Auton Robot (2018) 42:1067–1085
https://doi.org/10.1007/s10514-017-9671-8

Learning proactive behavior for interactive social robots

Phoebe Liu[1] · **Dylan F. Glas**[1] · **Takayuki Kanda**[2] · **Hiroshi Ishiguro**[3]

Received: 8 December 2016 / Accepted: 17 October 2017 / Published online: 6 November 2017

Abstract Learning human–robot interaction logic from example interaction data has the potential to leverage "big data" to reduce the effort and time spent on designing interaction logic or crafting interaction content. Previous work has demonstrated techniques by which a robot can learn motion and speech behaviors from non-annotated human–human interaction data, but these techniques only enable a robot to respond to human-initiated inputs, and do not enable the robot to proactively initiate interaction. In this work, we propose a method for learning both human-initiated and robot-initiated behavior for a social robot from human–human example interactions, which we demonstrate for a shopkeeper interacting with a customer in a camera shop scenario. This was achieved by extending an existing technique by (1) introducing a concept of a customer *yield action*, (2) incorporating interaction history, represented by sequences of discretized actions, as inputs for training and generating robot behavior, and (3) using an "attention mechanism" in our learning system for training robot behaviors, that learns which parts of the interaction history are more important for generating robot behaviors. The proposed method trains a robot to generate multimodal actions, consisting of speech and locomotion behaviors. We compared this study with the previous technique in two ways. Cross-validation on the training data showed higher social appropriateness of predicted behaviors using the proposed technique, and a user study of live interaction with a robot showed that participants perceived the proposed technique to produce behaviors that were more proactive, socially-appropriate, and better in overall quality.

Keywords Human–robot interaction · Data-driven learning · Learning by imitation · Social robotics · Service robots · Proactive behaviors

This is one of the several papers published in *Autonomous Robots* comprising the Special Issue on Learning for Human-Robot Collaboration.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10514-017-9671-8) contains supplementary material, which is available to authorized users.

✉ Phoebe Liu
phoebe@atr.jp

Dylan F. Glas
dylan@atr.jp

Takayuki Kanda
kanda@atr.jp

Hiroshi Ishiguro
ishiguro@sys.es.osaka-u.ac.jp

[1] ERATO Ishiguro Symbiotic Human-Robot Interaction Project ATR-HIL, 2-2-2 Hikaridai, Keihanna Science City, Kyoto, Japan

[2] ATR-IRC, 2-2-2 Hikaridai, Keihanna Science City, Kyoto, Japan

[3] ERATO Ishiguro Symbiotic Human-Robot Interaction Project IRL, Osaka University, 1-3 Machikaneyama, Toyonaka, Osaka, Japan

1 Introduction

The vision of humanoid robots providing service through natural conversational interaction, once a dream of science fiction, is now closer than ever to becoming a reality (Satake et al. 2015; Triebel et al. 2016; Jayawardena et al. 2016; Shiomi et al. 2009). With the arrival of commercial humanoid robot platforms like Pepper, social robots have begun to appear in commercial and public spaces. However, the problem of how to develop social interaction logic for conversational robots, including interactive dialog and interactive motion planning, is still a relatively young and unexplored research domain.

Some works in HRI have already demonstrated techniques for learning speech and motion behavior by imitation from human behavior captured from live interactions (Liu et al. 2016) and online games (Breazeal et al. 2013; Orkin and Roy 2007). These studies applied data-driven techniques to learn application logic through imitation of human behavior, as opposed to using a more traditional approach of manually designing interaction logic. As the availability of machine power for learning and the availability of large data sets increase, we propose that for situations where large amounts of example human–human interaction data is available, such data-driven approaches could produce more reliable interaction logic and require less effort than manual programming.

A typical approach to designing interaction logic for robots is to specify the robot's behavior in terms of responses to human actions or commands (Orkin and Roy 2009; Liu et al. 2016; Breazeal et al. 2013). Such approaches result in fundamentally passive systems, in which the robot only responds to explicit commands or actions from the human. However, many real social situations are mixed-initiative, and it is important for a robot not only to react to a person's actions, but to proactively take initiative as well. For example, a good museum guide not only answers questions about an exhibit, but should also ask questions back and provide interesting anecdotes about the exhibit to the visitor. Likewise, in a shopping scenario, a proactive shopkeeper would take the initiative to explain different product features to a customer.

Nevertheless, learning proactive behaviors in a data-driven way without hand-crafted rules or an explicit model of user's intention (Schrempf et al. 2005; Pandey et al. 2013) can be difficult, as rules for generating reactive versus proactive behavior can have different requirements. For example, in a shopping scenario, a reactive response to a customer's question may depend primarily on the customer's question itself, whereas a proactive behavior, in which the shopkeeper decides to take the initiative to do something (e.g. introducing a new product) as a result of the customer yielding his turn, may depend more strongly on interaction history or context. However, such contextual sensitivity is difficult to capture, and the naive injection of context information may introduce unnecessary noise, making the data too sparse and non-repeatable for the robot to learn an appropriate action. The question remains open as to how a robot can simultaneously and effectively learn the rules for generating both user-initiative and self-initiated actions.

In this work, we will address the question of how to learn both reactive and proactive robot behaviors from human interaction data. In previous work (Liu et al. 2016) we proposed a technique capable of learning social interaction logic for a robot in response to a human's speech and motion actions. However, that system is unable to generate proactive behavior, e.g. the robot does nothing unless the customer takes an action.

Thus, we propose three extensions to our previous work. First, we introduce a concept of a "yield action" enabling the robot to identify opportunities for a proactive action to be generated. Second, since proactive behaviors are often sensitive to the context of the interaction, we propose to incorporate **interaction history** as a training input. Third, we use an **attention mechanism** in our learning system, which has the ability to "attend" and learn which parts of the interaction history are important when predicting robot behaviors. In this work we will present this proposed architecture and demonstrate through offline analysis and live interactions with users that the proposed system can effectively reproduce proactive behavior learned from human interaction data.

2 Related work

Since learning both reactive and proactive behaviors for a social robot is novel, no previous study has reported an integrated method to address its whole process, although parts of the learning problem have been addressed to some degree. In this section, we report related works on some aspects of learning social behaviors.

2.1 Learning social behaviors from data

Several data-driven approaches have been applied to learning interactive behaviors for social robots. For example, Young et al. used learning from demonstration to generate real-time interactive paths for an animated characters and robots to match the style of interactive motion behaviors, based on a pattern-matching algorithm (Young et al. 2013, 2014).

Frameworks focused on crowdsourcing have been developed to enable learning of overall interaction logic from data collected from simulated environments, such as The Robot Management System framework (Toris et al. 2014) and The Mars Escape online game (Breazeal et al. 2013; Chernova et al. 2011). Remote users can interact collaboratively either in an online game, or through the web, and the interaction data are logged and used to develop HRI behaviors in a real autonomous robot. Our work complements these approaches by considering crowd-based data collected directly from human–human interaction using sensors in a physical environment, which presents unique challenges regarding resolving noise from sensor data, abstracting natural variations of human behavior, and discretizing actions for a robot to reproduce.

The use of real human interaction data collected from sensors for learning interactive behaviors has been investigated in some works. The robot JAMES was developed to serve drinks in a bar setting, in which a number of supervised (i.e. dialog management) and unsupervised learning techniques (i.e. clustering of social states) were applied to learn social

interaction (Keizer et al. 2014). Admoni and Scassellati proposed a model using empirical data from annotated human–human interactions to generate nonverbal robot behaviors in a tutoring application. The model can simultaneously predict the context of a newly observed set of nonverbal behaviors, and generate a set of nonverbal behaviors given a context of communication (Admoni and Scassellati 2014). Similar to these works, we use data from human–human interaction for learning robot behaviors, but we adopt a completely hands-off approach, with no human annotation needed for abstraction of social states or for robot behavior generation.

2.2 Proactive robot behaviors

Strategies for generating proactive robot behavior, in part, have been addressed in other works. In Rozo et al.'s work (2016), a robotic manipulator learns to complete a pouring and a handover task, in which they empirically predetermined six states the robot arm should be in. They achieve this by exploiting the temporal patterns (i.e. sequence of states) observed in the learning phase using an adaptive duration semi-Markov Model (ADHSMM) to generate state sequences and durations for the arm trajectory. Likewise, Huang et al. investigated proactive and reactive collaboration strategies that take account of real-time awareness of the task status of its user in performing handover actions between a human and robot manipulator (Huang et al. 2015). Other works focus on recognition of human intention in order to proactively decide when to complete the handover task (Schmid et al. 2007; Schrempf et al. 2005; Awais and Henrich 2012). For the most part, a typical objective for these foregoing works is to learn state sequences or durations using techniques like HMM, where the states are defined *a priori* based on domain knowledge of a specific, structured task. In contrast, our work addresses an open-ended problem of learning social interaction tasks in an unknown domain, where actions and states are not predetermined. The technique we propose begins from the problem of retrieving clusters from sensor data of unconstrained natural language and motion trajectories, and learns common transition patterns among them, including proactive behavior, using a deep neural network (DNN).

In the context of social robots, some works focus on how to better equip the robot to initiate interaction in a friendly and natural manner (Mutlu et al. 2009) or encourage people to initiate conversation (Robins et al. 2009; Hayashi et al. 2007). The use of proxemics has also been investigated for initiating interaction, such as feature representations for analyzing human spatial behaviors (Bauer et al. 2009) and developing generative model for approaching people (Satake et al. 2009) and maintaining spatial formation (Shi et al. 2011; Michalowski et al. 2006). Our work builds upon these studies by incorporating proxemics models for human–robot interaction, using them to support the higher-level goal of learning overall interaction logic, which combines proxemics, locomotion, and dialogue.

2.3 Learning from history

Some techniques have been developed for learning robot behaviors from history, such as goal-directed and habitual robot behaviors through a Bayesian dynamic working memory system (Viejo et al. 2015), or incorporating history in learning for mobile robots (Michaud and Matarić 1998; Mohammad and Nishdia 2012). Although our work also learns from history, we believe our work is closer to fields of language or dialog learning, where speech is a major part of the interaction.

Regarding learning from history for dialog in particular, many techniques involving deep neural networks have been developed recently for handling language-related tasks, which are inherently sequential and require some level of history or memory. Recurrent neural networks (RNN) (Mikolov et al. 2010) are often used for tasks like language processing, and Long Short-Term Memory (LSTM) (Hulme et al. 1991) techniques are often used for tasks such as word-by-word machine reading, where the meaning of a sentence must be interpreted in the context of previously encountered words (Cheng et al. 2016). A related technique, which we use in this work, is supplementing a neural network with an attention mechanism, which learns which part of an input sequence is important for predicting a response (Sukhbaatar et al. 2015; Bahdanau et al. 2014; Hermann et al. 2015). While several algorithms have been proposed for learning from history, it is still unclear how effective they can be in the problem space of learning human–robot multimodal interaction from noisy data, which is the main objective of our work.

3 Data collection

This section introduces our scenario for data collection, a camera shop, as well as the procedure and some observed behaviors of the participants.

3.1 Scenario

We chose a camera shop scenario for this study as an example of the kind of repeatable interaction for which this technique would be most useful. We set up a simulated camera shop environment in our laboratory with three camera models on display, each at a different location (Fig. 1), and we asked a participant to role-play a proactive shopkeeper. The shopkeeper interacted with participants role-playing customers, walking with the customers to different cameras in the shop, answering questions about camera features, and proactively

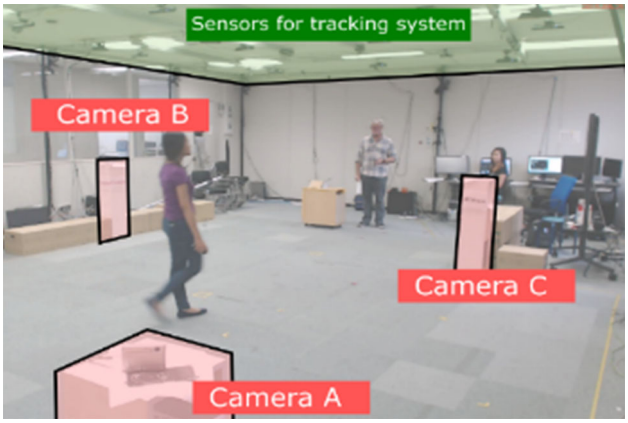


Fig. 1 Environment setup for our study, featuring three camera displays. Sensors on the ceiling were used for tracking human position, and smartphones carried by the participants were used to capture speech

introducing new cameras or features when the customers had no specific questions. We recorded the speech and motion data of both the shopkeeper and the customers during these interactions.

3.2 Sensors

To capture the participants' motion and speech data, we used a human position tracking system to record people's positions in the room, and we used a set of handheld smartphones for speech recognition.

The position tracking system used data from 20 Microsoft Kinect 1 sensors, arranged in opposing rows on the ceiling to minimize interference, with a lateral spacing of 1.9 m. The arrangement is similar to that shown in Glas et al. (2015). Particle filters were used to estimate the position of each person in the room based on point cloud data (Brscic et al. 2013).

Speech was captured via a smartphone with a hands-free headset, using the Android speech recognition API to recognize utterances and sending the text to a server via Wi-Fi. Users were required to touch the mobile screen to indicate the beginning and end of their speech. Although it would be ideal to passively collect speech data from microphones in the environment and automatically detect the start and stop of speech activity, reliable technologies to do this are not yet widely available.

Location data for the shopkeeper and the customer were recorded at a rate of 20 Hz. Speech data were recorded at the start and end of each speech event, as signaled by participants tapping on their Android phones.

3.3 Participants

The customer participants had varied levels of knowledge about cameras and were selected based only on English-speaking ability (due to the use of speech recognition in the study). We employed a total of 9 customer participants (8 male, 1 female, average age 34.1, s.d. 3.9).

To select a participant for the role of a proactive shopkeeper, we interviewed participants and observed trial interactions. We asked customer participants to provide feedback in terms of how proactive, helpful, and interested each shopkeeper was. We selected one shopkeeper participant (male, age 54) with a naturally outgoing personality and a great interest in cameras based on our interview with him, as well as the feedback from the customers. He played the shopkeeper in all interactions.

3.4 Procedure

For this data collection, the shopkeeper was encouraged to answer any questions the customer had, and also to take initiative in assisting the customer, either by introducing new camera features or presenting a different camera. The customer participants were instructed to browse as much or as little as they liked, and told that they could ask questions about cameras or simply listen to the shopkeeper's recommendations.

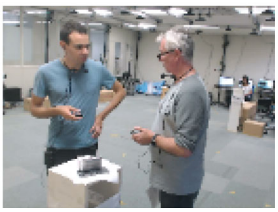
To create variation in the interactions, customer participants were asked to role-play in different trials as advanced or novice camera users, and to ask questions that would be appropriate for their role. Some camera features were chosen to be more interesting for novice users (color, weight, etc.) and others were more advanced (High-ISO performance, sensor size, etc.), although they were not explicitly labeled as such.

Customer participants were not given a specific target feature or goal for the interaction, as we were mostly interested in capturing the shopkeeper's proactive sales behavior. All participants were instructed to focus their discussion on the 8–10 features listed on the camera spec sheet, to minimize the amount of "off-topic" discussion.

Customer participants conducted 24 interactions each (12 as advanced and 12 as novice) for a total of 216 interactions. 17 interactions were removed due to technical failures of the data capture system and one participant who did not follow instructions. The final data set consisted of 199 interactions, with average duration of 3 min and 16 s per interaction. This includes a total of 2568 shopkeeper utterances (with an average of 19.53 words per utterance) and 2299 customer utterances (with an average of 10.88 words per utterance). This data set is available online.[1]

[1] http://www.geminoid.jp/dataset/camerashop/dataset-camerashop.htm.

Table 1 An example interaction from the data collection

C: Customer, S: Shopkeeper
(Shopkeeper and Customer talking about Sony)
C: Proably yeah.. I'm looking for something lightweight but good quality.
S: This Sony camera takes pictures that are almost as good as a top-end professional camera but it only has half the weight, please pick it up and have a try.
C: Yeah actually this weighs alright…how much is it?
S: This is only $550 so it's a quarter to a fifth of the price of a high-end professional camera
C: *(Silence for 5 s)*
S: the picture quality is almost as good as DSLR because it's a mirrorless camera …..
 

3.5 Observed behavior

Overall, the shopkeeper participant followed our suggestions and acted in a very proactive way. He often spoke in long, descriptive utterances and volunteered extra information when answering questions. In cases where a customer was silent or not asking questions, he frequently provided additional information about a camera or guided the customer to a new camera, so we considered his behavior to be fairly proactive and thus appropriate for this study.

This interaction data differed from that of the previous study (Liu et al. 2016) in a few ways. First, the shopkeeper's utterances tended to be much longer and more complex, sometimes talking about two or three topics in one sentence. Second, the shopkeeper often proactively spoke if some silence had elapsed after his last utterance. Third, the customers demonstrated more "backchannel" utterances. For example, a customer might say, "oh, ok," after listening to an explanation, but not ask a follow-up question. In such situations, the shopkeeper in this study often performed proactive behaviors, such as volunteering more information about the current camera or continued his previous explanation.

We performed an analysis of the customer utterances to identify whether an utterance required a response (such as a question or a request) or did not require a response (such as a backchannel utterance). We found that 527 (22.8%) of the customer's 2299 utterances did not seek a response from the shopkeeper. There were also 209 instances when the customer did not speak or move for some time, such as when reading the spec sheet or playing with the camera, and the shopkeeper took the initiative to perform some proactive behavior.

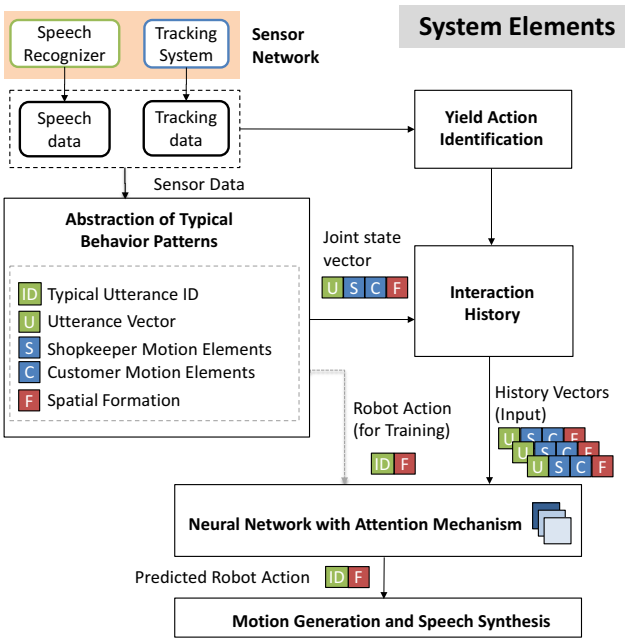


Fig. 2 Overview of the proposed system elements

Table 1 illustrates an example interaction. The customer first asks about a lightweight camera, prompting the shopkeeper to show the customer to the Sony camera. The shopkeeper then answers the customer's question about the price. Next, after several seconds of silence, the shopkeeper proactively presents more information about a different feature. Similar to the provided example, we observed that many customers used a variation of fillers (e.g. "you know", "like") and backchannel (e.g. "I see") in their utterances. In addition, some customers did not just ask direct questions, but also

provided other information (e.g. "Yeah actually this weighs alright how much is it?"). For these reasons, we consider the interaction data to be quite natural and fairly unconstrained.

4 Proposed technique

4.1 Overview

In order to reproduce both reactive and proactive behaviors for a robot, we used a sequence of techniques that enable behavior contents and interaction logic to be directly learned from noisy sensor data without human intervention. An overview of the techniques is shown in Fig. 2, which illustrates how behaviors are learnt from human–human interaction and generated in human–robot interaction. The key steps of the techniques are listed here:

1. *Abstraction of typical behavior patterns* (Sect. 4.2) Continuous streams of sensor data are abstracted into typical behavior patterns, and the corresponding *joint state vector* and *robot action* are defined.
2. *Defining yield actions* (Sect. 4.3) To enable the robot to generate proactive behavior, we introduce the concept of a *yield action*, which represents the moment when an interactant yields his turn and does nothing, allowing the robot to take initiative.
3. *Incorporating interaction history* (Sect. 4.4) We introduce interaction history by concatenating the last k joint state vectors to provide contextual information for generating proactive behavior.
4. *Learning to attend to history* (Sect. 4.5) To improve the efficiency of learning, we propose the use of an "attention" mechanism which ascribes weights to the relative importance of various steps of interaction history as inputs to learn appropriate behaviors.

In this work, we used the techniques presented in our previous study (Liu et al. 2016) for Step 1, while Steps 2–4 constitute the novel contributions of this work which enable proactive behavior generation.

4.2 Abstraction of typical behavior patterns

In order to learn effectively despite the large variation of natural human behaviors and noisy inputs from the sensor system, the continuous stream of captured sensor data needs to be discretized by time into behavior events, and then abstracted into common behavior patterns. Here we briefly describe our techniques:

- We used unsupervised clustering and abstraction to identify utterance vectors, typical utterances, stopping locations, motion paths, and spatial formations of both participants in the environment.
- An interaction is discretized into a sequence of actions, which are defined whenever: (1) a participant speaks an utterance and/or (2) a participant's motion target changes.
- For each action detected, the abstracted state of both participants at the time is represented as a *joint state vector*, with features consisting of their abstracted motion state the utterance vector of the current spoken utterance.
- For each observed shopkeeper action, we define a corresponding executable robot action, consisting of a typical utterance (e.g. ID 5) and a target spatial formation (e.g. *present Nikon*). When executed, this would cause the robot to speak the typical utterance "It's $68" associated with utterance ID 5 and execute a motion to attain the formation of *present Nikon*.

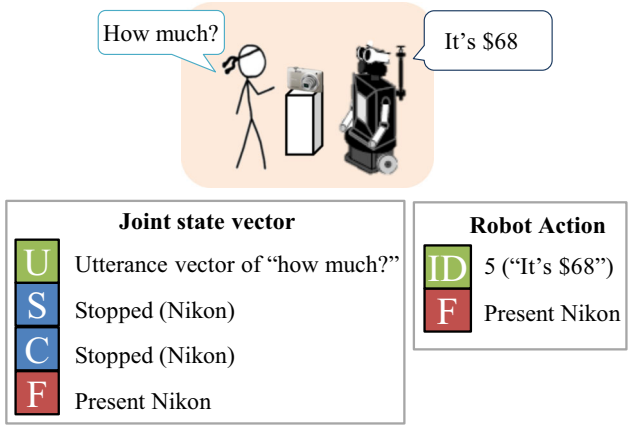


Fig. 3 Example of abstraction for joint state vector and robot action

Figure 3 shows an example of how *joint state vector* and robot action are abstracted from the sensor data. These data processing and abstraction techniques closely follow the procedure followed in our previous work (Liu et al. 2016), and additional details are presented in the "Appendix".

4.3 Definition of yield actions

To enable the robot to predict the timing when a proactive action should be generated, we define a *yield action*. A *yield action* represents a moment when an interactant is yielding the floor, providing an opportunity for a proactive behavior to be executed (Duncan 1974, 1972). In our training data, the customer was sometimes occupied with playing with the camera or reading the spec sheet, or sometimes just decided not to do anything, and thus did not speak or move for some time, indicating that the customer may have relinquished his turn. As observed in 209 instances from our training examples, the shopkeeper often seized the opportunity to do

something proactive, usually by introducing another feature or camera.

In the training data, we define the customer to have yielded his turn whenever we observe two consecutive occurrences of shopkeeper actions, based on the findings presented by Duncan (1972) and our observation that the shopkeeper proactively performed another action after his previous action. For example, after a shopkeeper speech action (e.g. answering a question), if the subsequent observed action is another shopkeeper speech action (e.g. talking about a camera feature), we can assume that a customer *yield action* has occurred between the two shopkeeper actions. Likewise, this strategy can be applied for the detection of a shopkeeper *yield action*.

The next task is to identify *yield actions* in the real-time system. Turn-taking is a complicated problem, involving gaze, prosodic, linguistic, and gestural signals as well as timing, but for the current study we make the simplifying assumption that we can detect a *yield action* using a timing threshold. This assumption has been made in HRI (Thomaz and Chao 2011; Chao and Thomaz 2011) and other spoken dialogue systems as well (Raux and Eskenazi 2008). To determine a time threshold for identifying yield actions, we computed the average amount of time elapsed between two consecutively observed shopkeeper actions in the training data. This value was calculated to be 3.52 s. Thus, in our system, we defined a customer *yield action* to occur if the customer did not begin speaking or moving within 3.52 s after the end of the previous robot action.

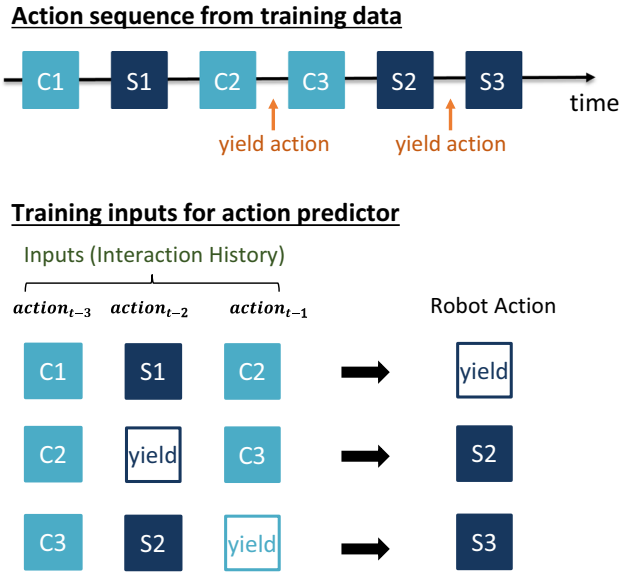


Fig. 4 Example of how actions are identified in the training data. A yield action is identified whenever two consecutive actions from the same participant without any action detected in between

4.4 Incorporating interaction history

Although single-step prediction might be sufficient for answering questions, there are many situations where context is important. For example, an answer to a customer's question such as, "how much does this cost," can be generated based on the most recent customer utterance and spatial location—information from interaction history is not necessary. However, after a customer *yield action* or a statement or backchannel utterance such as "Okay," or "I see", the customer's action does not contain information which uniquely determines a robot response. In such cases, an appropriate proactive shopkeeper action will depend to some degree on the previous interaction context. Some examples of history-dependent behavior include the following:

- After a customer *yield action*, the robot could continue to provide information about the last feature presented, or present a new feature not previously discussed. Both cases are dependent on the robot's previous utterance.
- There may be an inherent sequence to robot behaviors, e.g. first introducing and moving to a new camera, then offering for the customer to pick it up and try it, so the robot's second action depends on its previous action.
- When the customer answers a question, e.g. by saying "yes," the robot's next action depends on both the customer's answer and the question that was asked.

To address these cases, we propose the use of *interaction history* to enable the robot to determine an appropriate action for a given context. History can be represented in various ways, and including more information increases the dimensionality of the input vector and hence the difficulty of the learning problem. For the amount of training data available in our study, 3 steps of history seemed to be sufficient to enable the robot to learn proactive behaviors such as those described above.

Thus, we include the three most recent discrete actions as inputs to the classifier. Once an action is detected, a *joint state vector*, describing the state of both interactants at the time, is appended to the *interaction history*, which is kept at a fixed size of 3 steps. Figure 4 shows an example of how customer and shopkeeper actions from the training data are segmented into sets of 3 action vectors $(action_{t-3}, action_{t-2}, action_{t-1})$ to be used as inputs for training the behavior predictor. The subsequent shopkeeper action is represented as a robot action vector, and it is used as the training output for the predictor. In this way, interaction history segments are used to train the robot to predict an appropriate action.

4.5 Learning to attend to history

While including interaction history provides valuable context for predicting proactive behavior, it also increases complexity and noise, and thus considerably slows the rate of learning (Cover and Hart 1967). The inclusion of irrelevant information may thus hinder the robot's ability to learn correct behaviors.

To help the system learn more effectively, we can exploit the fact that some behaviors are more dependent upon specific steps of history than others. For example, answering a customer's direct question about a camera feature is primarily dependent only on the customer's most recent utterance, that is, $action_{t-1}$. On the other hand, when a customer yields the turn and the robot generates a proactive behavior, the decision is more likely to be dependent upon the robot's own previous action, $action_{t-2}$, and possibly also the customer's previous action, $action_{t-3}$. In the case where the customer says "yes" when the robot asks for confirmation, the decision may depend most heavily on $action_{t-2}$. If the predictor can be trained to focus only on the most relevant steps of history, it may be possible to improve the efficiency of learning.

To achieve this, we applied a recently introduced architecture in the deep learning field, a feed-forward deep neural network with an *attention mechanism* proposed by Raffel and Ellis (2015). For each possible training label, the *attention mechanism* takes each input in the sequence and learns an adaptive weighted average based on each input. This value can be thought as the "relevance" of the inputs, according to the context. Thus, this method has the capability to learn which part of interaction history is relevant for generating a robot action, and also the advantage of visualizing into the neural network to see which part of the history the network is attending to.

Figure 5a shows the schematic of the deep neural network, where the training input is the interaction history, consisting of an input sequence of the three most recent *joint state vectors*, $X = \{jsv_{t-3}, jsv_{t-2}, jsv_{t-1}\}$. The activation value of neuron j in layer l is defined in Eq. (1)

$$h_j^{(l)} = \sigma\left(\sum_k w_{j,k}^{(l)} \cdot h_k^{(l-1)} + b_j^{(l)}\right) \tag{1}$$

where $b_j^{(l)}, w_{j,k}^{(l)} \in \mathrm{R}$ are free parameters, $h_k^{(l-1)}$ is the activation (output) of neuron k in layer $l-1$, and σ is a nonlinear activation function.

The attention mechanism, a_j, is computed using a single layer perceptron and then a softmax operation to normalize the values between zero and one, as expressed in Eq. (2).

$$\begin{aligned} \gamma_j &= \tanh\left(W_a h_j^{(l)} + b_a\right) \\ a_j &= softmax\left(\gamma_j\right) \\ c &= \sum_{t=1}^{T} a_j h_j^{(l)} \end{aligned} \tag{2}$$

The idea is that once we have an activation value of neuron j in layer l, $h_j^{(l)}$, we can query each value asking how relevant they are to the current computation of the target class assignment. $h_j^{(l)}$ then gets a score of relevance which can be turned into a probability distribution that sums up to one via the softmax activation. We can then extract a context vector, c, that is a weighted summation of the activation value in layer l depending on how relevant they are to a target robot action (see Fig. 5b). Thus, the value of a_j, describes how much of each step in the *interaction history* should be considered for each robot action. For example, if a_{t-1} is a large number, this would mean that the DNN pays the most attention to the most recent step of the *interaction history*, and thus is important for predicting the robot action.

Here we describe the hyperparameters of our neural network. The dimension of the input layer is three sets of input neurons of size m ($m = 1244$) from the joint state vectors, followed by two leaky rectified hidden layers, an attention layer, and another leaky rectified hidden layer. The output layer is a softmax with the number of neurons equal to the number of possible robot actions (761), which represents the probability of a robot action given an interaction history input. The number of neurons for each hidden layer is 800. There was no pruning or dropout layer applied in our neural network architecture. The weights of $b_j^{(l)}, w_{j,k}^{(l)}$ is optimized by momentum-based mini-batch stochastic gradient descent, with batch size of 128, learning rate of 0.005, and momentum coefficient of 0.9, and learning decay is 10^{-9}. Initial weights for a neuron in layer l are sampled from a normal distribution, where the biases start at 0.

Figure 6 depicts an example interaction during online operation of the system. When a speech or *yield* action is detected, the interaction history, consisting of three *joint state vectors,* is sent as a query to the trained DNN, which updates an attention value for each input. The neural network then predicts the probability for each robot action and outputs the robot action with the highest probability for execution.

4.6 Examples of using the attention mechanism

Here, we would like to illustrate some examples of our system with the attention mechanism. One feature of the attention mechanism is that the value of $a\,(h_t)$ provides us with a way to visualize which step of the input sequences the neural network is attending to. The higher the value of $a\,(h_t)$ for a

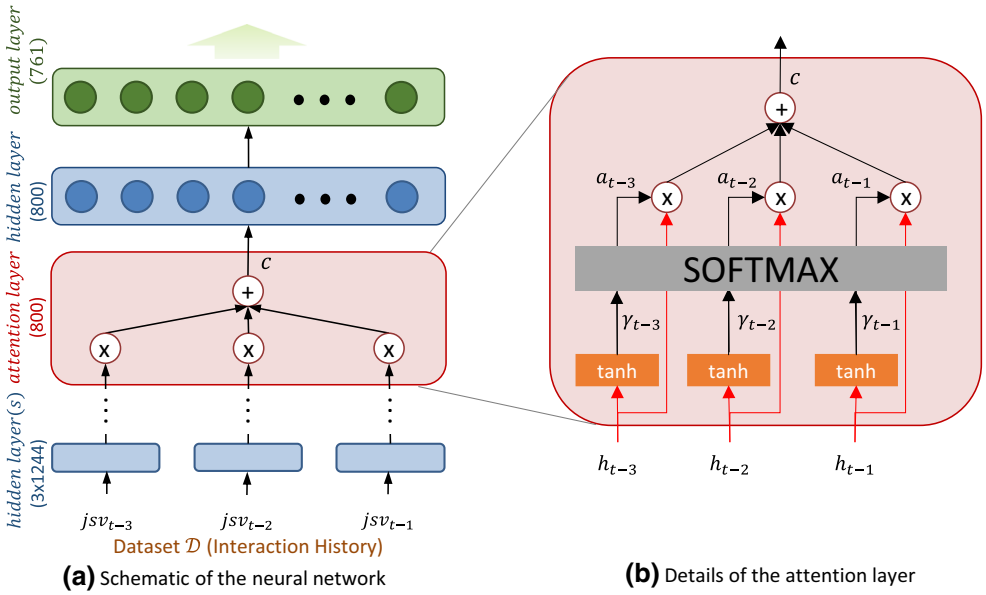


Fig. 5 **a** Schematic of the multilayer perception neural network: Interaction history is inputted to the neural network as joint state vectors and robot action as training target for the neural network. The output dimension of each layer is shown in parenthesis. **b** Details of the attention layer: The context vector is a weighted summation of the activation value in layer l and represents how relevant parts of the input is to predicating the robot action

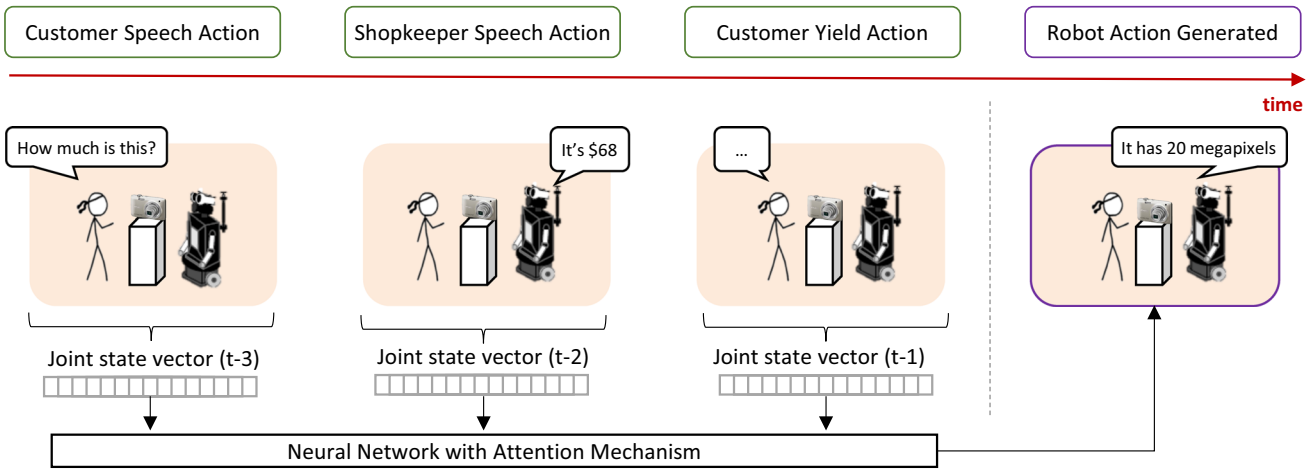


Fig. 6 An example of how actions are discretized and represented as joint state vectors in the interaction history during online operation of the system. A customer yield action is generated when no action has been detected for 3.52 s since the last robot action

certain step in the *interaction history*, the more it is considered for predicting a robot action.

Figure 7 shows these values for some example predictions, in which darker shades of blue represent higher attention values. For simplicity, only utterances are shown, although our system uses spatial data as well. These examples were generated by taking a sequence of three actions from the training data (customer—shopkeeper—customer) and feeding them into the trained DNN to predict an output shopkeeper utterance.

Example 1 illustrates a case where the customer asks a question. The attention model selects the most recent customer utterance as the most important factor for predicting the robot's answer. In Example 2, the attention model chooses the customer's previous utterance as the most relevant when customer says a "backchannel". We hypothesize that this is because the customer's previous question helps to define the set of proactive behaviors which would be appropriate in this context. Lastly, in Example 3, the system detects a customer *yield action*, and the attention model chooses the shopkeeper's previous utterance as the most relevant input.

We observed that the robot was able to learn the appropriate behavior due to interaction history, which would not have been possible if the robot was only to predict based on the most recent customer action, that is, the customer *yield action*.

These examples show some successful predictions, but we are not claiming that the attention mechanism will work for all situations. These examples were chosen because they illustrate that an attention model such as this could be a useful tool for visualizing a black-box system like a DNN.

5 Offline evaluation

Before evaluating our system with a live robot, we performed an offline evaluation of the behavior predictor through cross-validation with the training data, in order to confirm the effectiveness of the proposed inclusion of history and attention in the learning mechanism.

5.1 Evaluation procedure

A multi-fold cross-validation data set was generated by randomly selecting 10% of the data from the dataset, together with the following shopkeeper behavior which was to be predicted. The remainder of the training data, 2223 customer–shopkeeper–customer behavior sequences, excluding the selected sequences, was used for training the predictors. The test data from the multiple runs are aggregated together, for a total of 500 behavior sequences as evaluation data.

Five predictor variants were evaluated. All evaluations included the proposed detection of *yield action*, and the conditions differed by the type of classifier, the inclusion of history, and the use of the attention model.

1. *NB-1* A Naïve Bayesian classifier trained on the most recent single customer action. This was the classifier from the previous study, so we designated it as the baseline for comparison.
2. *NB-3* A Naïve Bayesian classifier trained with history (i.e. the most recent three steps of actions: customer–shopkeeper–customer).
3. *DNN-1* A DNN trained on the single most recent customer action.
4. *DNN-3* A DNN trained with history (i.e. the most recent three steps of actions: customer–shopkeeper–customer).
5. *DNN-3-AM* A DNN trained with history, which also incorporated an attention mechanism, as described above.

Normalized initiation, described by Ioffe and Szegedy (2015), was used to initialize the batch inputs of the DNN in (3)–(5). The networks were trained to minimize the cross entropy loss for 10,000 epochs between the target output and the observed output for the entire training set.

Example 1: Answering questions at Nikon (reactive)

	C: [*yield action*]
	S: its only $68 and great camera for all the family anyone can use it
	C: what color do you have for this camera?
Predicted: "this one comes in purple pink black silver and red."	

Example 2: Presenting unsolicited information (proactive)

	C: And what about the color of this camera?
	S: It comes in black, white, and silver.
	C: I see.
Predicted: "You can upload directly to Facebook through a wireless link."	

Example 3: Introducng Nikon at Sony (proactive)

	C: [*yield action*]
	S: over here we have the Nikon.
	C: [*yield action*]
Predicted: "picks up and take a few pictures if you like it set up to be point and shoot." (move to Nikon)	

Fig. 7 Examples of successful predictions using our attention mechanism technique with a history length of three. Shaded boxes show the relative weight of $a\,(h_t)$ from DNN assigned to each action, indicating its importance in predicting the final prediction. Darker shading indicates higher weight

To perform this comparison, we evaluated the "social appropriateness" of the predicted behaviors, rather than simple prediction accuracy, because many equally acceptable utterance behaviors exist in the data set. For example, "$2000", "it's only $2000", and "the camera body is only $2000", are all valid answers to the question of the price of one of the cameras. This approach is similar to the procedure used in Liu et al. (2016) for evaluating appropriateness of robot behaviors.

A human coder, naïve to the experimental conditions, rated each prediction as "acceptable" or "unacceptable". Unacceptable behaviors included factually incorrect responses, failures to answer a question, strange behaviors like moving to a new camera while a person was waiting for a response, and repetition of the previous behavior if not appropriate to do so.

As these ratings require subjective judgment, we confirmed the consistency of the coder's evaluations by asking a

Table 2 Results of manually-coded cross-validation comparison

Classifier	Behavior correctness (%)	*p* value
NB-1 (baseline)	56.2	–
NB-3	39.0	$< .001$
DNN-1	60.2	N.S.
DNN-3	61.8	N.S.
DNN-3-AM	62.4	$< .05$

The result of DNN-3-AM showed a significant difference when compared with the baseline system

second coder to independently rate the same data set. Their results were compared, and a Cohen's Kappa value of 0.80 was calculated, indicating very good interrater reliability, so we consider the coder's ratings to be reliable.

5.2 Results

To evaluate statistical significance of differences between the conditions, a chi-squared test was performed, comparing each of the classifiers against the NB-1 (baseline) classifier. The results of this comparison are shown in Table 2.

For the *NB-3* classifier, the chi-squared test showed significance [$\chi^2(1, N = 500) = 28.63$, $p < .001$] indicating that simply adding history to the Naïve Bayes classifier resulted in significantly worse performance than simple single-step prediction. For the *DNN-1* classifier, a chi-squared test did not show statistical significance, [$\chi^2(1, N = 500) = 1.46$, $p = .227$]. The performance of the *DNN-3* classifier again did not show a significant difference from the baseline in a chi-squared test, [$\chi^2(1, N = 500) = 2.75$, $p = .097$]. The proposed *DNN-3-AM* classifier provided the highest performance, and a chi-squared test showed a significant difference from the baseline, [$\chi^2(1, N = 500) = 4.45$, $p = .035$].

This evaluation shows that simply adding history as inputs to the original *NB-1* classifier resulted in significantly worse performance, whereas the proposed *DNN-3-AM* technique incorporating both history and the attention model, performed significantly better than the baseline predictor.

Although overall performance was lower than we had hoped, we believe performance would improve significantly with better speech recognition and more training data.

6 User study

To observe the effect of the new proposed features in live interaction, we conducted a user-study to compare the two conditions: (a) *proposed*, using customer *yield actions* and the DNN-3-AM classifier, and (b) *baseline*, a system using the NB-1 classifier and not using customer *yield actions*.

6.1 Hypothesis and prediction

In the evaluation experiment, we made the following hypotheses about the effects of our proposed techniques:

1. Identifying customer *yield actions* will lead to the user to perceive the *proposed* system as more proactive, since the robot is able to identify when it should take an action.
2. Using DNN-3-AM classifier will enable the robot to generate behaviors that are context-sensitive and therefore more contingent to the user's action in the *proposed* system, thus the robot will behave in a more socially-appropriate way.
3. Overall, this will lead users to perceive the interactions to be better in terms of quality using our *proposed* system, since proactive behavior and responding appropriately to the user's actions are desirable in service interactions.

6.2 Experiment setup

6.2.1 Participants

A total of 15 paid participants (11 male and 4 female, average age 31.3, s.d. 2.37) played the role of customer in the experiments. All of them were fluent English speakers.

6.2.2 Environment

The experiment was conducted in the same camera shop setting used for the data collection, with three digital cameras displayed in an 8 m × 11 m experiment space. The same sensor network was used for tracking, and the participants communicated with the robot using an Android phone for speech recognition.

6.2.3 Robot platform

For this experiment, we used Robovie 2, a humanoid robot with a 3-Degree-of-Freedom (DOF) head, two 4-DOF arms, and a wheeled base capable of moving at 0.7 m/s. For motion planning, the dynamic window approach (DWA) was implemented to avoid obstacles (Fox et al. 1997). The Ximera speech synthesis system (Kawai et al. 2004) was used to generate its speech.

Idle motion behavior was implemented in the robot for both conditions, consisting of small arm and head movements while idling, speaking, and moving (Shi et al. 2010). Automatic gaze tracking was also implemented, and the robot followed the customer with its gaze during all interactions.

6.2.4 Procedure

We compared the robot's performance between two conditions: *proposed* and *baseline*. For each condition, we asked participants to role-play for 4 trials. To create variation in the interactions, the participants were asked to role-play as: (1) a need-based customer (2 trials): who was looking for features as either someone familiar or unfamiliar with cameras, and (2) a quiet customer (2 trials): who was not looking for anything in particular and didn't have much to say, and was encouraged to read the spec sheets or play with the cameras. In all trials, they were encouraged to walk around the shop and show an interest in learning about camera features. The order of the conditions was counterbalanced and the order of the trials within each condition was randomized.

As in our data collection, participants were asked to pretend to be a first-time customer in the camera shop for every trial and the participants performed 2 sample interactions before the experiment to become familiar with the Android phone interface and confirm their understanding of the instructions.

After the 4 trials in one condition were completed, the participant answered a questionnaire. The procedure was repeated with the remaining condition (*baseline* or *proposed*).

6.3 Measurement

Before the experiment, we explained to each participant that the goal of this project was to create a proactive robot shopkeeper which could assist customers in a camera shop, and they were asked to evaluate how well the robot was able to demonstrate that proactivity. After the experiment, we had each participant fill out a written questionnaire, rating the following items on a 1–7 scale (1 being very negative and 7 being very positive):

- How proactive was the robot's behavior?
- How socially appropriate were the robot's behaviors?
- Overall evaluation

After the questionnaire was completed, the participants were interviewed to gain a deeper understanding of their opinions of the robot's behavior.

6.4 Results

6.4.1 Questionnaire results

Figure 8 shows questionnaire results from the participants. To compare each rating between the *proposed* condition and the *baseline* condition, we conducted a repeated-measures ANOVA for each of the three questions.

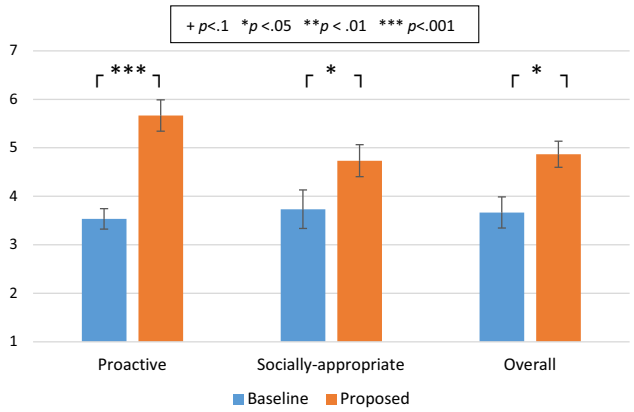


Fig. 8 Results of the robot behaviors in user study evaluation. The bar in the graph represents standard error

We verified that all of our predictions were supported, as this analysis found significant differences between the conditions for all ratings: "Proactivity" [$F(1, 14) = 28.332$, $p < .001$], "Social Appropriateness" [$F(1, 14) = 5.250$, $p = .038$], and "Overall evaluation" [$F(1, 14) = 7.875$, $p = .014$].

1. The results support our hypothesis that the participants would perceive the robot to be more proactive using the *proposed* system than the *baseline* system.
2. The results support our hypothesis that participants would perceive the robot to be more socially appropriate with our *proposed* system than the *baseline* system.
3. The results supported our hypothesis that the *proposed* system would lead to a better overall interaction than with a *baseline* system.

6.4.2 Qualitative observations

We observed a number of qualitative differences between the behaviors of the *proposed* robot and the *baseline* robot.

Approach The *proposed* robot would typically take the initiative to approach a customer standing at a camera. In contrast, the *baseline* robot typically waited at the service counter until the customer asked a question.

Introducing features and other cameras The *proposed* robot would proactively introduce camera features to the customer without being asked, e.g. saying: "pick it up see how light it is it is only 120 grams", or proactively lead the customer to a new camera. In contrast, the *baseline* robot would answer questions, but not take any initiative to talk about camera features or introduce new cameras. Rather, it stood silently by the customer when the customer had nothing to say to the robot.

Context-dependence We observed cases where the *proposed* robot was able to generate behaviors dependent on

context or interaction history. For example, in one case the *proposed* robot asked a customer who was looking to take travel pictures, "so you need a camera you can take anywhere use easily". With the customer's response of "yes yes I need that", the robot then introduced the smallest, most lightweight camera. We believe this illustrates the value of incorporating interaction history, as the customer's utterance itself contained no information about which camera would be appropriate.

The example transcript of the *proposed* robot interacting with a quiet customer shown in Table 3 illustrates how the robot was able to answer questions (reactive behavior) and proactively explain new features (proactive behavior). Additional examples of human–robot interactions can be seen in the accompanying video attachment.

6.4.3 Interview results

From our interview results, many participants thought both *proposed* and *baseline* robots were friendly. Many participants commented that they felt more engaged with the *proposed* robot because it proactively asked them questions (e.g. "what sort of pictures do you take?") and talked about camera features while they were playing with the camera. One participant said that he liked when the *proposed* robot initiated conversation, since he was unsure what to say to a robot in a shop. Many participants also commented that the *proposed* robot seemed more approachable, attentive, and aware.

It is interesting to note that some participants preferred the interaction style of the *proposed* robot more than the *baseline* robot. One participant said the *baseline* robot reminded her of a surveillance system, where the robot is watching to see if she has damaged any goods. Another participant felt annoyed by the *baseline* robot, as it followed him around the shop, but did not say anything to him when he was looking at the cameras.

7 Discussion

7.1 Contribution

In this study, we demonstrated that the robot was able to generate both reactive and proactive behaviors from examples of human–human interaction. We showed that the robot was able to not only answer questions, but also proactively assist the customer by introducing new features or a new camera. The robot was also able to respond based on interaction context, even when what the customer just said contained very little information (e.g. "yes please"). Through an offline evaluation and a user-study evaluation, we demonstrated that the robot was perceived as more proactive, more socially-appropriate, and better overall with our proposed techniques, as compared to a baseline system that did not use our techniques.

7.2 Identifying yield actions in turn-taking

In this study, we demonstrated that proactive behavior can be generated by identifying *yield actions* based on a timing threshold. While we demonstrated this approach to work well in our situation, we believe that this technique can be improved by including other ways of identifying *yield actions*. For example, nonverbal behaviors such as gaze and nodding have been investigated as turn-taking signals in both psychological (Duncan 1974; Gu and Badler 2006) and HRI studies (Rich et al. 2010; Mutlu et al. 2009). Thus, the detection of non-verbal feedback for a more natural turn-taking behavior in a robot could be interesting to explore in future work.

7.3 History representation

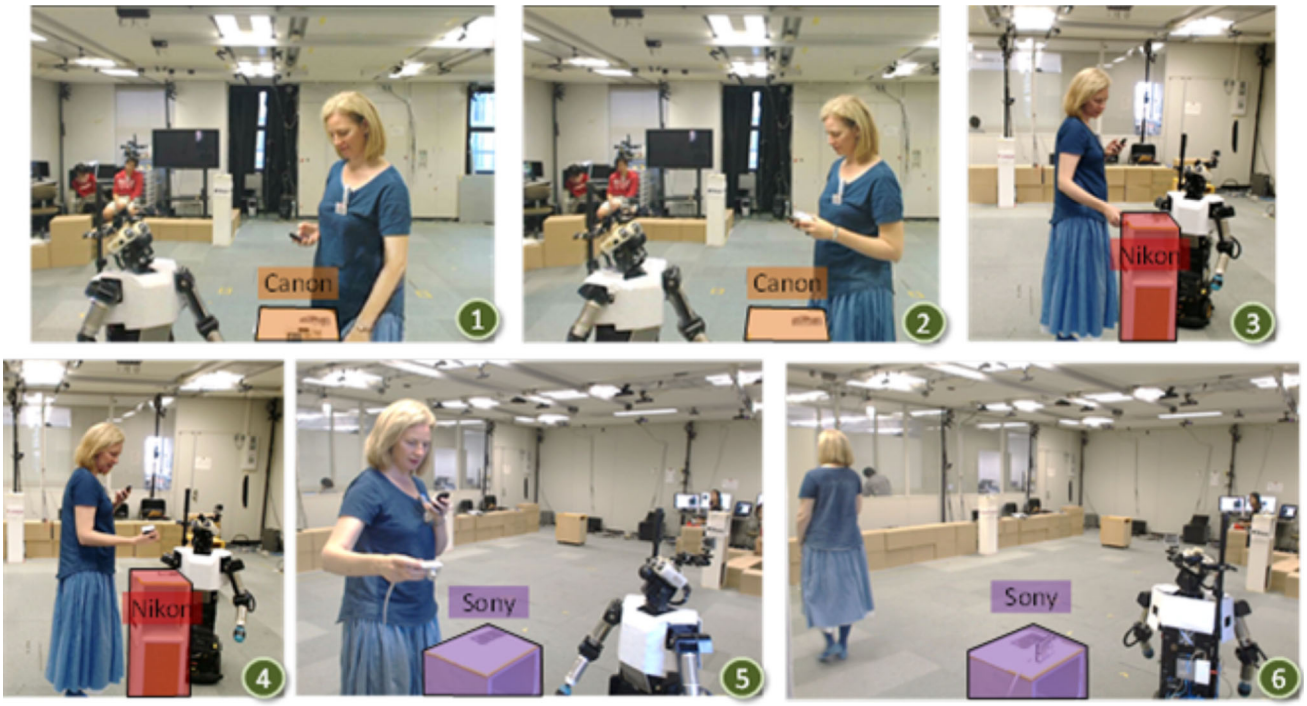
In our scenario, we demonstrated that the robot was able to reproduce the behaviors of a proactive shopkeeper with a fixed length of three history steps with our *proposed* system. While the choice of three history steps was enough for our scenario, we expect that additional benefits could be gained by increasing the length of history or otherwise representing long-term history in some way. For example, sometimes the customer would state their goal at the beginning of an interaction, "I am looking for a camera that is easy to carry around". Since only the immediate history was used for training and generating robot behavior, this information would be lost over time.

Choosing a history representation is a difficult problem. If the interaction history is too long, the robot may learn some additional context-dependent behavior, but it becomes more difficult for the system to learn to ignore history for simple question-answer exchanges. One possible future improvement may be to explicitly model a customer's intention or goals to capture this long-term history. Although such questions can be explored in future work, our current study has demonstrated that including just the immediate history reproduced reasonable proactive behaviors for the dataset we have.

7.4 Generalizability and scalability

We believe that this data-driven approach can be applied in domains where repeatable interactions can be captured, and where proactive behaviors are context-dependent. For instance, the task of an art museum tour guide robot includes answering questions about a particular artwork (e.g. facts about the artist), as well as proactively explaining about other interesting anecdotes about that piece (e.g. the medium used

Table 3 An example of the proposed robot interacting with a quiet customer in the user study



(1) [Customer and Robot at Canon]

Customer: "can you tell me a little bit about this camera?"

Robot: "you have full creative control it has every possible manual setting"

(2) **Customer** picks up the camera and plays with it

Robot: "we have one set up over here with a small zoom lens if you would like to try using it"

(3) **Customer:** "ok thank you" [walks to Nikon]

Robot: [moves to Nikon]

Customer reads the spec sheet

Robot:"would you like to take a couple pictures with at first"

(4) **Customer**: [plays with the camera]

Robot: "here is the optical zoom so you can see the effect it has on your picture too"

Customer: [continue playing with the camera]

Robot: "the two most important things with this kind of cameras that the pictures are great quality because everybody's going to look at them and then I can respond very quickly because it's all about being out with friends and family"

Customer: "yeah that's true"

Robot: "it comes in a range of colors it takes fantastic pictures it's really easy to use so you can focus on the photograph instead of all the ways that the camera can be set"

(5) **Customer**: "oh thanks" [walks to Sony]

Robot: [moves to Sony] "good afternoon how can I help"

Customer: [plays with the camera]

Robot: "it's an excellent camera that takes the same quality pictures as a top-end camera without the top and price"

(6) **Customer:** "okay well thanks so much for the information" then leaves the shop

Robot: returns to service counter while saying "no problem have a good afternoon"

or time period completed). We can also imagine a tourist center robot, where its tasks could include both answering questions about a tourist attraction (e.g. operating hours) and expatiating about other details (e.g. admission cost).

There may be some domains to which our approach cannot be generalized. These domains might require proactive behaviors that are dependent on subtle social cues or background knowledge. One example might be an educational robot that proactively teaches a language, where the lesson is tailored to the student's comprehension level. We imagine such domain would be difficult to learn with our current approach, since such framework containing the knowledge about a user (i.e. level of comprehension) is not represented in our system.

In terms of scalability with our proposed system, we believe that it will be able to scale up to more complex scenarios, for instance, when the number of cameras on display increases. The amount of training data required will be dependent on the number of social behaviors that need to be reproduced, the variability of the customer actions, and the reliability of sensing, thus training effort would scale linearly with the number of behaviors to be learned.

7.5 Limitations

While we have demonstrated a system for learning robot behaviors from a proactive shopkeeper, the offline evaluation shows there are some limitations to the current system. Below we discuss some limitations and possible strategies for future improvement.

Repeatability of actions This technique is designed to work for social scenarios containing many repeatable actions, and the most frequently-observed actions will be learned best. Actions that are very infrequent or unique in the training data will not be learned well. This is an inherent limitation of a learning-by-imitation approach, and it could be valuable to develop methods for quantifying the degree of repeatability in a set of interactions. This could be useful for judging when sufficient training data has been collected to reproduce an interaction, or for deciding whether this approach is applicable to a new social scenario.

Compound utterances The shopkeeper often spoke about multiple features in one utterance (e.g. "This has a 9 preset modes and it also has a 3200 ISO" and "This has 9 presets and is $550"), which means that utterances that are not exactly semantically similar may end up being clustered together, and consequently mapped to the same robot action. For future work, we envision improving the clustering algorithm (e.g. using a soft clustering algorithm to expose more information about the probability distribution of an utterance belonging to a robot action) or techniques in natural language processing to better handle more complex utterances.

Representing other modalities Modalities such as gaze and gesture are often important in social interaction. For example, the human shopkeeper sometimes introduced a camera by pointing to it instead of actually moving to that camera. This pointing behavior is not recognized by our sensors and thus not learned by the robot. Consequently, this led to some confusing situations where the robot would talk about a camera other than the one it was standing at. It would be interesting to incorporate additional perceptual (Nickel and Stiefelhagen 2007) and generative (Sugiyama et al. 2007) modules for additional modalities, such as pointing or gaze.

8 Conclusion

In this work we have successfully demonstrated a system designed to reproduce not only reactive behaviors for a robot (e.g. answering questions), but also proactive behaviors (e.g. providing unsolicited information) that are learned from human–human interactions. This was accomplished through three proposed techniques, including detection of yield actions, incorporating interaction history, and using an attention mechanism to learn which history steps are important for predicting the robot behavior. First, we demonstrated that our proposed technique was rated the highest in terms of behavior correctness among five different methods for predicting robot behaviors. Then, we validated our approach in a comparison user-study, which showed that participants perceived the proposed techniques to produce behaviors that were more proactive, socially-appropriate, and better in overall quality.

Social robots are now appearing in the real world, and we are seeing a growing market in the service industry for robots which interact with customers. In such situations, proactive behavior may prove necessary to enable robots to effectively engage with their customers and users. In this work we have successfully demonstrated one way in which a data-driven approach from our previous work can be extended to reproduce proactive behaviors from a human shopkeeper, and we believe that data-driven techniques like these will become a valuable tool for building real-world interaction logic for social robots.

Acknowledgements This work was supported in part by the JST ERATO Ishiguro Symbiotic Human-Robot Interaction Project, Grant Number JPMJER1401, and in part by JSPS KAKENHI Grant Number 25240042.

Compliance with ethical standards

Ethical approval This research was conducted in compliance with the standards and regulations of our company's ethical review board, which requires each experiment to be subject to a review and approval procedure according to strict ethical guidelines.

Appendix

Here we describe our data abstraction techniques to enable the learning of high-level interaction logic in human–robot interaction to be achieved in an entirely data-driven way, that is, without any kind of manual annotation or cleanup of the sensor data. This follows the work presented in Liu et al. (2016).

Defining input features

Here, we describe the features used in the *joint state vector*, including the abstraction of motion (consisting of *current location*, *motion origin*, and*motion target* of both participants, and a *spatial formation*), and an *utterance vector* of the current spoken utterance. The total dimensionality of the input features was 1244.

Motion abstraction We use motion abstraction to characterize a set of stopping locations, motion trajectories, and spatial formations which can be used to describe the motion of the customer or shopkeeper as a combination of discrete state variables rather than raw position or velocity data.

To begin the analysis, we segmented all trajectories in the training data into moving and stopped trajectories, based on a velocity thresholding technique presented in Guéguen (2001). We spatially clustered these trajectory segments to identify a discrete set of typical **stopping locations** and **motion trajectories** for each role (customer and shopkeeper).

For stopping locations, we used k-means clustering, identifying five stopping locations for the customer (i.e. the locations of the 3 cameras, the middle, and the door) and five for the shopkeeper (i.e. the locations of the 3 cameras, the middle, and the service counter).

For moving trajectories we used k-medoid clustering based on spatiotemporal matching using dynamic time warping.

We created rules for identifying a predetermined set of common **spatial formations** based on the distance between the interactants and their locations. The rules for spatial formations are similar to three existing HRI proxemics models: (1) *present object* (Yamaoka et al. 2008): both interactants were at stopping locations corresponding to the same camera, (2) *face-to-face* (Hall 1966): both interactants are within 1.5 m of each other but not at a camera, and (3) *waiting* (Kitade et al. 2013): if the shopkeeper was at the service counter and the customer was not.

In addition, we also identified the current spatial target for a particular spatial formation. The *formation target* for "present object" can be either Sony, Nikon, or Canon, whereas the *formation target* for the spatial formation "face-to-face" and "waiting" is 'none'.

Utterance vectorization We performed utterance vectorization of the customer and shopkeeper using common text-processing techniques. Specifically, we removed stop words, applied a Porter stemmer, enumerated n-grams up to 3, and performed Latent Semantic Analysis (Landauer et al. 1998) to reduce the dimensionality to 1000. To emphasize important keywords, we also used the AlchemyAPI cloud-based service[2] to automatically extract keywords from each utterance and represented the keywords separately in the vector (200 dimensions). By using this procedure, we were able to take any input utterance and represent it using a 1200-dimensional vector. Vectorization of customer and shopkeeper utterances were performed independently.

Defining robot actions

In our system, each observed shopkeeper action must correspond to a discrete robot action. A robot action consists of an utterance (represented by an ID number) with a corresponding target formation.

Shopkeeper utterance To reproduce shopkeeper speech with a robot, it is necessary to define a set of discrete utterance actions. Common utterances are frequently repeated in the training data (for example, variants of "How may I help you?" occur 188 times), but these instances often include slight differences due to speech recognition errors or individual variation. We used bottom-up hierarchical clustering based on lexical cosine similarity to group these similar utterances into 761 clusters corresponding to discrete robot speech actions.

From each shopkeeper utterance cluster, one utterance was selected for use in behavior generation. For each utterance, we compute the cosine similarity of its term frequency vector with every other utterance in the same cluster, and we sum these similarity values. The utterance with the highest similarity sum is chosen as the typical utterance to be used to generate robot speech. Notice the typical utterance can also be "none", which means that the robot does not output an utterance.

Target formation We use the same abstraction rule described earlier to represent a target spatial formation for the robot (i.e. *present product*, *face-to-*face, *waiting*, or *none*). This allows the robot to precisely calculate its target position and facing direction defined by the specefic HRI model, in accordance with its estimation of the customer's destination.

If the predicted target formation is different from the robot's current formation, the robot moves to attain the new target formation. Specifically, if the predicted formation is *face-to-face*, the robot approaches the customer; if the predicted formation is *waiting*, it returns to the service counter; if the predicted formation is *present-object*, the robot

[2] http://www.alchemyapi.com.

approaches the target object; and if the predicted formation is *none*, the robot stays where it is.

References

Admoni, H., & Scassellati, B. (2014). Data-driven model of nonverbal behavior for socially assistive human–robot interactions. In *Proceedings of the 16th international conference on multimodal interaction* (pp. 196–199), ACM.

Awais, M., & Henrich, D. (2012). Proactive premature intention estimation for intuitive human–robot collaboration. In *2012 IEEE/RSJ international conference on intelligent robots and systems* (pp. 4098–4103), IEEE.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Bauer, A., Klasing, K., Lidoris, G., Mühlbauer, Q., Rohrmüller, F., Sosnowski, S., et al. (2009). The autonomous city explorer: Towards natural human–robot interaction in urban environments. *International Journal of Social Robotics*, *1*(2), 127–140.

Breazeal, C., DePalma, N., Orkin, J., Chernova, S., & Jung, M. (2013). Crowdsourcing human–robot interaction: new methods and system evaluation in a public environment. *Journal of Human–Robot Interaction*, *2*(1), 82–111.

Brscic, D., Kanda, T., Ikeda, T., & Miyashita, T. (2013). Person tracking in large public spaces using 3-D range sensors. *IEEE Transactions on Human–Machine Systems*, *43*(6), 522–534. https://doi.org/10.1109/thms.2013.2283945.

Chao, C., & Thomaz, A. L. (2011). Timing in multimodal turn-taking interactions: Control and analysis using timed petri nets. *Journal of Human–Robot Interaction*, *1*(1), 1–16.

Cheng, J., Dong, L., & Lapata, M. (2016). Long short-term memory-networks for machine reading. arXiv preprint arXiv:1601.06733.

Chernova, S., DePalma, N., Morant, E., & Breazeal, C. (2011). Crowdsourcing human–robot interaction: Application from virtual to physical worlds. In *RO-MAN, 2011 IEEE, July 31 2011–Aug. 3 2011* (pp. 21–26). https://doi.org/10.1109/roman.2011.6005284.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, *23*(2), 283.

Duncan, S. (1974). On the structure of speaker–auditor interaction during speaking turns. *Language in Society*, *3*(02), 161–180.

Fox, D., Burgard, W., & Thrun, S. (1997). The dynamic window approach to collision avoidance. *IEEE Robotics & Automation Magazine*, *4*(1), 23–33.

Glas, D. F., Brščič, D., Miyashita, T., & Hagita, N. (2015). SNAPCAT-3D: Calibrating networks of 3D range sensors for pedestrian tracking. In *2015 IEEE international conference on robotics and automation (ICRA)* (pp. 712–719), IEEE.

Gu, E., & Badler, N. I. (2006). Visual attention and eye gaze during multiparty conversations with distractions. In *International workshop on intelligent virtual agents* (pp. 193–204), Springer.

Guéguen, L. (2001). Segmentation by maximal predictive partitioning according to composition biases. In O. Gascuel, & M.-F. Sagot (Eds.), *Computational biology. Lecture Notes in Computer Science* (Vol. 2066, pp. 32–44). Berlin: Springer.

Hall, E. T. (1966). *The hidden dimension*. London: The Bodley Head Ltd.

Hayashi, K., Sakamoto, D., Kanda, T., Shiomi, M., Koizumi, S., Ishiguro, H., et al. (2007). Humanoid robots as a passive-social medium—A field experiment at a train station. In *2007 2nd ACM/IEEE international conference on human–robot interaction (HRI), 9–11 March 2007* (pp. 137–144).

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., et al. (2015). Teaching machines to read and comprehend. In *Advances in neural information processing systems* (pp. 1693–1701).

Huang, C.-M., Cakmak, M., & Mutlu, B. (2015). Adaptive coordination strategies for human–robot handovers. In *Proceedings of robotics: Science and systems*.

Hulme, C., Maughan, S., & Brown, G. D. (1991). Memory for familiar and unfamiliar words: Evidence for a long-term memory contribution to short-term memory span. *Journal of Memory and Language*, *30*(6), 685–701.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.

Jayawardena, C., Kuo, I.-H., Broadbent, E., & MacDonald, B. A. (2016). Socially assistive robot healthbot: Design, implementation, and field trials. *IEEE Systems Journal*, *10*(3), 1056–1067.

Kawai, H., Toda, T., Ni, J., Tsuzaki, M., & Tokuda, K. (2004). XIMERA: A new TTS from ATR based on corpus-based technologies. In *Fifth ISCA workshop on speech synthesis*.

Keizer, S., Foster, M. E., Wang, Z., & Lemon, O. (2014). Machine learning for social multiparty human–robot interaction. *ACM Transactions on Intelligent Systems and Technology*, *4*(3), 1–32. https://doi.org/10.1145/2600021.

Kitade, T., Satake, S., Kanda, T., & Imai, M. (2013). Understanding suitable locations for waiting. In *Proceedings of the 8th ACM/IEEE international conference on Human–robot interaction* (pp. 57–64), IEEE Press.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284.

Liu, P., Glas, D. F., Kanda, T., & Ishiguro, H. (2016). Data-driven HRI: Learning social behaviors by example from human–human interaction. *IEEE Transactions on Robotics*, *32*(4), 988–1008. https://doi.org/10.1109/tro.2016.2588880.

Michalowski, M. P., Sabanovic, S., & Simmons, R. (2006). A spatial model of engagement for a social robot. In *9th IEEE international workshop on advanced motion control, 2006* (pp. 762–767). michalowski06: IEEE.

Michaud, F., & Matarić, M. J. (1998). Learning from history for behavior-based mobile robots in non-stationary conditions. *Machine Learning*, *31*(1–3), 141–167.

Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *Interspeech* (Vol. 2, p. 3)

Mohammad, Y., & Nishdia, T. (2012). Self-initiated imitation learning. Discovering what to imitate. In *2012 12th International conference on control, automation and systems (ICCAS), 2012* (pp. 726–732), IEEE.

Mutlu, B., Shiwa, T., Kanda, T., Ishiguro, H., & Hagita, N. (2009). *Footing in human–robot conversations: How robots might shape participant roles using gaze cues.* Paper presented at the Proceedings of the 4th ACM/IEEE international conference on Human robot interaction, La Jolla, California, USA.

Nickel, K., & Stiefelhagen, R. (2007). Visual recognition of pointing gestures for human–robot interaction. *Image and Vision Computing*, *25*(12), 1875–1884.

Orkin, J., & Roy, D. (2007). The restaurant game: Learning social behavior and language from thousands of players online. *Journal of Game Development*, *3*(1), 39–60.

Orkin, J., & Roy, D. (2009). Automatic learning and generation of social behavior from collective human gameplay. In *Proceedings of the 8th international conference on autonomous agents and multiagent systems-volume 1* (pp. 385–392). International Foundation for Autonomous Agents and Multiagent Systems

Pandey, A. K., Ali, M., & Alami, R. (2013). Towards a task-aware proactive sociable robot based on multi-state perspective-taking. *International Journal of Social Robotics*, *5*(2), 215–236.

Raffel, C., & Ellis, D. P. (2015). Feed-forward networks with attention can solve some long-term memory problems. arXiv preprint arXiv:1512.08756.

Raux, A., & Eskenazi, M. (2008). Optimizing endpointing thresholds using dialogue features in a spoken dialogue system. In *Proceedings of the 9th SIGdial workshop on discourse and dialogue* (pp. 1–10). Association for Computational Linguistics

Rich, C., Ponsler, B., Holroyd, A., & Sidner, C. L. (2010). Recognizing engagement in human–robot interaction. In *2010 5th ACM/IEEE international conference on human–robot interaction (HRI)* (pp. 375–382), IEEE

Robins, B., Dautenhahn, K., & Dickerson, P. (2009). From isolation to communication: a case study evaluation of robot assisted play for children with autism with a minimally expressive humanoid robot. In *Second international conferences on advances in computer–human interactions, 2009. ACHI'09* (pp. 205–211), IEEE.

Rozo, L., Silvério, J., Calinon, S., & Caldwell, D. G. (2016). Learning controllers for reactive and proactive behaviors in human–robot collaboration. *Frontiers in Robotics and AI*, *3*, 30.

Satake, S., Hayashi, K., Nakatani, K., & Kanda, T. (2015). Field trial of an information-providing robot in a shopping mall. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 1832–1839), IEEE.

Satake, S., Kanda, T., Glas, D. F., Imai, M., Ishiguro, H., & Hagita, N. (2009). How to approach humans? Strategies for social robots to initiate interaction. In *Proceedings of the 4th ACM/IEEE international conference on human robot interaction, La Jolla, California, USA* (pp. 109–116), ACM. https://doi.org/10.1145/1514095.1514117.

Schmid, A. J., Weede, O., & Worn, H. (2007). Proactive robot task selection given a human intention estimate. In *RO-MAN 2007—The 16th IEEE international symposium on robot and human interactive communication, 26–29 Aug. 2007* (pp. 726–731). https://doi.org/10.1109/roman.2007.4415181.

Schrempf, O. C., Hanebeck, U. D., Schmid, A. J., & Worn, H. (2005). A novel approach to proactive human–robot cooperation. In *ROMAN 2005. IEEE international workshop on robot and human interactive communication, 2005.* (pp. 555–560), IEEE

Shi, C., Kanda, T., Shimada, M., Yamaoka, F., Ishiguro, H., & Hagita, N. (2010). Easy development of communicative behaviors in social robots. In *2010 IEEE/RSJ international conference on intelligent robots and systems (IROS), 18–22 Oct. 2010* (pp. 5302–5309). https://doi.org/10.1109/iros.2010.5650128.

Shi, C., Shimada, M., Kanda, T., Ishiguro, H., & Hagita, N. (2011). Spatial formation model for initiating conversation. In *Proceedings of robotics: Science and systems VII*.

Shiomi, M., Kanda, T., Glas, D. F., Satake, S., Ishiguro, H., & Hagita, N. (2009). Field trial of networked social robots in a shopping mall. In *IEEE/RSJ international conference on intelligent robots and systems, 2009. IROS 2009. St. Louis, MO, USA, 10–15 Oct. 2009* (pp. 2846–2853). shiomi09: IEEE Press. https://doi.org/10.1109/iros.2009.5354242.

Sugiyama, O., Kanda, T., Imai, M., Ishiguro, H., & Hagita, N. (2007). Natural deictic communication with humanoid robots. In *2007 IEEE/RSJ international conference on intelligent robots and systems* (pp. 1441–1448), IEEE.

Sukhbaatar, S., Weston, J., & Fergus, R. (2015). End-to-end memory networks. In *Advances in neural information processing systems* (pp. 2440–2448).

Thomaz, A. L., & Chao, C. (2011). Turn-taking based on information flow for fluent human–robot interaction. *AI Magazine*, *32*(4), 53–63.

Toris, R., Kent, D., & Chernova, S. (2014). The robot management system: A framework for conducting human–robot interaction studies through crowdsourcing. *Journal of Human–Robot Interaction*, *3*(2), 25–49.

Triebel, R., Arras, K., Alami, R., Beyer, L., Breuers, S., Chatila, R., et al. (2016). Spencer: A socially aware service robot for passenger guidance and help in busy airports. In *Field and service robotics* (pp. 607–622), Springer.

Viejo, G., Khamassi, M., Brovelli, A., & Girard, B. (2015). Modeling choice and reaction time during arbitrary visuomotor learning through the coordination of adaptive working memory and reinforcement learning. *Frontiers in Behavioral Neuroscience*, *9*, 225.

Yamaoka, F., Kanda, T., Ishiguro, H., & Hagita, N. (2008). How close? A model of proximity control for information-presenting robots. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction, Amsterdam, The Netherlands* (pp. 137–144), ACM. https://doi.org/10.1145/1349822.1349841.

Young, J. E., Igarashi, T., Sharlin, E., Sakamoto, D., & Allen, J. (2014). Design and evaluation techniques for authoring interactive and stylistic behaviors. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *3*(4), 23.

Young, J. E., Sharlin, E., & Igarashi, T. (2013). Teaching robots style: Designing and evaluating style-by-demonstration for interactive robotic locomotion. *Human–Computer Interaction*, *28*(5), 379–416.

Phoebe Liu received her Ph.D. (2017) and M.Eng. (2013) in engineering from Osaka University, Osaka, Japan. She has been an internship researcher (2011–2017) and a researcher (2017–) at the Hiroshi Ishiguro Laboratories (HIL) at the Advanced Telecommunications Research Institute International (ATR) in Kyoto, Japan. Research interests: social human–robot interaction, machine learning.

Dylan F. Glas received his Ph.D. in Robotics from Osaka University (2013) and M.Eng. (2000) and S.B. (1997) degrees in Aerospace Engineering and Earth, Atmospheric, and Planetary Sciences from MIT. He is currently a Senior Researcher at Hiroshi Ishiguro Laboratories at ATR and a Guest Associate Professor at Osaka University. Research interests: autonomous human–robot interaction.

Takayuki Kanda received his B.Eng. (1998), M.Eng. (2000), and Ph.D. (2003) degrees in computer science from Kyoto University, Kyoto, Japan. He was an Intern Researcher at ATR Media Information Science Laboratories (2000–2003) and is currently a Senior Researcher at ATR Intelligent Robotics and Communication Laboratories. Research interests: intelligent robotics, human–robot interaction.

Hiroshi Ishiguro received a D.Eng. in systems engineering from Osaka University (1991). He is currently professor of the Department of Systems Innovation in the Graduate School of Engineering Science at Osaka University (2009–) and distinguished professor of Osaka University (2013–). He is group leader (2002–) of Hiroshi Ishiguro Laboratories at ATR and an ATR fellow. Research interests: distributed sensor systems, interactive robotics, android science.