Nonverbal Behavior Modeling for Socially Assistive Robots

Henny Admoni and Brian Scassellati
Department of Computer Science
Yale University
New Haven, Connecticut 06511 USA

Abstract

The field of socially assistive robotics (SAR) aims to build robots that help people through social interaction. Human social interaction involves complex systems of behavior, and modeling these systems is one goal of SAR. Nonverbal behaviors, such as eye gaze and gesture, are particularly amenable to modeling through machine learning because the effects of the system—the nonverbal behaviors themselves—are inherently observable. Uncovering the underlying model that defines those behaviors would allow socially assistive robots to become better interaction partners. Our research investigates how people use nonverbal behaviors in tutoring applications. We use data from human-human interactions to build a model of nonverbal behaviors using supervised machine learning. This model can both *predict* the context of observed behaviors and *generate* appropriate nonverbal behaviors.

Introduction

Socially assistive robotics (SAR) is a subfield of robotics that aims to design, construct, and evaluate robots that help people through social interactions (Feil-Seifer and Matarić 2005). Applications of SAR include educational tutoring (Kanda et al. 2004), eldercare (Wada and Shibata 2007), and therapy (Scassellati, Admoni, and Matarić 2012).

Efficient, intuitive human-robot communication is critical to SAR. People perform much of their communication nonverbally, using behaviors like eye gaze and gesture to convey mental state, to reinforce verbal communication, or to augment what is being said (Argyle 1972). Though these nonverbal behaviors are generally natural and effortless for people, they must be explicitly designed for robots. As SAR applications become more common, there is a growing need for robots to be able to use nonverbal communication.

Because people are so attuned to nonverbal communication, robot behavior must follow human expectations. If robots generate social behavior that is outside of the established communication norms, people will be confused or reject the robot interaction outright. Therefore, any approach to designing social behaviors for robots must be informed by actual human behavior.

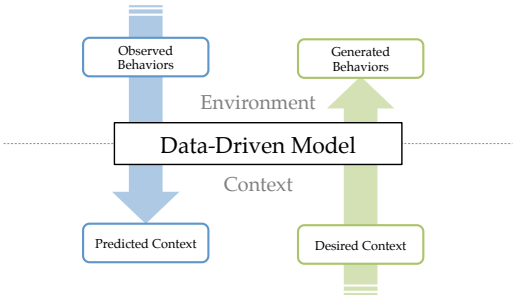

Figure 1: A model for nonverbal communication in SAR applications must be able to both predict the context of observed behaviors and generate appropriate behaviors for a desired context.

Fortunately, nonverbal communication is a particularly good candidate for supervised machine learning. Because nonverbal behaviors are inherently observable, these observations can be used to train a model of communication that can inform a robot's behaviors.

Our research focuses on improving human-robot interaction in SAR applications by modeling the complex dynamics of human communication. We do so by building models of human-human nonverbal communication and implementing these models to generate socially appropriate and communicative behavior for socially assistive robots. These data-driven models allow us to design robots that match people's existing nonverbal communication use.

Some researchers have begun to address this need for data-driven robot behavior models. For example, researchers have modeled conversational gaze aversions (Andrist et al. 2014) and gesture during narration (Huang and Mutlu 2013) based on analysis of human-human pairs. While these studies show promising advances in data-driven models of robot behavior, none of them deals directly with socially assistive applications, which require monitoring of—and feedback from—the interaction partner.

Modeling Human Interactions

Unlike previous work, our model is *bidirectional*, enabling both the *prediction* of a user's intent given observed nonverbal behaviors, and the *generation* of appropriate nonverbal

Category (*label*)	Vocabulary
Context (C)	Fact, spatial reference, demonstration, floor maintenance, interactive, other
Gaze (A)	Partner, referent, own gesture, other
Gesture (E)	Iconic, metaphoric, deictic, beat, demonstration, functional, other
Gesture style (S)	Sweep, point, hold, move
Affiliate (F)	Map, box, player token, partner face, partner hands, partner cards, etc.

Table 1: The coding vocabulary used to extract data from the human-human interaction video.

Figure 2: A frame from a human-human teaching interaction used to train the model.

communication behaviors for a robot (Figure 1).

There are three steps in the process of designing a data-driven generative behavior model: 1) collect data on nonverbal behaviors during human-human interactions, 2) train a predictive computational model with the human-human interaction data, and 3) develop a generative model for robot behaviors driven by the computational model from step 2.

To collect data about human interactions, we analyzed typical teaching interactions between pairs of individuals. One of the participants (the teacher) was asked to teach a second participant (the student) how to play a graph-building board game called TransAmerica. This game was chosen because the spatial nature of gameplay encouraged many nonverbal behaviors such as pointing.

We video and audio-record the teaching interaction (Figure 2), which was used as data for our model. Teaching interactions lasted approximately five minutes per dyad. To extract data from the video recordings, we manually annotated the nonverbal behavioral features identified in Table 1.

Each annotation can be described by a tuple (a, e, s, f_a, f_e) where $a \in A$ is gaze behavior, $e \in E$ is gesture behavior, $s \in S$ is gesture style (which indicates how the gesture was performed), and $f_a, f_e \in F$ are real-world objects or locations that gaze and gesture were directed toward, respectively. Each annotation has at least one non-null value in the tuple, though not all values need be non-null. Annotations are labeled with a context $c \in C$ that defines the subject or purpose of the communication.

With this representation, we can conceptualize the annotations as labeled points in high-dimensional space. New observations of nonverbal behavior can be classified using a k-nearest neighbor algorithm. To classify a new sample, the algorithm finds the k closest training samples and assigns the new observation a context based on a majority vote of c for those samples. This model allows our system to predict the context of new observations of nonverbal behaviors.

Generating Robot Behavior

To generate robot behavior, the system first identifies the desired context of the communication. Currently this is prespecified by labeling each robot utterance with a context and, optionally, an affiliate. For example, a segment of robot speech that refers deictically to the map, such as "you can build on any of these locations," is labeled with the spatial reference context and the map affiliate.

To select appropriate behaviors given the context, the system finds the largest cluster of examples of that context in the high-dimensional feature space, and selects the behaviors based on the tuple values in that cluster. In other words, the system finds the behaviors that were most often observed in that context. To generate more behavior variability, and to account for contexts in which there is more than one "right" behavior, the system can identify all of the tuples labeled with the context, and select behaviors by weighting the probability of selecting a set of tuple values by how many examples there are of those values labeled with the desired context.

In the teaching example, a spatial reference context was most often found with $a =$ referent, $e =$ deictic, $s =$ point, $f_a =$ map, and $f_e =$ map. Therefore, when performing the speech utterance labeled with the spatial reference context, the robot would make a deictic pointing gesture toward the map, while looking at the map.

Future Work

Real-time learning and adaptation remains a challenge of socially assistive robotics. People's preferences and knowledge change over time, and good SAR systems should be capable of adapting in real time based on continuously collected training samples. The current model is capable of such real-time adaptation given the appropriate training samples. However, classifying these samples online can be difficult. While there have been significant improvements in body posture recognition, gaze tracking, and natural language processing (for context recognition), real-time sensing is not yet reliable enough for this application.

Acknowledgments

This work is supported by NSF grants 1139078 and 1117801.

References

Andrist, S.; Tan, X. Z.; Gleicher, M.; and Mutlu, B. 2014. Conversational gaze aversion for humanlike robots. In *Pro-*

ceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI '14). ACM.

Argyle, M. 1972. Non-verbal communication in human social interaction. In Hinde, R. A., ed., *Non-verbal communication*. Oxford, England: Cambirdge University Press.

Feil-Seifer, D., and Matarić, M. J. 2005. Defining socially assistive robotics. In *Proceedings of the 9th International IEEE Conference on Rehabilitation Robotics*.

Huang, C.-M., and Mutlu, B. 2013. Modeling and Evaluating Narrative Gestures for Humanlike Robots. In *Proceedings of Robotics: Science and Systems*.

Kanda, T.; Hirano, T.; Eaton, D.; and Ishiguro, H. 2004. Interactive robots as social partners and peer tutors for children: A field trial. *Human-Computer Interaction* 19:61–84.

Scassellati, B.; Admoni, H.; and Matarić, M. 2012. Robots for use in autism research. *Annual Review of Biomedical Engineering* 14:275–294.

Wada, K., and Shibata, T. 2007. Living with seal robots—its sociophsychological and physiological influences on the elderly at a care house. *IEEE Transactions on Robotics* 23(5):972–980.