Interpreting Multimodal Referring Expressions in Real Time

Miles Eldon[1] and Stefanie Tellex[1]

***Abstract*— Robots that collaborate with humans must be able to identify objects used for shared tasks, for example tools such as a knife for assistance at cooking, or parts such as a screw on a factory floor. Existing work has addressed this problem in single modalities, such as natural language or gesture, but a gap remains in creating real-time multimodal systems that simultaneously fuse information from language and gesture in a principled mathematical framework. We define a multimodal Bayes' filter for interpreting referring expressions to object using language and gesture in real time. We collected a new RGB-D and audio dataset of people referring to objects in a tabletop setting and demonstrate that our approach successfully integrates information from language and gesture in real time to quickly and accurately identify objects continuously.**

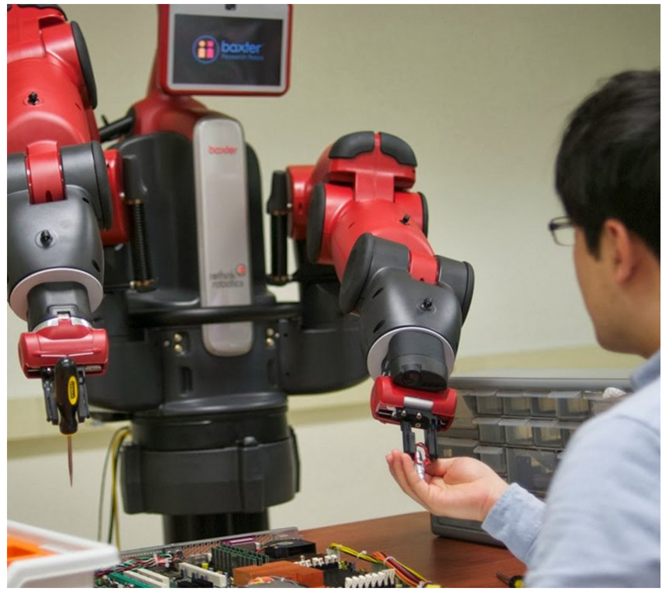

Fig. 1. Robots that collaborate with people need to understand their references to objects in the environment. For example, if a person asks for a tool using language and gesture, the robot needs to interpret the person's reference in order to pick up the correct tool.

I. INTRODUCTION

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, hand gestures, and body language such as eye gaze. These signals are provided continuously to the robot over time, and can change rapidly as a person responds to their environment. For example, Figure 1 shows a robot handing a tool to a human collaborator for a manufacturing task. For the robot to infer the correct tool to deliver, it must interpert a person's language and gesture over time.

Most existing approaches for interpreting language and gesture rely on unimodal models that do not integrate the two information sources, even though people fluidly use language and gesture together. 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. If a robot could exploit this temporal information, it can respond in real time to a person's input and indicate when it has understood; this backchannel feedback could enable a person to respond and provide additional clarifying input when the robot fails to understand.

To support these types of interactions, 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, hand gestures, and head gestures in real time to continously estimate the object a person is referring to in the real world. We demonstrate our model in simulation, as well as providing quantitative results on a real-world RGB-D corpus of people referring to objects in the environment. These results demonstrate that our approach quickly and accurately fuses multimodal information in real time to continously estimate the object a person is referring to.

II. RELATED WORK

Our approach is related to Holladay et al. [2014] but focuses on interpreting a person's gestures rather than enabling a robot to generate pointing gestures. A large body of work focuses on language understanding for robots [MacMahon et al., 2006, Dzifcak et al., 2009, Kollar et al., 2010, Matuszek et al., 2012]. Guadarrama et al. [2014] presents a framework for interperting open-domain references to objects but focuses on interperting language rather than language combined with gesture.

Many existing approaches for interpreting gesture rely on fixed vocabularies of gesture, such as "stop" or "follow" [Waldherr et al., 2000, Marge et al., 2011] without a princepled way for fusing information from language and gesture. Our work unifies language and gesture interpretation into a single mathematical framework, and focuses on parameterized gestures such as pointing.

Matuszek et al. [2014] presented a multimodal framework for interpreting unscripted references to tabletop objects using language and gesture. Our approach similarly focuses on tabletop objects but uses language, gesture, and head pose, and integrates these disparate data sources continuously

[1]Computer Science Department, Brown University

over time using a Bayes' filtering framework. This approach enables the robot to continuously process new information and produce an estimate that converges over time to the correct object as new information is observed from the person.

III. TECHNICAL APPROACH

Our aim is to estimate a distribution over the object that a person is referring to given language and gesture inputs. We frame the problem as a Bayes' filter [Thrun et al., 2008], where the hidden state, $\mathcal{X}$, is the set of m objects in the scene that the person is currently referencing. The robot observes the person's actions and speech, $\mathcal{Z}$, and at each time step estimates a distribution over $\mathcal{X}$:

$$p(x_t|z_0 \dots z_{0:t}) \tag{1}$$

To estimate this distribution, we take a Bayes' filtering approach and alternate performing a time update and a measurement update. The time update updates the belief that the user is referring to a specific subset of objects given previous information:

$$p(x_t|z_{0:t-1}) = \int p(x_t|x_{t-1}) \times p(x_{t-1}|\mathcal{Z}_{0:t-1}) \mathrm{d}x_{t-1} \tag{2}$$

The measurement update combines the previous belief with the newest observation to update each belief state:

$$p(x_t|\mathcal{Z}_{0:t}) = \frac{p(z_t|x_t) \times p(x_t|z_{0:t-1})}{p(z_t|z_{0:t-1})} \tag{3}$$

$$\propto p(z_t|x_t) \times p(x_t|z_{0:t-1}) \tag{4}$$

Algorithm 1 shows pseudocode for our approach.

A. Prediction Model

ST: Miles, can you check if this math is right?

We assume that a person is likely to continue referring to the same object, but at each timestep has a small probability, c, of transitioning to a different object:

$$p(x_t|x_{t-1}) = \begin{cases} 1-c & \text{if } x_t = x_{t-1} \\ c & \text{otherwise} \end{cases} \tag{5}$$

In our experiments, c has a value of XXX.

B. Observation Model

We assume access to an observation model of the form:

$$p(z_t|x_t) \tag{6}$$

Observations consist of a tuple consisting of a person's actions, $\langle l, r, h, s \rangle$ where:

- l represents the observed origin (l_o) and vector (l_v) for the left arm.
- r represents the observed origin (r_o) and vector (r_v) for the right arm .
- h represents the observed origin (h_o) and vector (h_v) for head.
- s represents the observed speech from the user, consisting of a list of words.

Formally, we have:

$$p(z_t|x_t) = p(l, r, h, s|x_t) \tag{7}$$

We factor assuming that each modality is independent of the others given the state (the true object that the person is referencing):

$$= p(l|x_t) \times p(r|x_t) \times p(h|x_t) \times p(s|x_t) \tag{8}$$

The following sections describe how we model each type of input from the person.

Gesture. We model pointing gestures as a vector through three dimensional space. We calculate the angle between the gesture vector and the vector from the gesture origin to the mean of each cluster, and then use the PDF of a Gaussian with trained variance to determine the weight that should be assigned to that object. Let $\Phi(< origin >, < point >, < point >)$ give the angle between the two points with the given origin.

$$p(l|x_t) \propto \mathcal{N}(\mu_l = 0, \sigma_l, \Phi(l_o, l_v, x_t)) \tag{9}$$

$$p(r|x_t) \propto \mathcal{N}(\mu_r = 0, \sigma_r, \Phi(r_o, r_v, x_t)) \tag{10}$$

Head Pose. Head pose is modeled in the same manner as arm gestures.

$$p(h|x_t) \propto \mathcal{N}(\mu_h = 0, \sigma_h, \Phi(h_o, h_v, x_t)) \tag{11}$$

Speech. We model speech with a simple bag of words model. We take the words in a given speech input and count how many words in this text match descriptors (denoted x_d) of specific objects.

$$p(s|x_t) = \prod_{w \in s} p(w|x_t) \tag{12}$$

C. Training Model Parameters

We train model parameters by fitting each factor to a ground-truth distribution.

Algorithm 1: Interactive Bayes Filtering Algorithm

Input: $bel(x_{t-1}), u_t, z_t$

Output: $bel(x_t)$

for x_t **do**

$\quad \bar{bel}(x_t) =$

$\quad bel(x_t) =$

end

ST: Miles – can you fill in the algorithm? We shouldn't just reproduce a simple Bayes' filter; it should get into what specific things we had to do to make it work with multimodal input.

Fig. 2. Scene from our dataset.

Language only	XX
Gesture only	XX
Head only	XX
Multimodal	XX

TABLE I
REAL-WORLD RESULTS

IV. EVALUATION

We evaluate our model in simulation, comparing the full model to versions without multimodal information. Additionally we assessed its performance on an RGB-D audio and video corpus of people referring to objects. Finally we created an end-to-end robotic demonstration, demonstrated in the video attachement to our paper, and available online[1].

A. Simulation Results

B. Real-world Corpus-Based Results

Our real-world experiments measured our algorithm's performance when a person referred to an object visually and with gesture. The person was seated roughly five feet from a table which had several objects, as shown in Figure 2. We instructed them to refer to the object as if they were talking to another person, using language and gesture. We indicted the object to refer to using a laser pointer, and we periodically shifted to a different object on a predetermined schedule. They wore a microphone to pick up high-quality audio, and we tracked their body pose using the NITE tracker [ope, 2014]. We used the Google Voice Recognition package to recognize speech. We used two Kinects: one pointed outward at the person, and one pointed down to recognize objects on the table.

C. Robotic Demonstration

Because our approach enables a robot to quickly and constantly monitor a person's references to an object, a robot can respond to these estimates in real time. We demonstrate this behavior by enabling Baxter to demonstrate its certainty about what object is being referenced, this eliciting more feedback from the person. When the robot is very unsure, its arm moves back and forth between the candidate objects. When it is sure, then it moves with more precision. The video shows this behavior, as a person provides more information about what object is being referred to by the person.

V. CONCLUSION

We have demonstrated a Bayes' filtering approach to interperting a person's multimodal language and gesture references to objects continoulsly in real time. Our approach enables a robot to understand a person's references to objects in the real world.

In the future we plan to expand our language model to incorporate models of compositional semantics and lower-level visual features so that the robot is not limited to prespecified object models.

[1] video reference

REFERENCES

Openni tracker. http://wiki.ros.org/openni_tracker, 2014.

J. Dzifcak, M. Scheutz, C. Baral, and P. Schermerhorn. What to do and how to do it: translating natural language directives into temporal and dynamic logic representation for goal management and action execution. In *Proceedings of the 2009 IEEE International Conference on Robotics and Automation*, pages 3768–3773, Piscataway, NJ, USA, 2009. IEEE Press. ISBN 978-1-4244-2788-8.

S. Guadarrama, E. Rodner, K. Saenko, N. Zhang, R. Farrell, J. Donahue, and T. Darrell. Open-vocabulary object retrieval. RSS, 2014.

R. M. Holladay, A. D. Dragan, and S. S. Srinivasa. Legible robot pointing. RO-MAN, 2014.

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

M. MacMahon, B. Stankiewicz, and B. Kuipers. Walk the talk: connecting language, knowledge, and action in route instructions. In *Proceedings of the 21st national conference on artificial intelligence*, volume 2 of *AAAI '06*, pages 1475–1482. AAAI Press, 2006. ISBN 978-1-57735-281-5.

M. Marge, A. Powers, J. Brookshire, T. Jay, O. C. Jenkins, and C. Geyer. Comparing heads-up, hands-free operation of ground robots to teleoperation. *Robotics: Science and Systems VII*, 2011.

C. Matuszek, E. Herbst, L. Zettlemoyer, and D. Fox. Learning to parse natural language commands to a robot control system. In *Proc. of the 13th Intl Symposium on Experimental Robotics (ISER)*, 2012.

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

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

S. Waldherr, R. Romero, and S. Thrun. A gesture based interface for human-robot interaction. *Autonomous Robots*, 9(2):151–173, 2000.