot interaction: a reflection on responsibility in medical robotics. Sci. Eng. Ethics **19**(1), 139–160 (2013)
8. Fang, H., Ong, S.K., Nee, A.Y.: Novel AR-based interface for human-robot interaction and visualization. Adv. Manuf. **2**(4), 275–288 (2014)
9. Ge, S.S., et al.: Design and development of nancy, a social robot. In: 2011 8th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), pp. 568–573. IEEE (2011)
10. He, H., Ge, S.S., Zhang, Z.: A saliency-driven robotic head with bio-inspired saccadic behaviors for social robotics. Auton. Rob. **36**(3), 225–240 (2013). https://doi.org/10.1007/s10514-013-9346-z
11. Kamp, H.: Discourse representation theory. In: Blaser, A. (ed.) IBM 1988. LNCS, vol. 320, pp. 84–111. Springer, Heidelberg (1988). https://doi.org/10.1007/3-540-50011-1_34
12. Katz, D., Kenney, J., Brock, O.: How can robots succeed in unstructured environments. In: In Workshop on Robot Manipulation: Intelligence in Human Environments at Robotics: Science and Systems. Citeseer (2008)
13. Ledbetter, S.: The chapman university survey on american fears. https://blogs.chapman.edu/wilkinson/2015/10/13/americas-top-fears-2015/
14. Lewis, J.R.: Psychometric evaluation of an after-scenario questionnaire for computer usability studies: the asq. ACM Sigchi Bull. **23**(1), 78–81 (1991)
15. Lindner, K., Blosser, G., Cunigan, K.: Visual versus auditory learning and memory recall performance on short-term versus long-term tests. Mod. Psychol. Stud. **15**(1), 6 (2009)
16. Littlewort, G., et al.: Towards social robots: Automatic evaluation of human-robot interaction by facial expression classification. In: Advances in Neural Information Processing Systems, pp. 1563–1570 (2004)
17. Nieto, D., Quesada-Arencibia, A., García, C.R., Moreno-Díaz, R.: A social robot in a tourist environment. In: Hervás, R., Lee, S., Nugent, C., Bravo, J. (eds.) UCAmI 2014. LNCS, vol. 8867, pp. 21–24. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-13102-3_5
18. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
19. Scheeff, M., Pinto, J., Rahardja, K., Snibbe, S., Tow, R.: Experiences with sparky, a social robot. In: Dautenhahn, K., Bond, A., Cañamero, L., Edmonds, B. (eds.) Socially Intelligent Agents. Multiagent Systems, Artificial Societies, and Simulated Organizations, vol 3. Springer, Boston (2002). https://doi.org/10.1007/0-306-47373-9_21
20. Sciutti, A., Mara, M., Tagliasco, V., Sandini, G.: Humanizing human-robot interaction: on the importance of mutual understanding. IEEE Technol. Soc. Mag. **37**(1), 22–29 (2018)
21. Thrun, S.: Toward a framework for human-robot interaction. Hum. Comput. Interact. **19**(1–2), 9–24 (2004)
22. Turkle, S., Breazeal, C., Dasté, O., Scassellati, B.: Encounters with kismet and cog: children respond to relational artifacts. Digit. Media Transformations Hum. Commun. (2006)

23. Wood, L.J., Zaraki, A., Robins, B., Dautenhahn, K.: Developing kaspar: a humanoid robot for children with autism. Int. J. Soc. Rob. 1–18 (2019)
24. Yan, F., Tran, D.M., He, H.: Robotic understanding of object semantics by referring to a dictionary. Int. J. Soc. Rob. 1–13 (2020)