Analysis of Robot Errors in Social Imitation Learning

Joshua Ravishankar
Graduate School of Informatics
Kyoto University
jravishankar@robot.soc.i.kyoto-u.ac.jp

Malcolm Doering
Graduate School of Informatics
Kyoto University
doering@i.kyoto-u.ac.jp

Takayuki Kanda
Graduate School of Informatics
Kyoto University
kanda@i.kyoto-u.ac.jp

***Abstract*—Data-driven imitation learning is a method that leverages human-human interaction data to effectively generate robot behaviors for human-robot interaction. However, interaction errors can occur that cause interaction breakdowns. Furthermore, these interaction errors do not occur in human-human interactions, and thus, the behavior generation model is left with no behaviors to imitate in order to effectively recover. To the end of building a robust error handling pipeline to facilitate interaction recovery, in this work we analyze error types in social imitation learning for human-robot interaction (HRI). We focus on two specific robot behavior generation systems: one with data abstraction and one without data abstraction. We categorize frequently occurring interaction errors from these systems into categories and summarize the resulting interaction patterns. Many of these errors lead to reduced interaction quality and sometimes lead to frustration and/or confusion in humans. Finally, we conclude that the existence of such errors necessitates an autonomous error detection and online interaction recovery method.**

***Index Terms*—human-robot interaction, errors in data-driven imitation learning, social robotics**

I. Introduction

The data-driven imitation learning approach has been increasingly explored as a method for endowing multi-modal, mobile robots with human-like interaction behaviors [5] [6] [7]. This approach is effective because it requires minimal input from humans for interaction design and data annotation and can learn robust behavior from natural interaction examples, which contain both sensor noise and natural human speech and behavior variation.

When employing this approach, there is a strong assumption that the speech and action trajectories learned from human-human interaction can serve as a valid basis for all possible interactions that a robot might encounter, enabling the robot to identify an appropriate response to any form of human behavior. For example, in a camera store scenario, using a broad range of human customer behaviors (e.g., window shopping, asking questions about specific camera features, etc.) and corresponding human shopkeeper responses as training data should lead to a robust, comprehensive robot shopkeeper.

M. Doering and T. Kanda are also with Advanced Telecommunications Research Institute International (ATR). This work was supported by JST, AIP Trilateral AI Research, Grant Number JPMJCR20G2, Japan.

However, through our analysis, we find that errors characterized by inappropriate or incomprehensible robot behavior occur, negatively affecting interaction quality and sometimes making the human customer frustrated or confused. These errors necessitate error recovery mechanisms in social imitation learning and data-driven HRI.

Previous work has classified trust-relevant failures in HRI broadly [2]. However, in this work, we focus on classifying system errors that appear in data-driven, social imitation learning specifically, drawing from real human-robot interactions in a camera store scenario. We contribute to the field of data-driven HRI by analyzing and categorizing interaction errors that occur in a previous study by Liu et al. [1]. Through this analysis, we aim to illuminate some types of errors that can occur when training a system on human-human interaction data. Though related research in HRI and natural language processing (NLP) has sought detection and recovery mechanisms for dialogue errors [8] [9], this work serves as the foundation towards the goal of autonomous detection of erroneous robot behaviors (both speech and locomotion) that does not require human input (manual labeling, etc.).

II. Dataset

We analyze a dataset of human-robot interactions that was produced in [1]. This dataset consists of interactions between hired participants role-playing as customers and a robot trained using imitation learning to perform the role of a shopkeeper within a camera store. The layout of the camera store can be seen in Figure 1. There were two systems trained to generate robot behaviors: a baseline system and a proposed system (details about each system will be explained in the following paragraphs).

In total, there were 17 participants who participated in two sets of 8 trials (one set per system) playing the following roles: a need-based customer, who is looking for a camera with a specific feature (3 trials), a curious customer, who is interested in multiple cameras (3 trials), and a window-shopping customer, who prefers to browse the store alone (3 trials), for a total of 272 interactions. Each interaction lasted anywhere from 1-5 minutes, and participants communicated with the robot using automatic speech recognition (ASR) through an Android phone interface. The resulting human-robot interaction dataset was multimodal, consisting of video,

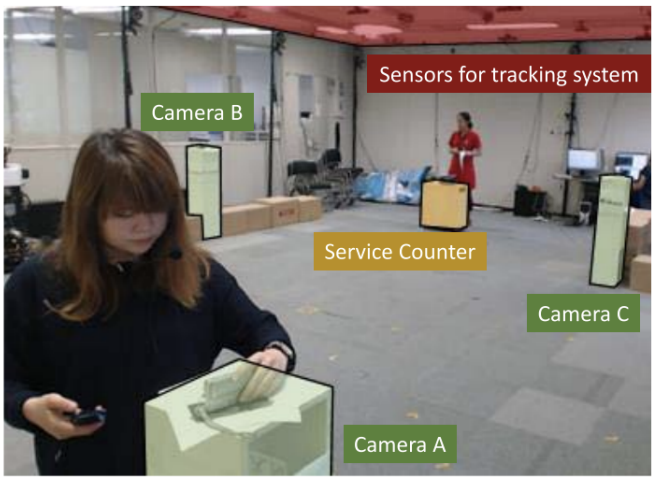

Fig. 1. Environment setup for HRI/HHI data collection in [1], featuring three camera displays. Sensors on the ceiling were used for tracking human position, and smartphones carried by the participants were used to capture speech. Figure reprinted from [3].

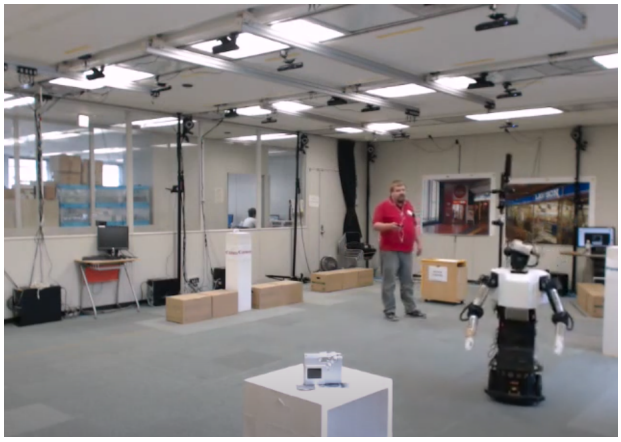

Fig. 2. An example of the "Incomprehensible Movement" error type. The robot does not verbally respond to and acknowledge the customer's greeting, so the customer stays in place next to the service counter while the robot moves to a camera.

audio, speech (ASR), and location data. This data was saved to a cloud database, from where it can be synchronized and replayed using a Java-based GUI tool. We use this GUI to perform our analysis in III.

The robot shopkeeper was trained using human-human interaction data in the same camera store setting. Each interaction had two participants, one role-playing as the customer and one role-playing as the shopkeeper. Before each interaction, customers were given a role to play (need-based, curious, or window-shopper). The shopkeeper was not informed of the customer's role, and was told to allow the customer to browse, answer any questions that arose, and introduce products when appropriate. To make the interactions repeatable, participants were instructed to restrict the scope of the interactions to focus on information about the cameras, avoiding any other topics (e.g. bartering over price). The resulting human-human interaction dataset was then used to train a proposed and baseline system to generate robot behaviors through imitation learning.

The proposed system uses data abstraction techniques which learn common behaviors in the interaction and reduce the dimensionality of the data, making the learning problem more tractable. These data abstraction techniques include speech clustering, motion clustering, and models of proxemics formations (which serve to capture the fact that human participants spend most of the interaction time in a few static spatial formations). A naive-Bayesian classifier is trained to take the abstracted, vectorized interaction state (which consists of speech, location, and motion information for both the customer and the shopkeeper) as input and produce the corresponding shopkeeper action vector (which is composed of an utterance and target spatial formation) as its target.

The baseline system does not use any forms of data abstraction, instead preserving the human-human interaction data as raw feature vectors consisting of raw x,y location data and unclustered vector representations of speech (using Latent Semantic Analysis). The baseline system generates robot speech and locomotion using a nearest-neighbor predictor as in [4]. During inference, whenever a customer action is detected, its raw feature vector is compared to all feature vectors from the training data. Once the best match is found, the subsequent shopkeeper action (comprising of speech and movement) from the training data is used as the robot action.

In our analysis, we do not separate the errors made by the two systems, and instead form a holistic set of categories that exemplify errors that can occur in *any* system that is trained on human-human interaction data.

III. Analysis

We analyze error types that can occur in data-driven, imitation-learning-based HRI in a customer-shopkeeper context. We define an error as any robot behavior that is socially unacceptable given the interaction context and leads to unnatural interaction patterns between the customer and (robot) shopkeeper. Acceptable robot behavior does not necessarily have to fall into the set of normal human shopkeeper behavior in the same scenario but must be deemed appropriate from the lens of a human observer.

To learn the types of errors that commonly occur, we parsed through the human-robot interaction dataset described in II and referred to evaluations made by a coder in [1] to identify a particular behavior as bad. The coder was asked to examine each action (speech or movement) made by the human participant, and to judge whether the robot's response to that action was appropriate. During our analysis, we identified reoccurring error types that regularly resulted in unnatural interaction patterns, grouping them based on the source of the error, the phase of the interaction in which they occurred (e.g when the customer initially enters, while the customer is asking questions, while the customer is browsing, etc.), and

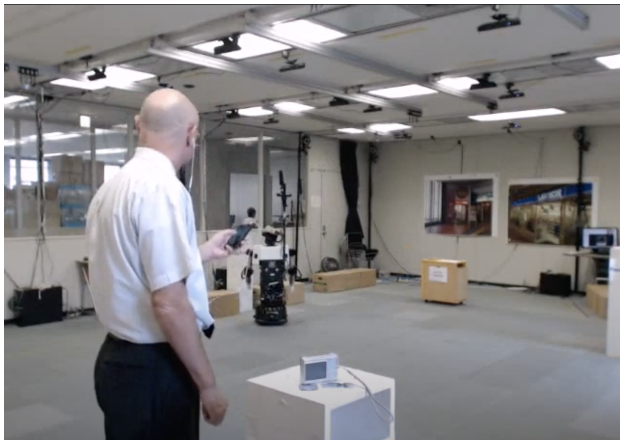

Fig. 3. An example of the "Misaligned Location" error type. The customer is positioned at the camera in the foreground hoping to ask the robot a question, but the robot is positioned near a different camera in the background.

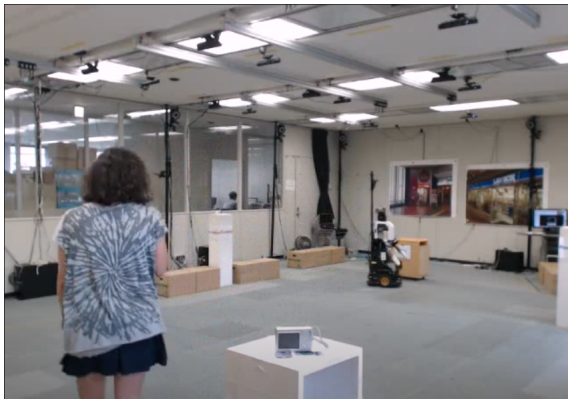

Fig. 4. An example of the "Sudden Disengage" and "Preemptive Goodbye" error types occurring in sequence. The customer is positioned at the camera in the foreground hoping to ask the robot a question, but the robot disengages. As the robot moves back to the service counter, it preemptively says "Thank you for coming", prompting the customer to leave the store with remaining unanswered questions.

the subsequent customer reactions. We also found that some of the error types can occur in tandem as shown in Figure 4.

We differentiate six different frequently occurring error types with respect to the robot behavior: (1) Incomprehensible Movement, (2) Repeated Greeting, (3) Preemptive Goodbye, (4) Misaligned Location, (5) Sudden Disengage, and (6) Incorrect Response.

Incomprehensible Movement. The customer enters the store, greets the robot, and asks the robot a question, but the robot does not respond. The robot then moves to a camera. The customer, confused by the lack of verbal response, either hesitantly follows the robot or stays in place. An example of this error is shown in Figure 2.

Repeated Greeting. The customer enters the store, and then the robot greets the customer and offers help. After a brief interaction, while the customer is browsing the store, the robot once again greets the customer and offers help. This greeting may be repeated three or four times despite the customer politely declining. We found that this behavior most frequently occurs when the customer plays the role of a window-shopper, but it can also occur when the customer returns to a mode of browsing after asking the robot questions.

Misaligned Location. The customer is standing nearby a particular camera (e.g., Sony) and asks a question about it. The robot moves to a different camera (e.g., Panasonic) and begins answering the question. The customer is often confused about which camera the robot is providing information about. Subsequently, the customer will continue asking questions despite the confusion, move to the robot's location, or move to a new target camera (e.g., Canon). The third type of customer response may be an attempt to reset the interaction and give the robot an opportunity to align itself with the customer. An example of this error is shown in Figure 3.

Preemptive Goodbye. The customer thanks the robot for answering questions during an exchange, and then moves to a new target camera. The robot preemptively says goodbye to the customer. In some cases, this prompts the customer to leave the store even though they were still browsing. We believe that the robot misinterprets the customer's "thank you" as a signal that the customer is leaving the store and/or makes an error in motion target estimation, believing the customer is heading for the door to exit the store. An example of this error is shown in Figure 4.

Sudden Disengage. The customer and robot are at the same target camera. The customer thanks the robot for answering a question or asks the robot a question about the target camera's features. The robot disengages from the customer. We believe that the "thank you" may trigger the robot to return to the service counter, while the question might trigger the robot to move to and introduce a new camera (instead of answering the question about the current target camera). In both instances, the robot disengages from the customer even though the customer may still have questions. An example of this error is shown in Figure 4.

Incorrect Response. The customer asks the robot a question about a particular camera's features. The robot either answers incorrectly or is verbally unresponsive. This leads the customer to either ask the same question again, ask a new question, or disengage from the current target camera.

IV. Conclusion and Future Work

In this paper, we aimed at consolidating the erroneous interaction patterns we have discovered in data-driven HRI for a shopkeeper robot to identify bottlenecks in data-driven imitation learning. Our efforts motivate the need for an autonomous error detection and interaction recovery system. We believe that human reactions to robot errors may provide a valuable signal both for error detection and some form of

online interaction recovery, and in future work we aim to build such a system, perhaps using a reinforcement learning framework.

REFERENCES

[1] P. Liu, D. F. Glas, T. Kanda, and H. Ishiguro, "Data-Driven HRI: Learning Social Behaviors by Example from Human-Human Interaction", IEEE Transactions on Robotics, vol. 32, pp. 988-1008, 2016.

[2] S. Tolmeijer, A. Weiss, M. Hanheide, F. Lindner, T. M. Powers, C. Dixon, M. L. Tielman, "Taxonomy of Trust-Relevant Failures and Mitigation Strategies", HRI '20: Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction, pp. 3-12, 2020.

[3] P. Liu, D. F. Glas, T. Kanda, and H. Ishiguro, "Two Demonstrators Are Better Than One—A Social Robot That Learns to Imitate People With Different Interaction Styles", IEEE Transactions on Cognitive and Developmental Systems, vol. 11, pp. 319-333, 2019.

[4] H. Admoni and B. Scassellati, "Data-Driven Model of Nonverbal Behavior for Socially Assistive Human-Robot Interactions", ICMI '14: Proceedings of the 16th International Conference on Multimodal Interaction, pp. 196–199, 2014.

[5] M. Doering, B. Drazen, and T. Kanda, "Data-Driven Imitation Learning for a Shopkeeper Robot with Periodically Changing Product Information", ACM Transactions on Human-Robot Interaction, vol. 10, pp. 1-20, 2021.

[6] M. Doering, P. Liu, D. F. Glas, T. Kanda, D. Kulić, and H. Ishiguro, "Curiosity Did Not Kill the Robot: A Curiosity-based Learning System for a Shopkeeper Robot", ACM Transactions on Human-Robot Interaction, vol. 8, pp. 1-24, 2019.

[7] P. Liu, D. F. Glas, T. Kanda and H. Ishiguro, "Two Demonstrators Are Better Than One—A Social Robot That Learns to Imitate People With Different Interaction Styles," IEEE Transactions on Cognitive and Developmental Systems, vol. 11, no. 3, pp. 319-333, 2019.

[8] T. Uchida, T. Minato, T. Koyama, and H. Ishiguro, "Who Is Responsible for a Dialogue Breakdown? An Error Recovery Strategy That Promotes Cooperative Intentions From Humans by Mutual Attribution of Responsibility in Human-Robot Dialogues", Frontiers in Robotics and AI, vol. 6, 2019.

[9] R. Meena, J. Lopes, G. Skantze, and J. Gustafson, "Automatic Detection of Miscommunication in Spoken Dialogue Systems", Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 354-364, 2015.