Data-Driven Model of Nonverbal Behavior for Socially Assistive Human-Robot Interactions

Henny Admoni
Department of Computer Science
Yale University
New Haven, CT 06520 USA
henny@cs.yale.edu

Brian Scassellati
Department of Computer Science
Yale University
New Haven, CT 06520 USA
scaz@cs.yale.edu

ABSTRACT

Socially assistive robotics (SAR) aims to develop robots that help people through interactions that are inherently social, such as tutoring and coaching. For these interactions to be effective, socially assistive robots must be able to recognize and use nonverbal social cues like eye gaze and gesture. In this paper, we present a preliminary model for nonverbal robot behavior in a tutoring application. Using empirical data from teachers and students in human-human tutoring interactions, the model can be both *predictive* (recognizing the context of new nonverbal behaviors) and *generative* (creating new robot nonverbal behaviors based on a desired context) using the same underlying data representation.

Categories and Subject Descriptors

I.2.9 [**Artificial Intelligence**]: Robotics

Keywords

Human-robot interaction; nonverbal behavior; tutoring

1. INTRODUCTION

Socially assistive robotics (SAR) focuses on building robots that help people through interactions that are inherently social [5]. Application areas for SAR include tutoring [8, 9], autism therapy [13], and elder care [16]. Social robots augment traditional human-human interactions in these areas by providing additional interactions that are impractical, time-consuming, or impossible to achieve with a person.

For example, a social robot can act as a peer tutor, helping students practice skills or solidify knowledge through one-on-one interactions outside of the classroom. By presenting itself as a peer, the robot can encourage students to practice previously-learned knowledge by re-teaching it to the robot. In this way, the robot provides educational support beyond what a classroom teacher has time for, and with potentially more consistent quality than a human peer.

ICMI '14 November 12–16 2014, Istanbul, Turkey

http://dx.doi.org/10.1145/2663204.2663263.

Figure 1: Screenshots from human-human teaching interaction videos. The student (top) displays gaze to the referent, while the teacher (bottom) displays gaze to the partner and a deictic gesture to the map.

For social robots to be effective communicators, they must understand the *context* of their human partner's communication, that is, the communicative goal or perspective. In the tutoring robot example, for instance, the robot must be able to recognize whether its partner is referring to a location in the environment, asking a question, or explaining some knowledge. Similarly, social robots must be able to convey the context of their own communication effectively.

The cues to understanding such context can come from speech, but often come from nonverbal behaviors like eye gaze [3] and gesture [10]. Gestures, for instance, reflect ideas that are not necessarily conveyed in speech [6], and teachers frequently use gestures to ground their spoken utterances to the objects of instruction [1, 12]. Eye gaze is critical for joint attention—simultaneous attention toward a particular object or location—which is fundamental for learning [15]. Therefore, the effectiveness of the tutoring robot, or any socially assistive robot, depends on its ability to recognize and utilize the nonverbal context clues that people use naturally.

In this work, we take a *data-driven* approach, using empirical data from human-human interactions to build a model of nonverbal robot behaviors. By training on previously-observed human behavior, we take advantage of the frequency and ease with which people use nonverbal behaviors to design more communicative robot behaviors.

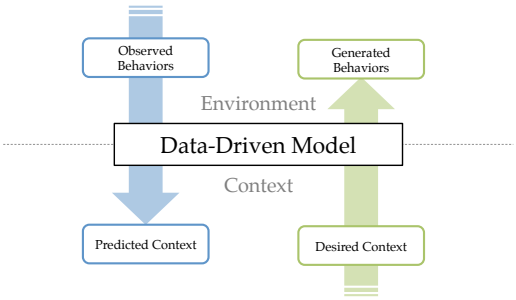


Figure 2: The model performs both context prediction (blue) and behavior generation (green).

Other work uses a similar data-driven approach for nonverbal behavior modeling. Researchers have generated robot behavior, such as gaze aversions [2] and narrative gestures [7], by analyzing videos of people conversing or telling stories. For virtual agents, empirical observation has driven gesture formation for iconic gestures [4] and narrative performance [11, 14].

However, much of this previous work focuses exclusively on the speaker's behaviors. In contrast, our work considers the behaviors of *both* interaction partners simultaneously. Tutoring is an activity with bi-directional communication—the teacher makes a statement, the student asks a question, the teacher replies, the student confirms—and peoples' nonverbal behavior is influenced by the behavior of their partner. For instance, joint reference is a common social behavior that involves one person deictically referring to an object or location, then repeatedly glancing between that referent and the partner's face to confirm that their partner understands the reference. With a view of both partners' behaviors, joint reference can emerge naturally from our model.

Our work also hinges on the idea that a model for nonverbal behavior should be simultaneously *predictive* and *generative*. In other words, the model should be able to both predict (or classify) the context of a newly observed set of nonverbal behaviors, and generate a set of nonverbal behaviors given a context of communication (Figure 2), without needing to collect and train on different sets of data. Some other work has this capability (such as [4]), but we elevate this to a central requirement for our system.

In this paper, we introduce the context and features that comprise our model and describe our preliminary data collection of real-world human-human teaching interactions. We then describe our model in terms of these features, detailing how it can both predict new contexts and generate new behaviors. We evaluate the model and show that it is effective at both of these tasks. We conclude with ideas for extensions of this model.

2. HUMAN-HUMAN INTERACTION

To create a model of nonverbal behavior, we first collected examples of nonverbal behavior during tutoring (Figure 1). We recruited two pairs of participants (mean age 22), randomly assigning one as teacher and the other as student, and recorded their interaction as the teacher taught the student how to play a board game called TransAmerica.

In TransAmerica, players must place game pieces representing railroad tracks along a grid overlaid on a map of the United States. Players score points for successfully building a track network that connects the cities specified in their randomly-selected hand of cards. We chose this game specifically because teaching the game involves spatial references, which encourage deictic gestures and demonstrations in addition to statements of facts and rules.

Neither student nor teacher had played the board game previously. Before the recorded interaction, the teacher was given a lesson on the game from an experimenter for approximately five minutes. The teacher was also provided with a rule sheet that described all of the rules of the game.

We audio- and video- recorded both teacher and student during this interaction, which lasted approximately five minutes per dyad. We then manually coded these recordings for five *predictors*: the *teacher's gaze*, the *teacher's gestures*, the *teacher's deictic references*, the *student's gaze*, and the *student's gestures* (Table 1). The student's deictic reference was infrequent, so we did not code for that predictor. We also coded the *context* of each utterance.

Values for gaze follow previous work [7], and represent possible gaze locations: to the *partner*, to the *referent* of current speech (regardless of who is speaking), to one's *own gesture*, or to some *other* location in the environment.

Values for gesture include those from established categorizations as well as additional values specific to physically-based teaching tasks. *Iconic*, *metaphoric*, *deictic*, and *beat* gestures are defined as in the literature [10]. *Demonstrations* involve physical movements that mimic the topic of speech. *Functional* movements are not intended for communication, but are used to accomplish game-related tasks such as dealing cards. Actions outside of these categories, such as brushing hair behind an ear, were categorized as *other*.

The deixis category encodes gesture types—*pointing* to a single target, *sweeping* over a range of targets, and *holding*—as well as gesture locations—the game *map*, *cards*, *playing pieces*, and *box*. Though every deixis value must have an associated gesture, not every gesture must have a deixis value. Deixis values can occur with any gesture, especially demonstrations and functional gestures.

The nine contexts each represent a particular kind of communication, and contexts are determined based on both speech and nonverbal behaviors. Contexts are mutually exclusive, though two sets of identical nonverbal behaviors may be classified as different contexts, for instance based on different speech during those behaviors.

The *rules* context indicates communication about the rules of the game. *Fact* contexts involve communication about facts that don't include game rules, such as "The name of the game is TransAmerica." An *expository* context indicates communication that elaborates on previous statements without providing new rules or facts. *Question* and *reply* contexts involve asking questions or providing direct answers, respectively. *Deixis* indicates communication that refers to physical locations or nearby objects. *Confirmation* contexts involve confirmation-seeking questions or statements such as "do you understand?" *Backchannel* contexts are utterances that indicate a listener's attention. *Filler* are non-meaningful communications that stand in for silence, often at the beginning of a new phrase.

We developed the list of contexts before examining the human-human interaction data, so that we would not be swayed by individual preferences for certain contexts. Interestingly, we did not note a single instance of confirmation context in the interactions we annotated, despite their ex-

Name	Values
Gaze (A)	partner, referent, own gesture, other
Gesture (E)	iconic, metaphoric, deictic, beat, demonstration, functional, other
Deixis (D)	point {map, own cards, partner cards, box}, sweep {map, box}, hold {cards, game piece, box}
Context (C)	backchannel, deixis, expository, fact, filler, question, reply, rules

Table 1: Model parameters and their values.

pected appearance in a teaching task. It is possible that a more experienced teacher might employ confirmation seeking behaviors, even though our current participants did not.

3. NONVERBAL BEHAVIOR MODEL

A model of nonverbal behavior should be able to classify the context given new observations of nonverbal behavior, as well as generate appropriate behaviors to suit a desired context (Figure 2).

We discretized the human-human interaction recordings into one-second segments. Each segment provides one observation $o \in O$, which is described by a tuple of predictors $o = \{a_T, e_T, d_T, a_S, e_S\}$ where $a_T, a_S \in A$ are the type of eye gaze exhibited by the teacher and student, respectively, $e_T, e_S \in E$ are the types of gestures exhibited by the teacher and student, respectively, and $d_T \in D$ is the deictic referent of the teacher's gesture in that segment. We chose one-second segments after observing the interactions, though the level of data granularity is flexible and may be adjusted for different applications.

Sometimes it is useful to take history into account, as well. An observation with history,

$$o_h = \{a_{T_t}, e_{T_t}, d_{T_t}, a_{S_t}, e_{S_t}, a_{T_{t-1}}, e_{T_{t-1}}, d_{T_{t-1}}, a_{S_{t-1}}, e_{S_{t-1}}\}$$

is defined by predictor values at current time t and predictor values from the previous time step $t-1$, if available. The set of observations including history is O_h.

Using this formulation, we can represent each observation as a point in high-dimensional space.

3.1 Predicting Context

Given a set of observations of nonverbal behavior, our system can predict the context of the communication. To do so, observations from the human-human interactions were used to train a prediction algorithm using k-nearest neighbors. In this algorithm, predictors are attributes and context is the class label. We can denote this as $label\,(o_h) = c$ for observation o_h and context c. Note that *label* is not a function, since identical observations can have different contexts.

To classify the context of a new observation, the algorithm performs operation

$$nclosest : (O_h, o_{new}, k) \rightarrow K \tag{1}$$

which takes a set of observations O_h, a new observation o_{new}, and a positive integer k and returns a set $K = \{o_{h_1}, \ldots, o_{h_k}\}$ containing the k closest observations to o_{new}.

Because the predictor values are categorical, rather than continuous, our KNN algorithm uses the Hamming distance to identify nearest neighbors. For each existing observation $o_{old} = \{x_1, \ldots, x_n\}$, the algorithm calculates the distance D between o_{old} and the new observation $o_{new} = \{y_1, \ldots, y_n\}$,

$$D = \sum_{i=1}^{n} h(x_i, y_i), \quad h(a,b) = \begin{cases} 0 & \text{if } a = b \\ 1 & \text{if } a \neq b \end{cases} \tag{2}$$

Once it has evaluated the k nearest neighbors, the model assigns o_{new} a context based on a majority vote of the contexts of the observations in K. Ties are resolved by selecting randomly. Since there may be several different behaviors applicable in the same context, we extend the context assignment to o_{new} such that the probability of context assignment is proportional to the number of observations with that context in K. In other words, the probability of assigning o_{new} a context c is $p(c) = \frac{count(label(o_h)=c)}{k}$ for each $o_h \in K$, where $count(x)$ is a function that returns the number of instances of x in the data.

We empirically determined that $k = 2$ was the most accurate value for our data, though k may vary by application. Our model examines the two most similar examples of previous behavior to judge a new behavior's context.

3.2 Generating Behavior

Given a context, the model can also select appropriate nonverbal behaviors. It does so by finding the largest cluster of examples for the context, then selecting the nonverbal behaviors that are most common in that cluster.

Mathematically, given a desired context $c_{des} \in C$, the model searches over all observations $o \in O$ for

$$\{\{a_T, e_T, d_T, a_S, e_S\} \mid \begin{array}{l} count\,(label(o) = c_{des}) > \\ count\,(label(o) = c_i)\,, c_{des} \neq c_i\} \end{array} \tag{3}$$

Since this can yield multiple qualifying sets of behavior, the model can weight its behavior choice based on the frequency of observations containing that behavior for the desired context. This allows behavior variability in proportion to observed examples.

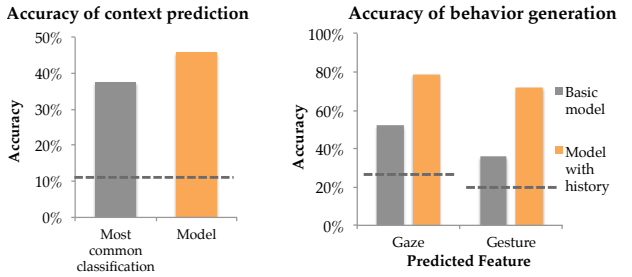
In effect, the model is replicating the most common behaviors it has observed for a given context. This follows the idea that people learn to communicate by mimicking observed behavior in given situations.

This behavior generation algorithm is amnesic because it does not account for history. To account for behavior from the previous time step, we use the history-aware representation of an observation, o_h. The process for generating new context (equation 3) remains the same, except that every step now uses o_h instead of o.

4. MODEL EVALUATION

Given a new observation containing the five predictors, how accurate is the model at identifying the correct context? We performed a 10-fold cross validation: combining all observations from both dyads, this validation segmented the data into 10 groups, trained the model on nine of those groups, and calculated the accuracy of context predictions using data from the remaining, untrained group (Figure 3a). On average, cross-validation accuracy was 45.9%. This value is significantly better than chance, which is 11.1% for nine classifications. It is also better than simply predicting the most common classification—rules—which only leads to an accuracy of 37.5% using the cross-validated model. Table 2 shows a confusion matrix for the cross-validated model.

Given a context, how well does the model generate gaze and gesture behaviors? To test this, we compared the rec-



(a) Accuracy of predicting context given behavior observations.

(b) Accuracy of generating new behaviors given context.

Figure 3: Results of model evaluation. The dashed line indicates chance.

10	13	23	1	11	2	2	21
4	14	16	0	6	4	0	15
6	13	59	3	29	8	3	28
1	0	4	0	1	2	0	1
1	13	36	2	32	1	3	22
5	6	7	2	8	26	0	5
1	2	4	0	1	1	2	4
7	19	47	3	25	6	3	72

Table 2: A confusion matrix for context prediction with the cross-validated model. Variable order is listed in Table 1.

orded human behavior for each observation in our data set against the most likely behavior generated by the algorithm for that observation's context (Figure 3b). When using the amnesiac generation method (that is, behavior generation that ignores any history), our system matches actual human gaze behavior 52.0% of the time, and human gesture behavior 36.0% of the time. This is an improvement over randomly selecting behavior values, which would yield 25.0% accuracy for gaze and 14.3% accuracy for gesture. Taking a single time-step of history into account significantly improves performance. The historically mindful generation method yields 78.8% accuracy for gaze behaviors and 72.0% accuracy for gestures.

5. DISCUSSION & FUTURE WORK

The next step in evaluating the model is to generate real robot behaviors and to measure how those behaviors are received in a human-robot interaction. We plan to use the generative portion of the model to create robot behaviors for a tutoring task similar to the one in this experiment. We will measure participants' acceptance and information recall when interacting with a robot using the current model versus a robot using a heuristic model or one that exhibits limited nonverbal behavior.

Given annotated data, our model is adaptable to new users and new tasks. However, data annotation remains a challenge. Manual annotations take time, and automatic annotations are not yet robust enough to correctly identify all of the features used by the model, particularly the context. However, automatic gaze and gesture detectors may ease some of the burden of manual annotation.

As a data-driven model, the effectiveness of the system depends on the quality of the data provided. This paper uses a small data corpus (two dyadic interactions), which we plan to increase for future evaluations. Even with this small corpus, however, the model can successfully predict and generate reasonable nonverbal behavior. While many nonverbal behaviors are consistent across people, atypical social behavior might necessitate a re-seeding of the model with new observations of that behavior.

We developed the model for a subset of the nonverbal behaviors that people use to communicate. Extending the model to include other features, such as head pose, might yield even greater accuracy and expressiveness.

6. ACKNOWLEDGMENTS

This work is supported by NSF grants 1117801 and 1139078.

7. REFERENCES

[1] M. W. Alibali and M. J. Nathan. Teachers' gestures as a means of scaffolding students' understanding: Evidence from an early algebra lesson. *Video Research in the Learning Sciences*, pages 349–365, 2007.

[2] S. Andrist, X. Z. Tan, M. Gleicher, and B. Mutlu. Conversational gaze aversion for humanlike robots. In *Proceedings of the 10th International Conference on Human-Robot Interaction (HRI '14)*. ACM, 2014.

[3] M. Argyle and M. Cook. *Gaze and Mutual Gaze*. Cambridge University Press, Oxford, England, 1976.

[4] K. Bergmann and S. Kopp. Modeling the production of coverbal iconic gestures by learning bayesian decision networks. *Applied Artificial Intelligence*, 24:530–551, 2010.

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

[6] S. Goldin-Meadow. The role of gesture in communication and thinking. *Trends in Cognitive Sciences*, 3(11):419 – 429, 1999.

[7] C.-M. Huang and B. Mutlu. Learning-based modeling of multimodal behaviors for humanlike robots. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, pages 57–64. ACM, 2014.

[8] T. Kanda, R. Sato, N. Saiwaki, and H. Ishiguro. Friendly social robot that understand human's friendly relationships. In *Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2215–2222, 2004.

[9] J. M. Kory, S. Jeong, and C. L. Breazeal. Robotic learning companions for early language development. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction (ICMI '13)*, pages 71–72, New York, NY, USA, 2013. ACM.

[10] D. McNeill. *Hand and Mind: What Gestures Reveal about Thought*. The University of Chicago Press, Chicago, 1992.

[11] M. Neff, M. Kipp, I. Albrecht, and H.-P. Seidel. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics*, 27(1):5:1–5:24, Mar. 2008.

[12] W.-M. Roth. Gestures: Their role in teaching and learning. *Review of Educational Research*, 71(3):365–392, 2001.

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

[14] M. Stone, D. DeCarlo, I. Oh, C. Rodriguez, A. Stere, A. Lees, and C. Bregler. Speaking with hands: Creating animated conversational characters from recordings of human performance. In *Proceedings of ACM SIGGRAPH*, pages 506–513, New York, NY, USA, 2004. ACM.

[15] M. Tomasello and M. J. Farrar. Joint attention and early language. *Child Development*, 57(6):1454–1463, 1986.

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