Developing Evaluation Criteria for Interaction Quality in Guiding Daily Life Activities

Yoshiaki Mizuchi
Tamagawa University
Tokyo, Japan
mizuchi@eng.tamagawa.ac.jp

Tetsunari Inamura
Tamagawa University
Tokyo, Japan
inamura@lab.tamagawa.ac.jp

ABSTRACT

Interactive service robots must be able to effectively communicate with users when teaching work procedures and seeking assistance in case of errors. To develop and enhance such interaction capabilities for service robots, the evaluation of interaction quality is crucial. However, assessing interaction quality in daily life scenarios poses significant challenges, primarily due to its subjective nature. In this paper, we present case studies on the evaluation of interaction quality in the context of guiding daily life activities.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Virtual reality**; • **General and reference** → **Metrics**; **Evaluation**.

KEYWORDS

human-robot interaction, interaction quality, evaluation metrics, virtual reality

ACM Reference Format:
Yoshiaki Mizuchi and Tetsunari Inamura. 2023. Developing Evaluation Criteria for Interaction Quality in Guiding Daily Life Activities. In *Workshop at the HAI'23.* ACM, New York, NY, USA, 2 pages.

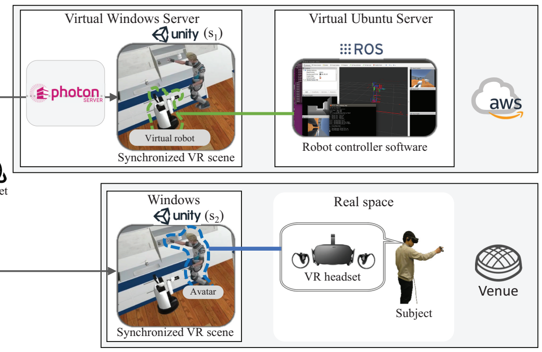

Figure 1: VR platform for online HRI competition.

1 INTRODUCTION

In our daily lives, we frequently find ourselves in situations where we need to convey information, procedures, and tasks to others. This ability to effectively interact with people is crucial for interactive service robots, enabling them to teach work procedures, share roles in cooperative tasks, and request assistance when encountering errors. The ability to evaluate interaction quality, based on the robot's behavior and human responses, holds promise for enhancing robot decision-making and fostering the development of more sophisticated interactive intelligence.

However, evaluating interaction quality in the context of daily life scenarios poses several notable challenges. These include (i) the time and cost associated with conducting interaction experiments to collect data for analysis and evaluation, (ii) the ambiguity surrounding which evaluation metrics humans prioritize when guided by robots in their actions, and (iii) the inefficiency of questionnaire-based subjective evaluations, especially in situations requiring prompt results.

To address these challenges, we have proposed a Human Navigation competition task in which a robot has to guide human behavior in a VR scene. Furthermore, we investigated evaluation metrics for interaction quality and a determination process of criteria for fair and efficient evaluation. In this paper, we present case studies and examine the potential of interaction quality evaluation.

HAI '23, December 4–7, 2023, Gothenburg, Sweden
2023.

2 INTERACTION QUALITY EVALUATION

2.1 Cloud-based VR Platform

To efficiently conduct multimodal interaction experiments that may be challenging to perform in real-world environments, we have introduced a cloud-based virtual reality (VR) platform called SIGVerse [2]. This VR platform has been utilized to conduct the robot competition, collect behavioral data, and reproduce the interaction data.

Figure 1 illustrates the software configuration utilized in an online competition [6]. The Robot Operating System (ROS) software controls the virtual robots, while control commands and emulated sensor data are exchanged between Unity and ROS in real-time [3]. Test subjects log in as avatars within a virtual environment using VR headsets and hand-tracking controllers. They can move around the virtual space using hand-controller joysticks and interact with virtual objects through head-mounted displays and hand-tracking controllers. By synchronizing events across each VR environment, such as avatar body movements, robot actions, speech interactions, and object interactions, over the internet, robots and test subjects in different locations can engage in interactive experiences.

2.2 Task Design

The interaction scenario is designed for a robot competition task called Human Navigation [1], in which a virtual robot is tasked with guiding a non-expert human user to locate a target object and transport it to a designated destination. Figure 2 shows a screenshot

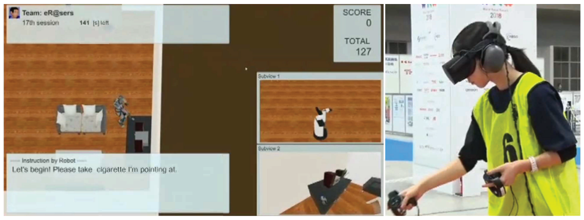

Figure 2: Competition system and a test subject in the Human Navigation task.

of a competition system and a test subject in the Human Navigation task.

The virtual robots play the role of generating natural language and gesture instructions to guide human behavior, while human users (i.e., test subjects) are responsible for following the robot's instructions. Prior to each session, the test subject is not provided with information about room layouts, the target object, or the destination. At the beginning of each session, only the robot is informed about the positions and orientations of furniture, graspable objects, the target object, and the destination area. Subsequently, the virtual robot must generate instruction sentences and gestures based on this positional information. Therefore, if the instructions given by the robot are vague, the human may become disoriented and fail to complete the task. Consequently, the interaction quality can be evaluated based on the interaction behavior exhibited by both the human and the robot.

2.3 Designing Evaluation Metrics

Interaction quality should be evaluated based on human subjectivity and perception, rather than solely on the time taken or the level of achievement. However, the evaluation metrics that humans prioritize when robots guide their behavior have not been thoroughly examined. To address this gap, we conducted an investigation into the metrics that humans use to evaluate interaction quality [5].

In this study, we used interaction behavior data of different quality collected through three approaches: human-robot interaction involving instruction generators developed for a robot competition, unconstrained oral human-human interactions, and constrained human-human interaction using a GUI-based instruction generator. To identify factors that humans prioritize to score the interaction quality, we asked the evaluators to review playback videos of the collected interaction data and to answer what they based their evaluation of the interaction quality on in an open-ended form. We organized, reduced, and merged the initial items into 17 evaluation items. Subsequently, we requested evaluators to rate interaction data based on the provided items and performed a factor analysis.

This study revealed four primary evaluation metrics:

- The robot's ability to guide human users according to their behavior,
- The appropriateness of explanations regarding features and attributes
- The conciseness of instructions
- The frequency of instructions and the amount of information provided

2.4 Formulation of Evaluation Criteria

The conventional approach to evaluating interaction quality involves subjective evaluation through the use of questionnaires. However, subjective evaluation is often time-consuming and may not be practical in situations requiring prompt results. To address this limitation, we explored a method for formulating evaluation criteria that can approximate subjective evaluation results using interaction behavior data [4].

This study has demonstrated that it is possible to reasonably approximate subjective scores for interaction quality by considering explanatory variables derived from interaction behavior data.

3 CONCLUSIONS

In this paper, we have presented a series of case studies on interaction quality evaluation within the framework of the 'Human Navigation' competition task, where robots provide guidance for daily life activities.

Recent advancements in the field of large language models have significantly improved the generation of instructions. While challenges related to the accuracy of spatial information and real-time responsiveness to human behavior still exist, generating instructions that effectively guide everyday actions is becoming increasingly attainable. However, even with these advanced instruction generation capabilities, the necessity of conducting reasonable and efficient evaluations of interaction quality remains paramount to ensure that interactions align with the user's subjective experience.

While our previous work relied on average evaluator ratings, it is essential to acknowledge that the prioritization of evaluation metrics varies among users. From the perspective of interaction optimization, introducing personalized assessments tailored to each user's individual preferences becomes a critical consideration for the advancement of this field.

ACKNOWLEDGMENTS

This work was supported by JST [Moonshot R&D][Grant Number JPMJMS2034], a project JPNP20006 commissioned by the New Energy and Industrial Technology Development Organization (NEDO), and ROIS NII Open Collaborative Research 2023-(23FP05).

REFERENCES

[1] Tetsunari Inamura and Yoshiaki Mizuchi. 2018. Competition design to evaluate cognitive functions in human-robot interaction based on immersive VR. In *RoboCup 2017: Robot World Cup XXI, Lecture Notes in Artificial Intelligence*, Vol. 11175. 84–94.

[2] Tetsunari Inamura and Yoshiaki Mizuchi. 2021. SIGVerse: A Cloud-Based VR Platform for Research on Multimodal Human-Robot Interaction. *Frontiers in Robotics and AI* 8 (2021), 158. https://doi.org/10.3389/frobt.2021.549360

[3] Yoshiaki Mizuchi and Tetsunari Inamura. 2017. Cloud-based multimodal human-robot interaction simulator utilizing ROS and unity frameworks. In *IEEE/SICE International Symposium on System Integration*. 948–955.

[4] Yoshiaki Mizuchi and Tetsunari Inamura. 2020. Optimization of criterion for objective evaluation of HRI performance that approximates subjective evaluation: a case study in robot competition. *Advanced Robotics* 34, 3-4 (2020), 142–156. https://doi.org/10.1080/01691864.2019.1698462

[5] Yoshiaki Mizuchi, Yusuke Tanno, and Tetsunari Inamura. 2023. Designing Evaluation Metrics for Quality of Human-Robot Interaction in Guiding Human Behavior. In *Proceedings of the International Conference on Human-Agent Interaction.*

[6] Yoshiaki Mizuchi, Hiroki Yamada, and Tetsunari Inamura. 2023. Evaluation of an online human-robot interaction competition platform based on virtual reality – case study in RCAP2021. *Advanced Robotics* 37, 8 (2023), 510–517. https://doi.org/10.1080/01691864.2022.2145235