Nodding in Conversations with a Robot

Christopher Lee, Neal Lesh, Candace L. Sidner
Mitsubishi Electric Research Labs
Cambridge, MA 02139
{sidner, lee, lesh}@merl.com

Louis-Philippe Morency, Ashish Kapoor, Trevor Darrell
MIT
Cambridge, MA 02139
{lmorency, kapoor, trevor}@mit.edu

ABSTRACT
In this demo we describe our ongoing efforts to build a robot that can collaborate with a person in hosting activities. We illustrate our current robot's conversations, which include gestures of various types, and report on extensions to the robot's existing gestural abilities to be able to recognize nodding in conversations.

Author Keywords
Collaboration, conversation, human-robot interaction, nods.

ACM Classification Keywords
H5.m. Information interfaces and presentation: Miscellaneous (Robotics), H.5.2 Natural Language.

INTRODUCTION
This demo reports on recent extensions to our research toward developing the ability for robots to be engaged as they participate with humans in a collaborative interaction for hosting activities. Engagement is the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake. Engagement is supported by conversation (that is, spoken linguistic behavior), ability to collaborate on a task (that is, collaborative behavior), and gestural behavior that conveys connection between the participants.

In order to narrow our research efforts, we have focused on hosting activities. Hosting activities are a class of collaborative activity in which an agent provides guidance in the form of information, entertainment, education or other services in the user's environment. In applying our research, physical robots, serving as guides, replace human hosts in the environment. To do so, our goals include understanding the nature of human-to-human engagement, especially the role of gestures [5]. We then apply our findings to robots interacting with people.

CURRENT EFFORTS

In previous work at MERL, we have developed a stationary robot (depicted in Figure 1), that takes the form of a penguin and can participate in conversations with users about visiting a laboratory and about participating jointly with the robot in demonstrating lab equipment (6). The robot uses a dialogue system [4], speech recognition and synthesis, face detection [7], sound location, speech detection and object detection algorithms to track the visitor's face, to participate in spoken conversation, to turn to look and point at objects of interest and in a limited way, to notice where the visitor is looking. A excerpt of a typical demo conversation is given in Figure 2 (the full conversation is too long for this short report).

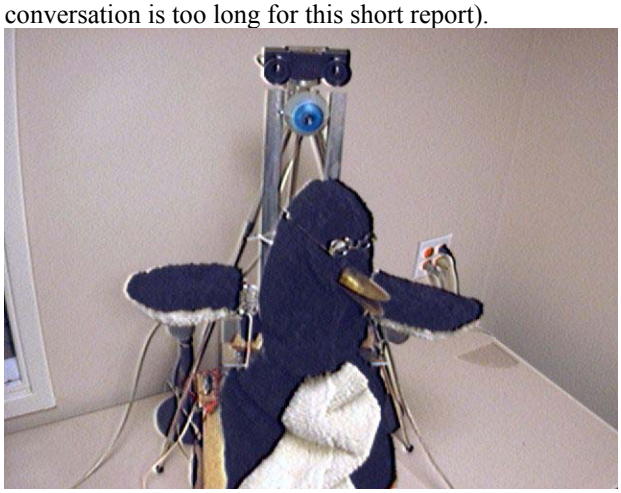

Figure 1. Mel the robotic penguin

M: Hi, I'm Mel a robotic penguin.
U: Hi.
M: What's your name?
U: Sam.
M: Sam, I'd like to show you a demo. OK?
U: No.
M: But it's really interesting. Come on. Try it!
U: Okay . (USER CAN NOD HERE)
M: It would be really nice to know what kind of progress people are making in their dining. So Paul Dietz created a new product called I Glassware for this. By the way, Sam, do you know Paul?
U: No.
M: Well, Paul is a member of MERL. This demo is one of his creations.
U: Okay. (USER CAN NOD HERE)

M: I Glassware stands for Instrumented glassware! Sam, right there <*robot turns head to table*> is the I Glassware cup <*robot turns head to user*>. And near it, is the table readout. The word re-fill on the readout means that it is empty……..

Figure 2. Portion of a typical interaction between Mel and a person

The robot however cannot pick up on a significant face-to-face conversational gesture: nodding. People nod in face-to-face conversation to indicate their agreement, or to backchannel[1] rather than speak. Furthermore, people even nod to robots despite their ability to pick up on this signal. In recent experiments [6] with human participants with our robot, 21 of 33 human participants nodded at our robot during the demonstration conversation even though the robot could not interpret their nods.

Production of nods by conversational partners would make it possible for robots to produce a more natural conversational style that extends beyond just interpreting nods as well as or in addition to spoken assents. Human conversation is not a series of sentences strung together to constitute a turn in the dialogue. Rather, utterances are in smaller chunks, some of which don't follow the classic noun phrase plus verb phrase form. They punctuated by brief pauses during which conversational partners use linguistic or non-verbal responses to ground the utterance [1]. A robot that interprets nods can produce this more typical form of language, thereby obviating the need to speak in long sentences that make up a turn.

In collaboration with colleagues at MIT, the MERL robot has been modified to interpret nods from its human partners. The robot uses a stereo camera. The head orientation and position of each visible speaker is estimated independently using adaptive view-based appearance models created online [3] from a two-frame registration algorithm. This tracking technique has bounded drift and can track heads undergoing large motion for long periods of time. Using a stereo camera for head pose tracking makes it less sensitive to lighting variations.

To detect head nods and head shakes in real time, an existing detection technique based on pupil tracking [2] was extended to use the output from the head pose tracker, providing better performance. A continuous Hidden Markov Model was trained to detect head nods and head shakes using recorded sequences of 11 subjects interacting with a simple character displayed on the screen. The complete system can robustly estimate the head position and orientation of each visible speaker as well as head nods and head shakes. The full conversation excerpted in figure 2 will be demonstrated with nodding by the human and robot participants.

[1] Backchannels are non-verbal "utterances" offered instead of a spoken utterances to indicate that the speaker may continue because the hearer is at least following the speaker's utterances.

In the conversation when the robot expects a back channel or an agreement to a robot command, the robot looks for head nods. The robot also nods slightly in response, in part to entrain participants to nod to the robot. For example in the conversation above, the robot looks for nods for a backchannel response; the second "okay," can be accomplished with just a head nod, and the robot interprets it accordingly. Speech and vision have been coordinated to fuse the recognition of nodding with spoken utterances**.** Otherwise a human participant could nod "yes" and say "no," as one participant in our previous experiments actually did.

This demo presents a robot that can do something totally new—to interpret and offer nods in collaborative conversation.

FUTURE WORK

Future work on Mel includes testing it with nodding with human participants to determine its ability to recognize nodes and whether it can encourage people to nod in the interaction? We will compare new results to previously collected data where robot did not nod nor recognize nods. The success of that endeavor will lead us to explore the production of conversational turns made up of smaller chunk utterances with the ability to recognize backchannels from human participants.

REFERENCES

1. Clark, H.H. *Using Language*, Cambridge University Press, Cambridge, 1996.
2. Kapoor, A. and Picard, R. A real-time head nod and shake detector, in *Proceedings from the Workshop on Perspective User Interfaces*, 2001.
3. Morency, L.-P.; Rahimi, A.; Darrell, T.; Adaptive view-based appearance models, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 1:8-20, 2003.
4. Rich, C., Sidner, C.L. and Lesh, N. COLLAGEN: Applying Collaborative Discourse Theory to Human-Computer Interaction, *AI Magazine*, *Special Issue on Intelligent User Interfaces*, AAAI Press, Vol. 22: 4: 15-25, 2001.
5. Sidner, C.L., Lee, C. and Lesh, N. Engagement when looking: behaviors for robots when collaborating with people, in *Diabruck: Proceedings of the 7th workshop on the Semantic and Pragmatics of Dialogue*, I. Kruiff-Korbayova and C.Kosny (eds.), pp. 123-130, 2003.
6. Sidner, C.L., Kidd, C.D., Lee, C. and Lesh, N. Where to look: a study of human-robot interaction, in submission, 2003.
7. Viola, P. and Jones, M. Rapid Object Detection Using a Boosted Cascade of Simple Features, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 905-910, 2001.