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 and a state for no item being referenced. The robot observes the person's actions and speech, $\mathcal{Z}$, and at each timestep estimates a distribution over $\mathcal{X}$:

$$p(\mathcal{X}_t|\mathcal{Z}_0 \ldots \mathcal{Z}_{0:t}) \tag{1}$$

The time update is the probability that the person will change the object they are referring to at the next time step:

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

The measurement update incorporates an estimate of the updated state based on new observations of the person's actions:

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

$$\propto p(\mathcal{Z}_t|x_t) \times p(x_t|\mathcal{Z}_{0:t-1}) \tag{4}$$

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 probability of a gesture by examining every three dimensional particle (denoted as q) in an object and calculating the angle (function Φ) between the vector formed by the gesture and the vector formed with that particle. We then use a Gaussian distribution with a variance (σ) found during training to calculate the probability of seeing that angle difference. We then take the product of each of these points and normalize it so that small objects with few particles don't always dominate the probability distribution. The probability of each gesture given the state is as follows:

$$p(l|x_t) = [\prod_{q \in x_t} \mathcal{N}(\mu_l = 0, \sigma_l, \Phi(l_o, l_v, q))]^{(\frac{\sum_{x' \in \mathcal{X}} len(x'_p)}{len(x_p)})} \tag{8}$$

$$p(r|x_t) = [\prod_{q \in x_t} \mathcal{N}(\mu_r = 0, \sigma_r, \Phi(r_o, r_v, q))]^{(\frac{\sum_{x' \in \mathcal{X}} len(x'_p)}{len(x_p)})} \tag{9}$$

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

$$p(h|x_t) = [\prod_{q \in x_t} \mathcal{N}(\mu_h = 0, \sigma_h, \Phi(h_o, h_v, q))]^{(\frac{\sum_{x' \in \mathcal{X}} len(x'_p)}{len(x_p)})} \tag{10}$$

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.

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

[1]Computer Science Department, Brown University

IV. EVALUATION
V. CONCLUSION
VI. REFERENCES