Interpreting Multimodal Referring Expressions in Real Time

Miles Eldon[1] and Stefanie Tellex[1]

***Abstract*— Identifying objects for shared tasks, such as a knife for assistance at cooking, or a screw used to assemble a part on a factory floor, is a key part of many human-robot collaborative tasks. Robots that collaborate with people must be able to understand their references to objects in the environment. 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 mathemtical framework. We define a multimodal Bayes' filtering approach to interpreting referring expressions to object using language and gesture. 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.**

I. INTRODUCTION

II. RELATED WORK

[?]

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, where the hidden state, $\mathcal{X}$, is the set of m objects in the scene that can be referenced. The robot observes the person's actions and speech, $\mathcal{Z}$, and at each timestep estimates a distribution over $\mathcal{X}$:

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

The time update updates the belief that the user is in any specific state given all 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:

$$\begin{aligned} p(x_t|\mathcal{Z}_{0:t}) &= \frac{p(z_t|x_t) \times p(x_t|z_{0:t-1})}{p(z_t|z_{0:t-1})} && (3)\\ &\propto p(z_t|x_t) \times p(x_t|z_{0:t-1}) && (4)\end{aligned}$$

A. Observation Model

We assume access to an observation model of the form:

$$p(z_t|x_t) \tag{5}$$

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.

$$p(z_t|x_t) = p(l, r, h, s|x_t) \tag{6}$$

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

Gesture. We model gesture 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.

$$p(l|x_t) \propto \Phi(l_o, l_v, x_t) \tag{8}$$

$$p(r|x_t) \propto \Phi(r_o, r_v, x_t) \tag{9}$$

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

$$p(h|x_t) \propto \Phi(h_o, h_v, x_t) \tag{10}$$

ME: More concise way to show this? They are all the same besides variance. I guess we could just do $\prod_{g\in\{h,l,r\}}$

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.

$$p(s|x_t) = \prod_{w\in s} p(w|x_t) \tag{11}$$

IV. EVALUATION

V. CONCLUSION

VI. REFERENCES

[1] Computer Science Department, Brown University