

Can a social robot train itself just by observing human interactions?

Dylan F. Glas, Phoebe Liu, Takayuki Kanda, *Member, IEEE*, Hiroshi Ishiguro, *Senior Member, IEEE*

Abstract— In HRI research, game simulations and teleoperation interfaces have been used as tools for collecting example behaviors which can be used for creating robot interaction logic. We believe that by using sensor networks and wearable devices it will be possible to use observations of live human-human interactions to create even more humanlike robot behavior in a scalable way. We present here a fully-automated method for reproducing speech and locomotion behaviors observed from natural human-human social interactions in a robot through machine learning. The proposed method includes techniques for representing the speech and locomotion observed in training interactions, using clustering to identify typical behavior elements and identifying spatial formations using established HRI proxemics models. Behavior logic is learned based on discretized actions captured from the sensor data stream, using a naïve Bayesian classifier, and we propose ways to generate stable robot behaviors from noisy tracking and speech recognition inputs. We show an example of how our technique can train a robot to play the role of a shop clerk in a simple camera shop scenario.

I. INTRODUCTION

Machine learning has been applied to several elements of HRI, e.g. to mimic gestures and movements [1] or to learn how to direct gaze in response to gestural cues [2]. So far, little effort has been made towards using machine learning for the overall generation of robot motions and spoken utterances in a conversational interaction. Yet, many of the challenges posed by conversational interaction resemble the kinds of problems where machine learning is typically applied, i.e., decision-making under uncertainty in a high-dimensional space. In particular, unconstrained speech recognition is highly noisy (a problem not faced by chatbots), and there can be a lot of natural variation between semantically-similar speech or motion behaviors conducted by different individuals.

For dialogue systems to be useful and robust, they often require tens of thousands of utterance rules to be created. To minimize design effort, it would be ideal to train such systems from human-human interaction data, rather than manually authoring the rules. We have been researching ways to automate the collection of human-human interaction data, and to use machine learning to characterize the elements of those interactions and reproduce the observed human behaviors.

Some work has investigated learning-by-imitation approaches for reproducing free-form human actions in the context of video games, and this work is conceptually similar

* This research was supported in part by JSPS KAKENHI Grant Number 25240042 and in part by JST, ERATO.

D. F. Glas, P. Liu, and T. Kanda are with ATR Intelligent Robotics and Communication Labs., Kyoto, Japan. H. Ishiguro is with the Intelligent Robotics Laboratory, Graduate School of Engineering Science, Osaka University, Toyonaka, Japan. D. F. Glas and H. Ishiguro are also with the Ishiguro Symbiotic Human-Robot Interaction Project, ERATO, JST. (corresponding author's phone: +81-774-95-1405; fax: +81-774-95-1408; e-mail: dylan@atr.jp).

in some ways to the “restaurant game” work of Orkin et al. [3] and the “Crowdsourcing HRI” work of Breazeal et al. [4].

In this paper we will present some of the ways that we have applied machine learning techniques to the problem space of reproducing social interactions based on data collected by sensors, such as those shown in Fig. 1. Much of the work we present here is based on our previous paper from RO-MAN 2014 [5], and a journal paper documenting an improved version of our system is currently under review, so this paper will focus on providing a high-level view of our approach, presenting transcripts of example interactions, and sharing some of our challenges and successes in this process.

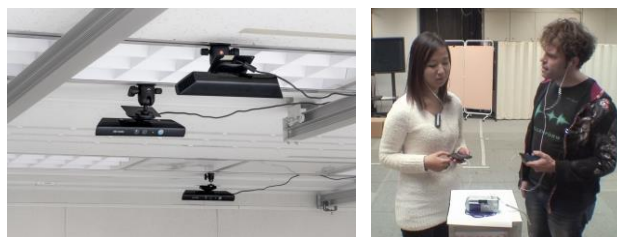


Figure 1. Sensors used in this study. Left: ceiling-mounted Kinect sensors for position tracking. Right: capturing speech data with smartphones.

II. SCENARIO AND OVERVIEW

Our overall strategy was to use a purely data-driven approach for generating both robot behaviors and the rules which trigger them. Although it may seem that for simple scenarios, better results might be attained by hand-coding robot behaviors, the principle of keeping the process purely data-driven is important for scalability of the technique.

A. Scenario

The scenario we used for this study was a customer-shopkeeper interaction in a camera shop, the objective being to train the robot to reproduce the actions of the shopkeeper. This scenario presented many opportunities for movement to different locations, as well as conversational content that depended on the location context. For example, the answer to “how much does this camera cost?” is different depending on which camera the customer is looking at.

B. Data Collection

To perform learning from interactions in real-world environments such as an actual retail shop, it would be desirable to capture behavior data using only passive sensing techniques, so as to interfere as little as possible with the natural interactions. To this end, we used a position tracking system based on ceiling-mounted Kinect sensors to capture people’s positions and movement [6]. However, since accurate speech recognition is not yet easily achieved using environmentally-mounted microphones, we used handheld smartphones to capture their speech (Fig. 1). Participants

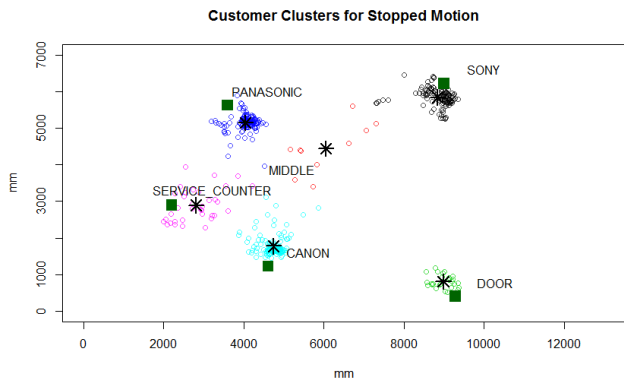


Figure 2. Typical stopping location clusters for the customer. Asterisks show cluster centers, and squares show the locations of known objects.

tapped the smartphone display before and after speaking, and their speech was recognized using the Google speech API.

With this system, we recorded 178 example human-human interactions to be used as training data. Live data from the same sensor system was later used for conducting online interactions with the robot. Such use of a sensor network to augment a robot’s on-board sensing is common in “Network Robot System” applications for social robotics [7].

C. Learning Strategy

Our basic learning strategy can be summarized as follows:

1. Discretize actions of shopkeeper and customer in time.
2. Use abstraction techniques to represent customer actions as a feature vector.
3. Represent shopkeeper actions as discrete executable robot actions.
4. Train predictor with customer-shopkeeper action pairs.
5. In online system, call predictor whenever customer action is detected, and execute predicted robot action.

The following sections will elaborate on these steps and introduce the techniques we used for processing the noisy sensor data into representations that are useful for machine learning and robot behavior generation.

III. ABSTRACTION OF FEATURES

A. Spatial

Rather than using raw (x, y) positions for representing spatial location, we identified a discrete set of typical stopping locations in the room. To do this, we segmented trajectories from the data collection by using velocity thresholding to separate walking from stopped segments.

We then used unsupervised k-means clustering to group the stopped segments into typical stopping locations for each person (see Fig. 2), and we represented each moving segment as a transition between two stopping locations. For the most part, these points corresponded to known objects in the room (the door, the service counter, and three cameras of different brands: Sony, Panasonic, and Canon), so in this paper we refer to them by those labels. As an example of a typical movement



Figure 3. Spatial formations detected in this study. From left to right, they are *waiting*, *face-to-face*, and *present object*.

action, we might see the customer moving from the *door* location to the *canon* location.

The discretized locations of the customer and shopkeeper, and a state variable representing whether or not they were moving, were combined into a feature vector $F_{spatial}$.

B. Formations

Next, we modeled each interaction as consisting of a sequence of stable interaction states which last for several turns in a dialogue, recognizable by distinct spatial formations such as talking face-to-face or presenting a product. The modeling of interaction states helps to generate locomotion in a stable way, to specify robot proxemics behavior at a detailed level, and to provide context for more robust behavior prediction.

We identified three interaction states related to existing HRI models: *present object*, based on the work of Yamaoka et al. [8], *face-to-face*, based on interpersonal distance defined by Hall [9], and *waiting*, inspired by the modeling of socially-appropriate waiting behavior by Kitade et al. [10]. Examples of these states are shown in Fig. 3. Discrete variables representing the interaction state and the target location, if any, were added to a feature vector $F_{formation}$.

C. Speech

To represent speech as a vector for use in machine learning, we used several common speech-processing techniques, including removal of stopwords, a Porter stemmer, the generation of n-grams to capture word sequence, generation of a term frequency-inverse document frequency (TF-IDF) matrix, and Latent Semantic Analysis (LSA), a dimensionality-reduction technique for text similar to principal components analysis.

After this processing, each captured utterance was represented as a vector of approximately 350 dimensions. We designated this vector as F_{speech} .

IV. DEFINING ROBOT ACTIONS

For each observed shopkeeper action, it was necessary to create a corresponding robot action, incorporating speech and locomotion. As an example, consider the case where the shopkeeper was observed to say, “It comes in red and silver,” while presenting the Sony camera to the customer.

A. Locomotion

Locomotion behaviors were defined in terms of achieving a target interaction state. Thus, in the above example, the target interaction state would be *present product* (an interaction state corresponding to a spatial formation) at *sony* (a location known based on the clustered stopping locations).

To execute this action, the robot must first determine whether it is in the target state. If not, it moves towards the destination most likely to achieve that state, using the proxemics model for *present product* and the projected position of the customer to choose a target location to move to in order to achieve *present product* at *sony*.

B. Speech

In order to reproduce speech behaviors, we faced the difficult problem that speech recognition results were significantly corrupted by speech recognition errors.

An analysis of 400 utterances from the training interactions showed that 53% were correctly recognized, 30% had minor errors, e.g., “can it should video” rather than “can it shoot video,” and 17% were complete nonsense, e.g. “is the lens include North Florida.”

Since nearly half of the captured utterances contained errors, we needed some strategy to minimize the impact of these errors on the speech generated by the robot. We clustered the shopkeeper’s speech utterances using dynamic hierarchical clustering [11] to group the observed shopkeeper utterances into clusters representing unique speech elements. 166 clusters were obtained from 1233 shopkeeper utterances.

Next, we analyzed each cluster to identify the utterance with the greatest similarity to other utterances in that cluster, in order to minimize the likelihood that it contained recognition errors. For this step, it was important to use the actual text strings rather than their vectorized representations.

Finally, we extracted a typical utterance for each cluster to be defined as a robot speech action, which would usually be a paraphrase of the actual utterance. The example above might map to the phrase, “We have red and silver available.”

C. Execution

For robot locomotion, the dynamic window approach was used for obstacle avoidance [12]. Speech was synthesized with Ximera software [13]. The robot’s gaze was always directed towards the customer, and idle behaviors were generated based on whether the robot was speaking, stopped, or moving [14].

V. TRAINING THE PREDICTOR

A. Discretizing actions

As described in Section II-C, the basic procedure of our learning approach was to first identify discrete action events for the shopkeeper and customer in the training data, and then to train a predictor to predict an appropriate robot (shopkeeper) action each time a human (customer) action was detected.

Actions were generated whenever one person started moving, which we detected by velocity thresholding, or when someone spoke, which was detected by the system whenever a new utterance was output by the speech recognizer.

B. Train classifier

We then considered all instances where a customer action was followed by a shopkeeper action, as shown in Fig. 3. These action pairs were used to train a Naïve-Bayesian classifier to predict a discrete robot action based on a vector characterizing the customer action.

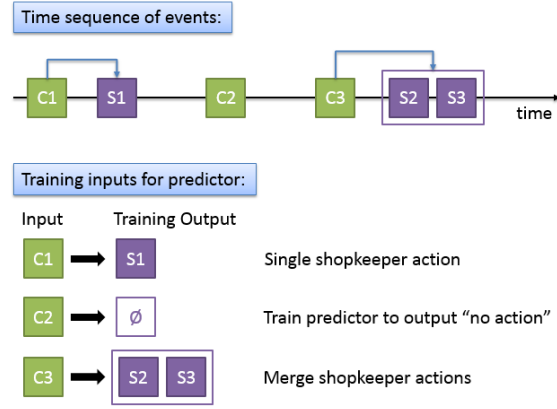


Figure 4. Correspondence of action pairs.

We trained the classifier using the feature vector comprised of $F_{spatial}$, $F_{formation}$, and F_{speech} for each customer action as a training input, and we used the subsequent robot action corresponding to the shopkeeper action as its training class.

The naïve-Bayesian classifier is a generative classification technique which uses the formula below to classify an instance that consists of a set of feature-value pairs.

$$a_{NB} = \arg \max_{c_j \in C} P(a_j) \prod_i P(f_i = v_i | a_j)$$

a_j , denotes a robot action, and f_i denotes a feature in the feature vector. The naïve-Bayesian classifier picks a robot action, a_{NB} , that maximizes the probability of being classified to the robot action given the value v_i for each feature f_i .

Each feature has different dimensionality. Thus, the model can be extended to:

$$v_i = \{t_{i1}, t_{i2}, \dots, t_{im}\}$$

$$a_{NB} = \arg \max_{c_j \in C} P(a_j) \prod_i \left(\prod_k P(t_{ik} \text{ appears in } f_i | a_j) \right)^{w_i}$$

We would like to give higher priority to values in the features that are more discriminative in classifying robot action. Information gain tells us how important a given feature in the joint state vector is. Therefore a weighting factor w_i is applied for each feature f_i , calculated as the information gain ratio, that is, the ratio between information gain and intrinsic value for each feature over all training examples.

VI. EXAMPLE INTERACTION

To demonstrate the kinds of interactions that the robot can reproduce using this technique, we provide the transcripts of two interactions between recruited participants and our robot in Tables I and II. These interactions illustrate several important capabilities of our system.

A. Successes

In this scenario, the primary function of the robot is to provide information about the various features of the cameras, and as these examples show, the robot was able to correctly answer questions about features such as price, color, weight, storage capacity, and battery life of three different cameras.

TABLE I: TRANSCRIPT OF EXAMPLE INTERACTION 1






























(1-1)		<p>(Scene 1-1) <i>Robot is standing at service counter.</i> <i>Customer enters and approaches panasonic.</i></p>
(1-2)		<p>(Scene 1-2) <i>Robot begins approaching panasonic.</i> Robot: hi is there anything I can help you with today</p>
(1-3)		<p>Customer: Ah yes, this camera, uh, how's the battery life?</p>
(1-4)		<p>(Scene 1-3) Robot (arriving at panasonic): long battery life of 9 hours.</p>
(1-5)		<p>Customer: Ah I see. And... the weight?</p>
(1-6)		<p>Robot: this camera is 150 grams and it can fit in your pocket</p>
(1-7)		<p>Customer: Oh, wow. That's very convenient. Thank you.</p>
(1-4)		<p><i>Customer turns away and begins walking towards canon.</i></p>
(1-5)		<p>Robot: OK <i>Robot drives back to service counter.</i></p>
(1-6)		<p>Customer (at canon): How about this one? How's the storage memory?</p>
(1-7)		<p>(Scene 1-4) <i>Robot begins moving towards canon.</i></p>
(1-5)		<p>Robot (while moving): yes it can store 10000 photos</p>
(1-6)		<p>Customer: hmm, how much is it?</p>
(1-7)		<p>(Scene 1-5) <i>Robot arrives at canon.</i> Robot: this is \$400</p>
(1-5)		<p>Customer: Ok. What colors does it come in?</p>
(1-6)		<p>Robot: intense grey red and brown colors</p>
(1-7)		<p>Customer: Ah ok, well, actually... <i>Customer walks back to panasonic.</i></p>
(1-5)		<p>(Scene 1-6) <i>Robot turns and follows customer to panasonic.</i></p>
(1-6)		<p>Customer: What's the optimal zoom on this camera?</p>
(1-7)		<p><i>Robot arrives at panasonic and turns toward the customer.</i> Robot: 5x optical zoom</p>
(1-5)		<p>Customer: Oh, ok. Thank you. <i>Customer turns to leave.</i></p>
(1-6)		<p>(Scene 1-7) Robot: no problem <i>Robot returns to service counter.</i></p>

TABLE I: TRANSCRIPT OF EXAMPLE INTERACTION 2

(2-1)		<p>(Scene 2-1)</p> <p><i>Robot is standing at service counter. Customer enters and approaches canon.</i></p>
		<p>Robot: hi, my I help you</p>
		<p>Customer: Yes, I'm looking for a camera with interchangeable lenses.</p>
(2-2)		<p>(Scene 2-2)</p> <p><i>Robot begins moving towards sony. Customer follows.</i></p>
		<p>Robot: Over here we have my favorite which is the Sony NEX 5 which is a mini SLR and has 28 replaceable lens.</p>
		<p>Customer: 28 replaceable lens?</p>
(2-3)		<p>(Scene 2-3)</p> <p>Robot (arriving at sony): This is \$600.</p>
		<p>Customer: Do you have a cheaper one?</p>
(2-4)		<p>(Scene 2-4)</p> <p><i>Robot turns and begins moving towards canon. Customer follows.</i></p> <p>Robot: (while moving) Rite Aid this is because the mirrorless camera is actually lighter. This is a little bit heavier but has good performance in low light.</p>
		<p>(Scene 2-5)</p> <p>Customer (as they arrive at canon): Oh, really?</p>
		<p>Robot: Yes.</p>
(2-5)		<p>(Scene 2-6)</p> <p><i>Customer turns to face towards sony.</i></p>
		<p>Customer: I think the Sony camera is very good.</p>
		<p>Robot: 400 dollars.</p>
		<p>Customer: (laughs)</p>
(2-6)		<p>(Scene 2-7)</p> <p>Customer: Thank you. See you. Goodbye.</p>
		<p><i>Customer turns to leave.</i></p>
		<p>Robot: No problem.</p>
(2-7)		<p><i>Robot returns to service counter.</i></p>

The robot responds not only to speech, but also to motion cues from the customer. In Scene 1-1, when the customer enters and approaches *panasonic*, the robot responds by offering to help and approaching the same camera. Later, in Scene 1-5, she walks from *canon* to *panasonic*, and the robot follows her to the new camera.

These interactions also illustrate how the robot is able to perform movement and speech at the same time. In Scene 1-4, the customer asked a question to the robot while it was at the service counter, and it predicted that it should provide the answer and establish the *present product* formation at the *canon* location. Thus, it spoke the predicted utterance, while at the same time driving to that target location. Scenes 2-2 and 2-4, show other examples of the robot speaking while moving.

The system is also robust to phrasing and recognition errors. For example, in Scene 1-6, the customer misspoke and said “*optimal zoom*” rather than “*optical zoom*.” Because the system was trained from noisy speech recognition data, it is quite robust to small errors like this, and it was able to correctly answer the question regardless of that error.

B. Challenges and Limitations

We considered the robot’s performance in these example interactions to be quite acceptable overall. However, it is important to consider the challenges and limitations of the system and of the approach in general.

First, some minor phrasing issues can be seen in the example interactions. In Scene 1-4, the robot says “yes it can store 10000 photos,” where the word “yes” would not have been warranted. Likewise, in Scene 1-6, the robot said “5x optical zoom,” whereas a human probably would have said something more grammatically complete, like “it has 5x optical zoom”. Several very minor errors like this occurred because our system has no knowledge of semantic meaning or grammatical structure.

The robot sometimes spoke strange utterances because of speech recognition errors. In Scene 1-5, the robot says “*intense gray, red, and brown colors*,” a phrase derived from a speech recognition error in the training data when the shopkeeper had said, “*it has gray, red, and brown colors*.” Similarly, the phrase “*my I help you*” in Scene 2-1 was an error in the recognition of “*may I help you*”, and “*Rite Aid*” in Scene 2-4 came from incorrect recognition of “*Right, and*” in the training data. Interestingly, most of these mistakes went unnoticed by the participants and even the experimenters. We attribute this to the fact that many speech recognition errors resulted in words that were phonetically similar to the correct ones, and people unconsciously corrected the errors.

One limitation of this approach is the fact that it contains no representation of history, so for example we sometimes observed situations where the customer would approach one camera, the robot would say, “may I help you,” then the customer would say “no, thanks,” and move to another camera. Frequently the robot would then repeat, “may I help you?” because the predictor did not consider interaction history.

Finally, we have shown this technique to be effective in the kinds of interactions where the robot must directly respond to a human’s actions. We believe that this will cover a wide range of human-robot interaction scenarios, but it might not be

expected to perform so well in contexts where the robot needs to be more proactive.

VII. CONCLUSIONS

In this study, we showed a proof of concept that a purely data-driven approach can be used to reproduce social interactive behaviors with a robot based on sensor observations of human-human interactions.

Overall, we were quite satisfied with the performance of the system, and we think that the scalability of a data-driven approach gives it the potential to transform the way social behavior design is conducted in HRI. Once passive collection of interaction data becomes practical, even a single sensor network installation could provide enormous amounts of example interaction data over time, an invaluable resource for the collection and modeling of social behavior. We believe that with today’s trends towards big-data systems and cloud robotics, techniques like this will become essential methods for generating robot behaviors in the future.

REFERENCES

- [1] B. M. Scassellati, "Foundations for a Theory of Mind for a Humanoid Robot," Massachusetts Institute of Technology, 2001.
- [2] Y. Nagai, "Learning to comprehend deictic gestures in robots and human infants," in *Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on*, 2005, pp. 217-222.
- [3] J. Orkin and D. Roy, "The restaurant game: Learning social behavior and language from thousands of players online," *Journal of Game Development*, vol. 3, pp. 39-60, 2007.
- [4] C. Breazeal, N. DePalma, J. Orkin, S. Chernova, and M. Jung, "Crowdsourcing Human-Robot Interaction: New Methods and System Evaluation in a Public Environment," *Journal of Human-Robot Interaction*, vol. 2, pp. 82-111, 2013.
- [5] P. Liu, D. F. Glas, T. Kanda, H. Ishiguro, and N. Hagita, "How to train your robot - teaching service robots to reproduce human social behavior," in *Robot and Human Interactive Communication, 2014 ROMAN: The 23rd IEEE International Symposium on*, 2014, pp. 961-968.
- [6] D. Brščić, T. Kanda, T. Ikeda, and T. Miyashita, "Person Tracking in Large Public Spaces Using 3-D Range Sensors," *Human-Machine Systems, IEEE Transactions on*, vol. 43, pp. 522-534, 2013.
- [7] D. F. Glas, S. Satake, F. Ferreri, T. Kanda, H. Ishiguro, and N. Hagita, "The Network Robot System: Enabling social human-robot interaction in public spaces," *Journal of Human-Robot Interaction*, 2012.
- [8] F. Yamaoka, T. Kanda, H. Ishiguro, and N. Hagita, "How close?: model of proximity control for information-presenting robots," in *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, Amsterdam, The Netherlands, 2008, pp. 137-144.
- [9] E. T. Hall, *The Hidden Dimension*. London, UK: The Bodley Head Ltd, 1966.
- [10] T. Kitade, S. Satake, T. Kanda, and M. Imai, "Understanding suitable locations for waiting," in *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*, 2013, pp. 57-64.
- [11] P. Langfelder, B. Zhang, and S. Horvath, "Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R," *Bioinformatics*, vol. 24, pp. 719-720, 2008.
- [12] D. Fox, W. Burgard, and S. Thrun, "The dynamic window approach to collision avoidance," *Robotics & Automation Magazine, IEEE*, vol. 4, pp. 23-33, 1997.
- [13] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda, "XIMERA: A new TTS from ATR based on corpus-based technologies," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.
- [14] C. Shi, T. Kanda, M. Shimada, F. Yamaoka, H. Ishiguro, and N. Hagita, "Easy development of communicative behaviors in social robots," in *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, 2010, pp. 5302-5309.