Video Demonstration of a Robotic Assistant Interpreting Multimodal Expressions

Miles Eldon, David Whitney, Stefanie Tellex
Brown University

In order for humans and robots to collaborate in complex tasks, robots must be able to understand people's references to objects in the external world. To refer to objects, people use a combination of language, gesture, and body language such as eye gaze and pointing. People provide these signals continuously, and a person's reference can quickly change based on new information about the domain. Moreover, a human listener responds to these signals as they are given using *backchannels*, for example nodding their head when they understand and looking confused or interrupting to ask a question when they do not. Clark [1996] refers to this continuous dance as *joint activity* and compares language use to playing a duet because of its collaborative nature, where both parties act to establish common ground and reduce uncertainty.

Despite the importance of real-time response to multimodal input, existing unimodal models do not integrate information from language and gesture [Matuszek et al., 2014, Tellex et al., 2011, Kollar et al., 2010]. Approaches that fuse information from language and gesture [Matuszek et al., 2014] do not take into account that information appears to the system over a period of time. These approaches make it impossible for a robot to provide back-channel feedback.

To provide a foundation for these capabilities, we propose a Bayes' filtering approach for interpreting multimodal information from language and gesture [Thrun et al., 2008]. Our framework relies on a factored observation probability that fuses information from language and gesture at a rate of 14Hz to continuously estimate the object a person is referring to in the real world. We assume each observation z is of the form $< l, r, h, s >$ where l, r, and h are the vectors for the left arm, right arm, and head respectively, while s is the recognized speech for that time step. At each time step we update our belief that the user is referencing an object x with respect to these observations.

In the beginning of the video, the system quickly updates its distribution based on varying information from different types of expressions over time. For example, an ambiguous utterance such as "I want a spoon," provides some information, but not enough to discern which of two spoons is desired. However, when this utterance is followed by a gesture, as in the video, the system immediately recognizes the desired object. We demonstrate our by providing quantitative results on a real-world RGB-D corpus of people referring to objects with language and gesture. These results, which appear in Table 1, show that our model can correctly determine the true object being reference by a user 90% of the time in real time.

Table 1: Real-world Results

Random	25%
Language only	46.15%
Gesture only	80.0%
Head only	18.46%
Multimodal (Language and Gesture)	**90.77%**
Multimodal (All)	61.54%

Without a robot, it is clear to see how the system interprets the user, but there is little for the user to interpret in return. With the addition of a robot, we can incorporate back channels from the robot to the user. In the video, the robot has a happy face when it has determined the desired object and a confused face otherwise. That, in addition to physical gestures from the robot help the human user interact with Baxter in a smooth way, adding clarification only when necessary.

References

H. H. Clark. *Using Language*. Cambridge University Press, May 1996. ISBN 0521567459.

T. Kollar, S. Tellex, D. Roy, and N. Roy. Toward understanding natural language directions. In *Proceedings of HRI-2010*, 2010.

C. Matuszek, L. Bo, L. Zettlemoyer, and D. Fox. Learning from unscripted deictic gesture and language for human-robot interactions. 2014.

S. Tellex, T. Kollar, S. Dickerson, M. Walter, A. Banerjee, S. Teller, and N. Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proc. AAAI*, 2011.

S. Thrun, W. Burgard, and D. Fox. *Probabilistic robotics*. MIT Press, 2008.