Monitoring and Guiding User Attention and Intention in Human-Robot Interaction

Aaron St. Clair, Ross Mead, and Maja J. Matarić, *Senior Member, IEEE*

***Abstract*—A robot interacting with humans and attempting to generate effective social interaction and intervention behaviors benefits greatly from being able to understand and predict the underlying intentions of actions in context. Related work on collaborative discourse suggests that intention can be described in terms of either goal-directed task completion or communicative behavior directed to other collaboration partners. This paper describes early work on a generalizable framework for estimating the attentional space of a human interaction partner, providing context for grounding action in terms of intentions, and using this model to perform contextualized robotic intervention and ambiguity resolution. We describe an experiment aimed at applying and validating the framework in a simple collaborative human-robot interaction scenario.**

I. Introduction

One of the key challenges for the development of autonomous robots capable of effective interaction with humans is accurately detecting and reacting to human activity in a variety of interaction contexts. Our work is motivated by, but not limited to, socially assistive robotics (SAR) [1], an area of human-robot interaction focusing on helping people through social interaction. Preliminary work in USC's Interaction Lab has shown that SAR has the potential to benefit a diverse variety of target populations, including stroke patients, children with autism, and the elderly [2, 3].

Most existing systems for modeling human activity in interactive settings are closely tied to a specific control system, robotic platform, and task model. Adapting these systems to new task environments and robot embodiments is inherently difficult and often requires re-implementing the system from the bottom up. By imbuing robots with some form of social intelligence, we aim to unify common interaction mechanisms across a wide variety of populations and platforms. Toward these ends, we are investigating a general framework for monitoring the attentional space of a user, contextualizing specific user actions according to intent, and using this as a basis for formulating practical robot actions for intervening and communicating robot intention.

Manuscript received March 31, 2010. This work was supported in part by a National Science Foundation (NSF) Graduate Research Fellowship, as well as NSF grants CNS-0709296, IIS-0803565, and IIS-0713697.

A. St. Clair, R. Mead, and M. J. Matarić are with the University of Southern California, Los Angeles, CA 90089 USA. (phone: 213-740-6245; fax: 213-821-5696; email: {astclair, rossmead, mataric}@usc.edu).

II. User Monitoring

Human activity in social contexts is extremely complex. Determining the intended effect of human action as observed from sensor data is difficult; it requires filtering and segmenting the input serial data streams, assigning one of a number of possible explanations for any given action, and contextualizing the intended action in terms of the interaction.

Drawing from collaborative discourse theory [4], we model the interaction of the user and the robot as a simplified collaboration, in which the robot maintains a model of the user, which is then used in both human action recognition and robot action planning. Breazeal *et al.* (2004) have used a related approach to structure collaborative learning, in which a robot learns a task model from guided human examples [5]. While our approach uses similar theoretical underpinnings, it is aimed at developing a general social communication framework that can be applied in many task environments, learned or otherwise.

A. Attention

User attention is modeled by constructing a probability distribution over salient world objects. Our current approach takes into account the user's position and head direction, extracted from camera and laser rangefinder sensors, as well as the relative saliency of world objects, to assign probability mass. For example, objects in the person's field of view are assigned relatively higher weights. The saliency map used can be specified *a priori* if the task domain is relatively static and well specified in advance, or it can be computed in real time using scene analysis [6].

B. Intention

Predicting human action and mapping it to underlying intention is a difficult problem since human activity is inherently complex. Neuroscience evidence suggests that humans accomplish this feat using mirror neurons to recognize actions while recruiting their own intentions for the recognized action and ascribing them to others [7]. Using a model of the current task and the estimated attentional space, we constrain the space of possible future actions and provide context to explain why a user might perform a particular action at a given time.

Our current focus is on deictic gestures—such as pointing, head orientation, and eye gaze—since they are well understood as a means of establishing joint attention, and are easily identified and physically grounded in terms of world objects. To compute possible target objects of a pointing gesture with respect to the user, we can utilize a Bayesian classifier to combine an error model of human pointing and the attentional distribution as a prior. We are investigating methods for recognizing attention and action stemming from more complex intentions, and distinguishing actions that are task-oriented, such as reaching, from those that are communication-oriented, such as pointing and other social gestures.

III. User Intervention

By monitoring user intentions and anticipating their effect on the success of an individual or collaborative task, a robot may determine that it is appropriate for it to intervene. Such an intervention may be deemed necessary to improve task performance or to prevent undesirable actions from being taken by the user. Directing user attention and intention must be done in as clear a way as possible to maintain a successful interaction between the robot and user. It is therefore crucial that potential ambiguity be minimized or resolved.

A. Intervention

In this preliminary work, the robot plans and executes an intervention strategy over possible proxemic and deictic actions. *Proxemics* here refers to the manipulation of robot position and orientation with respect to the human [8]. The robot must situate itself in the appropriate "social space" to maximize the effectiveness of subsequent communicative actions. Once the robot has positioned itself, it utilizes deictic gestures—such as pointing, head orientation, and eye gaze—to focus the attention of the user to a particular object or region, thus attempting to establish joint attention [9]. Intent is then communicated by exploiting the theory of perceived *affordances*, which suggests how an object may be interacted with [10]. This reliance on affordances constrains the interaction to simple tasks; however, in future work, we will investigate more complex representation and communication of intent [11].

B. Ambiguity Resolution

In the ideal case, the appropriate application of social distance and deictic gestures would result in a clear user interpretation of the task objective and, thus, a successful intervention; however, in the real world, such communication is often noisy and potentially ambiguous. To resolve such ambiguity, the robot engages in perspective-taking, considering the viewpoint of the human observer, as well as previous user activity. We utilize a naïve Bayes approach to estimate the clarity of a human's interpretation of potential robot actions over the attentional space. We then select a robot intervention strategy by applying gradient decent to find a global minimum with regard to ambiguity.

IV. Implementation

We are in the process of collecting human interpretation data based on interactions with a physical robot. An experiment will be conducted to evaluate the accuracy of automated user attention-intention monitoring. From this, we can produce a probabilistic model of error in user perception of robot deictic gestures. This model will then be validated in a collaborative task to demonstrate the efficacy of robot intervention and ambiguity resolution strategies with a human user.

A. Robot Platform

The system is being implemented on the Bandit III robot platform available in the Interaction Lab, shown in Fig. 1. Bandit is an upper-torso humanoid robot with 17 degrees of freedom: 7 in each arm (shoulder forward and backward, shoulder in and out, elbow tilt, elbow twist, wrist twist, grabber open and close; left and right arms), 2 in the head (pan and tilt), 2 in the lips (upper and lower), and 1 in the eyebrows. These degrees of freedom allow the robot to be highly expressive through individual and combined motions of the head, face, and arms. An extensive gesture and facial expression library has been developed to enhance the interactive experience. The robot is closer to human-scale than many other humanoid platforms; mounted atop a Pioneer P2 base, the entire robot stands one meter tall, making it an adequate choice for robot interaction. An overhead camera and on-board laser rangefinder facilitate human and robot pose tracking.

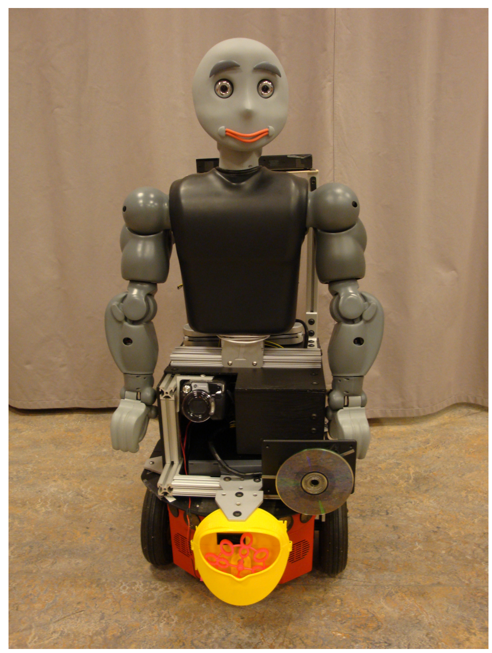

Fig. 1: The robot platform: Bandit upper-torso humanoid on a Pioneer base.

B. Experiment Design

We are investigating a concrete application of this framework within the realm of deictic gesture. The experimental design is a two-phased approach aimed at producing an empirical error model of both human gesture perception accuracy and robot gestural accuracy, and then applying these models using our attention, intention, and ambiguity resolution framework to allow a robot to engage in a simple collaborative task with a human partner.

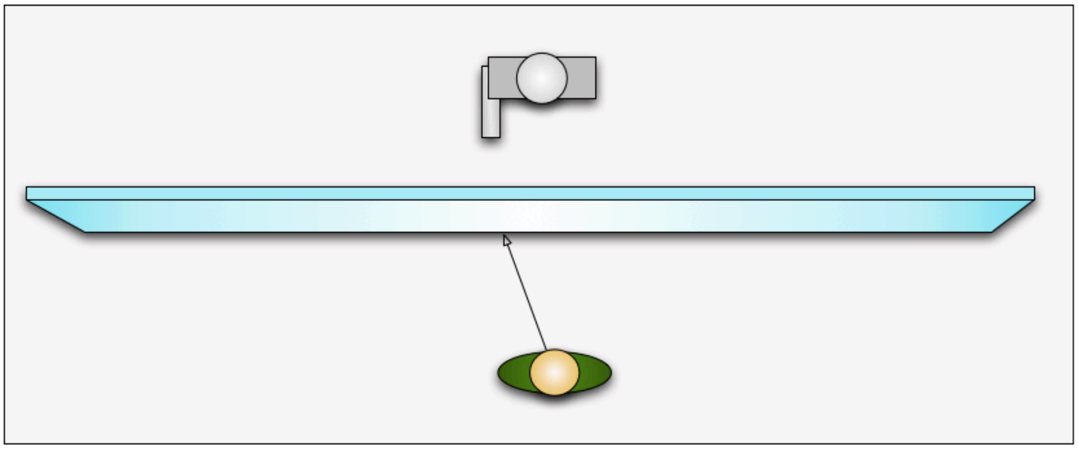

Fig. 2: The robot makes a deictic gesture on one side of the screen, and the user indicates the perceived point on the other side.

1) Building perceptual models: We have begun preliminary experiments aimed at building an error model for human perception based on different robot pointing modalities, including head, arm, and combined head and arm gestures. Each gesture's accuracy is evaluated in experiments with human participants, in which the angle, distance and point location are systematically varied. Target points are assumed to be on an approximately 2.5 x 3.5 m (8 x 12 ft) transparent acrylic screen. The participant is seated on one side of this screen, opposite the robot, as illustrated in Fig. 2. As the robot gestures to each point, the participant uses a laser pointer to indicate his perception of the target location on the screen. An experimenter then marks these points with a fiducial marker, and they are recorded using an upward-pointing laser rangefinder, yielding measurements accurate to within centimeters.

In the first iteration of this experiment all users were given a random set of points on a regular grid and we to aimed to use within subject comparison to determine how error responds to changes in the state variables. After conducting an analysis of variance from data collected from 11 participants we determined that errors were generally on the order of 30-60cm (1-2 feet), but were unable to draw deeper conclusions due to high variance in the sampled distributions. To address these issues we have redesigned the experiment to perform a between subjects comparison, with more participants. This is realized by interspersing a constant "calibration point," within the randomly distributed points presented to a user, to ensure that within-user accuracy is consistent. From this output, we construct an error model parameterized by human-robot-point locations and angles. During this experiment, we are also monitoring the head orientation of the participant in order to empirically determine how head direction can be used to model attention.

2) Validating attention, intention and intervention: Using this perceptual error model, we will conduct a further experiment to validate the estimated attentions and intentions within the context of a collaborative game-playing scenario. The game involves a robot indicating to the user a series of targets within a cluttered office environment; the user must then visit these targets in a specified order; this task is similar to, but less constrained than, that of our previous work [12], and was chosen specifically for comparison and analysis. The error model of human perception will be used to determine the position and orientation from which the robot should point to a target to ensure that the gesture is specified in a minimally (or maximally, for testing purposes) ambiguous manner. We are also investigating the use of a similar intention model to determine, at any point in time, the most likely intended target of a user, allowing the robot to intervene by redirecting the user, if necessary, to correct potential errors.

References

[1] D. Feil-Seifer and M. J. Matarić, "Defining socially assistive robotics," *Proceedings of the International Conference on Rehabilitation Robotics (ICORR-05)*, Chicago, Illinois, 2005.

[2] A. Tapus, M. J. Matarić, and B. Scassellati, "The grand challenges in socially assistive robotics," *IEEE Robotics and Automation Magazine*, vol. 14, no. 1, pp. 35-42, 2007.

[3] D. J. Feil-Seifer and M. J. Matarić, "Toward socially assistive robotics for augmenting interventions for children with autism spectrum disorders," *11th International Symposium on Experimental Robotics 2008*, vol. 54, pp. 201-210, Athens, Greece, July 2008.

[4] B. Grosz and C. Sidner, "Attention, intentions, and the structure of discourse," *Computational Linguistics*, vol. 12, no. 3, pp. 175–204, 1986.

[5] C. Breazeal, A. Brooks, J. Gray, G. Hoffman, C. Kidd, H. Lee, J. Lieberman, A. Lockerd, and D. Chilongo, "Tutelage and collaboration for humanoid robots," *International Journal of Humanoid Robotics*, vol. 1, no. 2, pp. 315–348, 2004.

[6] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[7] S. Blakemore and J. Decety, "From the perception of action to the understanding of intention," Nature Reviews Neuroscience, vol. 2, no. 8, pp. 561–567, 2001.

[8] M. L. Knapp and J. A. Hall, *Nonverbal communication in human interaction*, 7th ed., Boston, MA: Wadsworth Publishing, 2009.

[9] M. M. Louwerse and A. Bangerter "Focusing attention with deictic gestures and linguistic expressions", *Proceedings of the XXVII Annual Conference of the Cognitive Science Society (CogSci 2005)*, Stresa, Italy, July 21–23 2005.

[10] D. A. Norman, *The design of everyday things*, New York: Basic Books, 2002.

[11] R. Mead, J. B. Weinberg, and M. J. Matarić, "An ontology-based multimodal communication system for human-robot interaction in socially assistive domains", *IEEE ICRA2010 Workshop on Multimodal Human-Robot Interfaces*, Anchorage, AK, May 2010.

[12] R. Mead and M. J. Matarić, "The power of suggestion: teaching sequences through assistive robot motions," *Proceedings of The 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI-09)*, San Diego, CA, pp. 317-318, March 2009.