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

Definitions

- Hidden State Space
  - $\mathcal{X}$: the set of m objects in the scene that can be referenced and a state for no item being referenced
  - An example would be: $x =$ Object 1 is being referenced
  - Each object in the scene has an associated set of three dimensional points (x_p) and set of descriptor keywords (x_d)
- Observations
  - $\mathcal{Z}$: the set of four-tuple $\{l, r, h, s\}$
  - 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
- 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?**
- $\mathcal{N}$: a function that returns the probability of a sample under a Gaussian distribution given a mean and variance
  - Applied as $\mathcal{N}(\mu, \sigma, \text{sample})$
- 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

Equations

- Time Update
  - An equation used to determine the probability that $\mathcal{X}_t = x$ given only previous belief states
  - $P(\mathcal{X}_t = x|\mathcal{X}_{t-1}....\mathcal{X}_0) = \sum\limits_{x' \in \mathcal{X}} \mathcal{T}(x, x') * bel(\mathcal{X}_{t-1} = x')$
- Measurement Update
  - An equation used to determine the belief that $\mathcal{X}_t = x$ given observation $\mathcal{Z}_{t-1}$
  - $P(\mathcal{X}_t = x|\mathcal{Z}_{t-1}) = P(X_t = x|l_{t-1}) * P(X_t = x|r_{t-1}) * P(X_t = x|h_{t-1}) * P(X_t = x|s_{t-1})$
  - $P(\mathcal{X}_t = x|l_{t-1}) = \prod\limits_{p \in x_p} \mathcal{N}(\mu_l = 0, \sigma_l, \Phi(l_o, l_v, p))$
  - $P(\mathcal{X}_t = x|r_{t-1}) = \prod\limits_{p \in x_p} \mathcal{N}(\mu_r = 0, \sigma_r, \Phi(r_o, r_v, p))$
  - $P(\mathcal{X}_t = x|h_{t-1}) = \prod\limits_{p \in x_p} \mathcal{N}(\mu_h = 0, \sigma_h, \Phi(h_o, h_v, p))$
  - $P(\mathcal{X}_t = x|s_{t-1}) = \dfrac{\sum\limits_{w \in s_{t-1}} \mathcal{I}(w, x_d)}{\sum\limits_{x \in \mathcal{X}} \sum\limits_{w \in s_{t-1}} \mathcal{I}(w, x'_d)}$
- 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

[1]Computer Science Department, Brown University

system. In *Proc. of the 13th Intl Symposium on Experimental Robotics (ISER)*, 2012.