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

[Matuszek et al., 2012]

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 $\mathcal{X}$:

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

ST: Time update and measurement update should go here, with z and not the tuple. Then the observation model section should explain how each model works.

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

$$P(x_t | \mathcal{Z}_{0:t-1}) = \quad (2)$$

ST: Should be an integral over the previous state

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

$$P(x_t | \mathcal{Z}_{0:t}) \quad (3)$$

ST: Should be Bayes's rule.

[1] Computer Science Department, Brown University

A. Observation Model

We assume access to an observation model of the form:

$$p(z_t | x_t) \quad (4)$$

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.

$$p(z_t | x_t) = p(l, r, h, s | x_t) \quad (5)$$
$$p(z_t | x_t) = p(l | x_t) \times p(r | x_t) \times p(h | x_t) \times p(s | x_t) \quad (6)$$

Gesture. We model gesture ....
ST: Fill in equations here
Head Pose.
ST: Fill in equations here
Speech.
ST: Fill in equations here

- Transition Function
  - $\mathcal{T}$: a function such that $\mathcal{T}(x_a, x_b)$ is equivalent to the probability that x_a transitions to x_b
  - **ST: Model with Poisson?**
- Φ: a function that, given an origin and two points, returns the angle between the two points
  - Applied as $\Phi(\text{origin}, p_1, p_2)$
- $\mathcal{I}$: an indicator function applied as $\mathcal{I}(\text{word}, \text{corpus}$ that returns 1 if the word is in the corpus and 0 otherwise
- len: the number of items in a list

Equations

- Time Update
  - An equation used to determine the probability that $\mathcal{X}_t = x$ given only previous belief states
  - $P(\mathcal{X}_t = x | \mathcal{X}_{t-1} \ldots \mathcal{X}_0) = \sum_{x' \in \mathcal{X}} \mathcal{T}(x, x') * bel(\mathcal{X}_{t-1} = x')$
  - This function computes the new probability for a state by multiplying the probability for each past state times transition probability and then summing all these new probabilities.
- Measurement Update

- An equation used to determine the belief that $\mathcal{X}_t = x$ given observation $\mathcal{Z}_{t-1}$

- $P(\mathcal{X}_t = x|\mathcal{Z}_{t-1}) = P(X_t = x|l_{t-1}) * P(X_t = x|r_{t-1}) * P(X_t = x|h_{t-1}) * P(X_t = x|s_{t-1})$

- $P(\mathcal{X}_t = x|l_{t-1}) = [\prod\limits_{p \in x_p} \mathcal{N}(\mu_l = 0, \sigma_l, \Phi(l_o, l_v, p))]^{(\frac{\sum_{x' \in \mathcal{X}} len(x'_p)}{len(x_p)})}$

- $P(\mathcal{X}_t = x|r_{t-1}) = [\prod\limits_{p \in x_p} \mathcal{N}(\mu_r = 0, \sigma_r, \Phi(r_o, r_v, p))]^{(\frac{\sum_{x' \in \mathcal{X}} len(x'_p)}{len(x_p)})}$

- $P(\mathcal{X}_t = x|h_{t-1}) = [\prod\limits_{p \in x_p} \mathcal{N}(\mu_h = 0, \sigma_h, \Phi(h_o, h_v, p))]^{(\frac{\sum_{x' \in \mathcal{X}} len(x'_p)}{len(x_p)})}$

- These three equations are all very similar. They basically compute the probability of each point being seen given the vector of the gesture and multiply this out for all points. Finally, it is raised to a power equal to the ratio of total points to the number of points in the given object to normalize so that small objects don't have much larger probabilities.

- $P(\mathcal{X}_t = x|s_{t-1}) = \dfrac{\sum\limits_{w \in s_{t-1}} \mathcal{I}(w, x_d)}{\sum\limits_{x \in \mathcal{X}} \sum\limits_{w \in s_{t-1}} \mathcal{I}(w, x'_d)}$

- This simply computes the ratio of words in the spoken phrase that are in the descriptors of the object to the number of the words that match any object

- Belief Update
  - An equation that produces the probability that $\mathcal{X}_t = x$ given all past observations and belief states, namely the product of the measurement and time updates
  - $bel(X_t = x) = P(\mathcal{X}_t = x|\mathcal{Z}_{t-1}) * P(\mathcal{X}_t = x|\mathcal{X}_{t-1}....\mathcal{X}_0)$
  - $bel(\mathcal{X}_0 = x) = \frac{1}{m+1}$ (uniform initialization of belief)

IV. EVALUATION

V. CONCLUSION

VI. REFERENCES

REFERENCES

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.