Towards Imitation Learning of Human Interactive Motion

Yongqiang Jiang
Graduate School of Informatics
Kyoto University
yongqiang@robot.i.soc.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*—Current imitation learning methods for designing robot motion are limited because they rely on human input of environment information (e.g. objects indicated by deictic gestures), which is time-consuming when designing robot motions for real-world scenarios. To solve this problem, we propose a novel method for finding the references of pointing gestures by using heat map information extracted from motion and speech clustering of human-human interaction data. To develop this method we setup an array of skeleton-sensing depth sensors to record human motion during natural interaction in an in-lab camera shop scenario. The results show that our method can find the positions of the objects being pointed to (cameras in the camera shop). Eventually, we aim to design a robot system that automatically learns to imitate all socially appropriate interactive motions (gestures, body postures, etc.).**

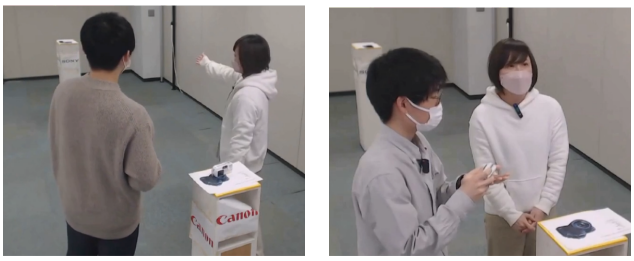
***Index Terms*—human-robot interaction, data-driven learning, imitation learning, gesture, social interaction**

I. Introduction

Humans use many types of interactive motions for effective social interaction [1], such as gestures and body postures, which contain important social information (Fig. 1). Therefore, social robots should also be able to perform socially appropriate interactive motions; however, current state-of-the-art techniques for designing robot motions are labor-intensive, which limits the possibility of applying them to social robots deployed in the real world. In contrast to previous approaches, we aim to design a system that automatically learns interactive motions, and the logic for when it is socially appropriate to perform a motion, from human-human interaction data collected via a passive sensor network, without manual data labeling and manually designed motions.

In this paper, we present our work in progress on learning interactive motions, specifically focusing on the problem of learning pointing gestures and their references (which objects in the room are being pointed at). Deictic gestures are unique because not only must the motion of the gesture be learned, but also where in the surrounding environment they are referring to. Previous approaches to imitation learning of gestures require this information to be specified and input ahead of time [2, 3], but we propose a technique using motion clustering and pointing heat maps to identify the pointed targets automatically.

Y. Jiang, M. Doering, and T. Kanda are also with the Advanced Telecommunications Research Institute International (ATR). This work was supported in part by JST, AIP Trilateral AI Research, Grant Number JPMJCR20G2, Japan (problem formulation) and in part by JST CREST under Grant Number JPMJCR17A2, Japan (data collection).



(a) (b)

Fig. 1: Interactive motions used in a camera shop. (a) The shopkeeper used a pointing gesture to provide position information. (b) The shopkeeper stood in a respectful posture to show a positive attitude towards the customer.

Furthermore, in this work, for the purpose of a proof of concept, we collected a new dataset of human-human interactions in an in-lab camera shop scenario via a passive sensor network, where participants role-played as customers and shopkeepers. Using this dataset we demonstrate our automatic pointing gesture reference identification technique and applied motion clustering to learn a variety of other interactive motions. This work focuses on the unique problems of learning to imitate pointing gestures, but the eventual goal is to develop a generalized method for finding environment references for all kinds of interactive motions and for a robot to imitate these interactive motions in real-time human-robot interaction.

II. Related Works

A. Human-designed Motion Model

In the field of social robotics, many studies have used manually designed motion models on their robot to accomplish complex interaction tasks such as models for deictic gestures [4, 5, 6], head movement [7, 8], and arm movement [3, 9]. Among these works, Okuno et al. [4] designed pointing gestures for a navigation robot by modeling the speech, motion, and timing of human behaviors from observation. Liu et al. [2] discussed the difference between casual pointing and precise pointing and designed a robot with polite deictic gestures. Huang and Mutlu [3] modeled four different types of gestures

for a teaching robot, which are able to refer to pictures projected on the wall.

However, as these studies manually design their models, it would require huge human effort to apply them in the real world, where novel, unanticipated interactive motions may be used. Furthermore, when referring to objects or directions from the environment with gestures like pointing, they require human input of the environment information, such as the position of the object. In contrast, in our less labor-intensive approach, we use motion clustering and a novel reference finding method to automatically learn motions and their reference in the environment from data without extensive human input.

B. Data-driven Motion Model

Data-driven motion models generate motions for robots by learning from multi-modal data. Yoon et al. [10] proposed a model to generate human upper body gestures from a given speech text, and their model is trained on human gesture data from online videos. The model designed by Plappert et al. [11] could output whole-body motions and be trained from motion capture data. Shlizerman et al. [12] presents their model for output skeleton movement of playing instruments for music audio input by training with music playing videos.

In contrast to our work, these related works are designed for single-speaker scenarios. Their training data and implementations do not include the complex turn-taking interaction context we may have in our target interaction scenario. Our essential goal is to learn the interaction logic of performing interactive motions in human-robot interactions. Moreover, the related works do not focus on the problem of using motions to refer to objects or directions in the environment, which is the main focus of this paper.

III. Data Collection

The first step of the imitation learning approach is to collect human-human interaction data in a target scenario. Ideally, data for imitation learning should be collected in the real world, in situations where the robot will eventually be deployed. But, interactions in the real world are complex, which would require great quantities of data to train the system. Furthermore, depth sensors and microphones are still not precise enough to record data in the real world while maintaining the quality and naturalness of data. Therefore, to replicate real-world data as closely as possible, and to provide a proof of concept, we setup a data-collection sensor network in the lab to collect human-human interaction data.

The sensor network was set up in a 7 m x 8 m room with some commercially available sensors, including web cameras, Kinect Azure depth sensors, and blue-tooth microphones. Fig. 2 shows the layout of our experiment room and sensor network. We set up 15 Kinect sensors to cover the whole room from different angles so that participants were allowed to move anywhere in the room. The passive skeleton sensing sensor network does not need the participants to hold or wear sensor equipment so they were free to perform interactive motions naturally. The human motion data was recorded as

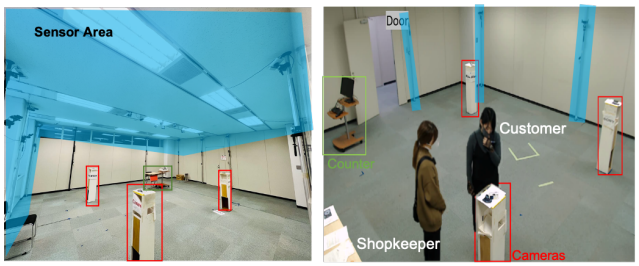


Fig. 2: Layout of the experiment room and our sensor network. Our sensors are in the blue areas in the pictures.

sequences of joint positions (skeletons) (which could eventually be mapped to robot joint positions for imitation). The use of the blue-tooth microphone audio recording system with interchannel suppression [13] and online speech recognition system (Google Cloud[1]) allow us to record speech data without requiring the participants' hands and leave them free to perform interactive motions. And the speech data was recorded as utterances (text) which could be used as robot speech after processing in the imitation learning approach.

Our target scenario is a camera shop, in which participants playing the roles of customer and shopkeeper interact using a variety of interactive motions. As shown in Fig. 2, we set up three cameras on pedestals in the room and a service counter near to the door. We hired actors to perform one-to-one customer-shopkeeper interactions. Three participants, who had prior experience in customer service, took turns role-playing as shopkeepers. We recruited 30 participants (aged from 19 to 60, 13 male and 17 female) to role-play as customers.

During the data collection, each customer participant role-played 12 interactions with one of the shopkeeper participants, in which they were allowed to ask about the cameras in the room. Furthermore, they role-played different customer types to collect data in a variety of situations. Finally, 394 trials of interaction data (mean 8 min., SD 3.5 min.; 3170 min. total) were collected.

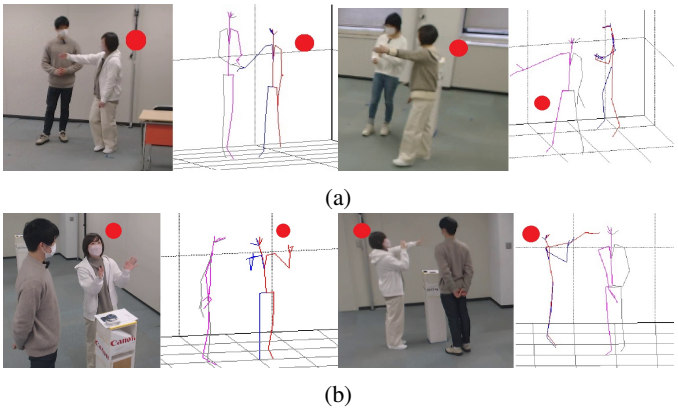
IV. Data Processing

We aim to build a fully unsupervised data-driven pipeline to generate robot behaviors with socially appropriate motions from only passively collected human-human interaction data. In this pipeline, the participants' typical behaviors will be used to train a classifier for generating robot behaviors. Currently, we are working on processing data to extract the typical behaviors in preparation for the later steps, including neural network training. In this section, we will introduce our methods for extracting participants' typical motion behaviors, including motion clustering and finding the environment references for pointing gestures.

A. Motion Clustering

People tend to use a set of typical behaviors for responding to similar interaction contexts. For example, in the camera

[1]https://cloud.google.com/



(a)

(b)

Fig. 3: Examples of the motion clustering results. The red dot marks the shopkeeper. (a) Pointing gestures from cluster 24 (left) and 28 (right). (b) Iconic gestures from cluster 2 (left) and 25 (right).

shop scenario, the shopkeeper will use a typical behavior to thank the customer for coming in by saying "welcome to our store" with a nod or a bow. These typical behaviors (speech, motion, etc.) and interaction patterns are what the imitation learning approach intends to learn [6]. We intend to learn the interactive motions from the skeleton data collected in Sec. III. But, it is difficult to learn directly from the raw skeleton data because it is noisy and high dimensional. Therefore, we used an unsupervised method to cluster different types of interactive motions into typical behavior clusters, which also resulted in reducing the noise and dimensionality.

When humans perform interactive motions, it is a continuous process that should be treated as time series rather than separate frames without time information. Therefore, we used time series K-means [14] to group the skeleton clips into 30 different clusters. The time series K-means uses dynamic time warping to calculate distances between clips, which represent the similarity of the clips over time. And the skeleton data is cut to one-second clips based on our observation that one second was enough time to perform a meaningful motion, and was not so long that separate motions would be mixed. We tested the number of clusters from 5 to 100 and selected 30 because 30 is the point when the inertia (the sum of the squared distance between the Centroid and each point of the cluster) starts decreasing in a linear fashion, which means adding another cluster doesn't give much better modeling of the data (Elbow method [15]). Since we focus more on human upper body motions, we selected 10 different features for the clustering related to the human hand and head movement, including the distance between wrists, elbows, and shoulders; the speed of hand movement; and head orientation. Since our goal is to imitate the behavior of the shopkeeper, the motion clustering is performed only on the shopkeeper to find the typical motions performed by the shopkeeper. In this work, we clustered the skeleton data from only a single shopkeeper participant to reduce the variation in the motion data. We leave exploring ways to deal with individual differences in interactive motions (from multiple shopkeepers) for future work.

TABLE I: Categories of Motion Clusters for One of the Shopkeepers

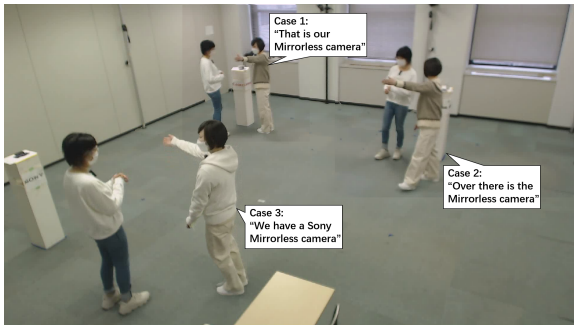
Category	Cluster ID	Total Duration (seconds)
Idle Motion	1, 3, 12, 15, 16, 21, 26, 29	53867
Deictic Gesture	9, 10, 11, 14, 19, 23, 24, 28	6509
Beat Gesture	4, 5, 6, 7, 17, 18, 20, 22, 27	17269
Iconic Gesture	2, 8, 13, 25, 30	4378

We manually checked the clustering results and found that each of the 30 clusters was well grouped as motions in the same cluster having similar hand and head movements. The 30 clusters could be categorized into several types, including deictic gestures, iconic gestures, beat gestures, and idle motions (mainly respectful standing and walking). Fig. 3 shows several examples of the motion clustering results and Table I shows the categories and total durations of the 30 clusters categorized by our manual check. We can see that the shopkeeper will frequently use different gestures during the interaction. The deictic gestures (6509 seconds) are used a lot in our scenario, which implies the importance of correctly imitating them with the robot. Among the observed iconic gestures, we found some motion clusters that might be unique in the camera shop scenario, for example, cluster 25 contains the motions the shopkeeper performed with her hands to describe how the light goes through a lens for explaining the function of a single lens reflex camera.

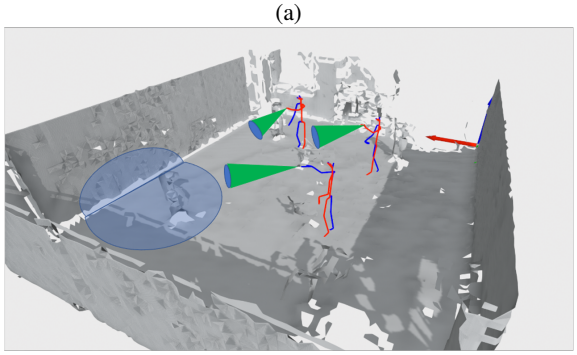
B. *Finding Environment Reference*

The environment references should be given to the robot to ensure its motion is consistent with its speech and its surrounding environment. Especially for pointing gestures, the pointing gestures provide position information of the environment. And the position information needs to be consistent with robot speech to perform socially appropriate behaviors. For example, we don't want the robot to say "this is an apple" while pointing to a random position. Since our aim is to design a system that can automatically learn to imitate interactive motions, we propose a method to automatically learn the environment reference for pointing gestures.

The information extracted from our speech data is used to find the environment reference for pointing gestures. During our data collection, we collected both motion data and speech data which are well synchronized in time. As we clustered the speech with the same process proposed in [16], utterances with similar meanings are grouped in the same cluster, so the speech clusters could provide timely information about when the participants are talking about the similar topics, thus could help to find the environment reference of pointing gestures. For example, "mirrorless camera" is a particular feature of the



(a)



(b)

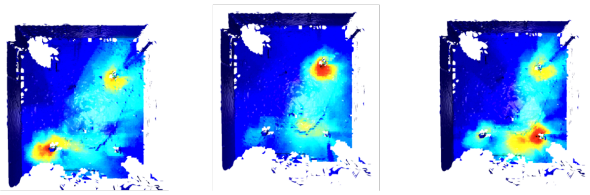
Fig. 4: Calculation of pointed area. (a) three cases the shopkeeper points while talking about "mirrorless camera". (b) use the pointing cone to find the pointed area.

Sony camera in our setup, and pointing gestures that are used when participants are talking about "mirrorless camera" have a high probability that is used to refer to the Sony camera.

Based on this assumption, we used the speech clusters to divide each pointing motion cluster into sub-clusters for determining the reference of pointing gestures. For a pointing cluster M_i and speech cluster S_j, we select motions from M_i, which have the same timestamp with the utterances from S_j, and put the selected motions to the same sub-motion-cluster M_{i,S_j}. For example, motions that have the same timestamp as utterances about a "mirrorless camera" will be in the same sub-motion cluster (Fig. 4a).

Furthermore, to link the pointing gestures and the speech data with an actual position in the room, for each sub-motion cluster M_{sub}, we used the pointing heat map to calculate the probability that each location of the room is pointed at. As shown in Fig. 4b, the pointing cone is a cone starting from the hand and pointing in the direction of the line passing through the head and that hand, which is suggested as a better description for pointing direction in [2]. Each pointing cone will have an intersection with the room map (the blue area in Fig. 4b). And by adding all intersections calculated from the same sub-motion-cluster, we can get a heat map like in Fig. 5, in which areas that are pointed at most frequently are 'hotter' (red).

Fig. 5 shows the heat maps of three sub-clusters for cluster 24, which is one of the pointing clusters. With the generated



(a) (b) (c)

Fig. 5: Pointing heat maps generated from sub-motion clusters of cluster 24 divided by topics of (a) Nikon camera, (b) Sony camera, and (c) Canon camera. The room map is generated from our sensor network and is on the top view of the room. Areas on the map that are more frequently pointed at are shown as closer to red (hotter). The hottest area in the above pictures, (a) bottom-left (b) top-right, and (c) bottom-right, are actually the position of each camera in the room.

heat maps we could determine the position of the pointing gesture for the robot in the imitation learning approach. For example, when the robot decides to perform a pointing gesture from motion cluster 24 and talk about "mirrorless camera", we can use the hottest area in Fig. 5b as the position the robot should point at. This should result in a robot action with synchronized motion, speech, and environment reference. The average distance between the hottest area on each heat map and the actual position of the closest camera is 0.21 m, which shows that our method can find the position of each camera.

Currently, we still need some minor human input of product information (e.g. "mirrorless camera" is a feature of Sony camera) to generate the above heat maps. But our aim for building the data-driven system is to find this kind of information automatically from speech data. We will discuss more of this part in our future works.

V. Conclusion

In this paper, we explained our current efforts toward a data-driven approach for imitating human interactive motions. Our main novel contribution is a method to find the environment references for pointing gestures by combining motion and speech data. The results obtained show that we are able to find the environment references for pointing gestures.

In our future work, we intend to use the imitation learning approach to train a neural network for decision-making for our robot, in which the input of the neural network represents the interaction state and the output is the shopkeeper action used for the robot. Finally, we hope that our imitation learning approach will reduce the cost of creating interactive motions for social robots in different social situations.

References

[1] P. Wagner, Z. Malisz, and S. Kopp, "Gesture and speech in interaction: An overview," *Speech Communication*, vol. 57, pp. 209–232, 2014.

[2] P. Liu, D. F. Glas, T. Kanda, H. Ishiguro, and N. Hagita, "It's not polite to point generating socially-appropriate deictic behaviors towards people," in *2013*

8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2013, pp. 267–274.

[3] C.-M. Huang and B. Mutlu, "Modeling and evaluating narrative gestures for humanlike robots." in *Robotics: Science and Systems*, 2013, pp. 57–64.

[4] Y. Okuno, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita, "Providing route directions: design of robot's utterance, gesture, and timing," in *2009 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2009, pp. 53–60.

[5] O. Sugiyama, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita, "Natural deictic communication with humanoid robots," in *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2007, pp. 1441–1448.

[6] 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, no. 4, pp. 988–1008, 2016.

[7] C. Breazeal, C. D. Kidd, A. L. Thomaz, G. Hoffman, and M. Berlin, "Effects of nonverbal communication on efficiency and robustness in human-robot teamwork," in *2005 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2005, pp. 708–713.

[8] B. Mutlu, J. Forlizzi, and J. Hodgins, "A storytelling robot: Modeling and evaluation of human-like gaze behavior," in *2006 6th IEEE-RAS International Conference on Humanoid Robots*. IEEE, 2006, pp. 518–523.

[9] M. Salem, S. Kopp, I. Wachsmuth, K. Rohlfing, and F. Joublin, "Generation and evaluation of communicative robot gesture," *International Journal of Social Robotics*, vol. 4, no. 2, pp. 201–217, 2012.

[10] Y. Yoon, W.-R. Ko, M. Jang, J. Lee, J. Kim, and G. Lee, "Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4303–4309.

[11] M. Plappert, C. Mandery, and T. Asfour, "Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks," *Robotics and Autonomous Systems*, vol. 109, pp. 13–26, 2018.

[12] E. Shlizerman, L. Dery, H. Schoen, and I. Kemelmacher-Shlizerman, "Audio to body dynamics," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7574–7583.

[13] C. T. Ishi, C. Liu, J. Even, and N. Hagita, "Hearing support system using environment sensor network," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 1275–1280.

[14] R. Tavenard, J. Faouzi, G. Vandewiele, F. Divo, G. Androz, C. Holtz, M. Payne, R. Yurchak, M. Rußwurm, K. Kolar, and E. Woods, "Tslearn, a machine learning toolkit for time series data," *Journal of Machine Learning Research*, vol. 21, no. 118, pp. 1–6, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-091.html

[15] R. L. Thorndike, "Who belongs in the family," in *Psychometrika*. Citeseer, 1953.

[16] M. Doering, D. F. Glas, and H. Ishiguro, "Modeling interaction structure for robot imitation learning of human social behavior," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 3, pp. 219–231, 2019.