Multichannel Robotic Interaction With Gesture and Language

Anonymous for submission.

Abstract Humans communicate about objects using language and gesture, fusing information from multiple modalities and responding to those communications in real time. For robots to make the leap from factories into homes, they must be able to rapidly understand human communication and respond through the use of backchannels, gestures and language that seek to clarify goals in joint activities. Existing work has addressed this problem in single modalities, such as natural language or gesture, or fused modalities to make reactive, but non-realtime, systems, but a gap remains in creating systems that can continuously and simultaneously fuse information from language and gesture and respond to that information in real time. To address this problem, we define a multimodal POMDP for interpreting and responding to a users referring expressions. Finding this formalized POMDP to be intractable in the real world, we then provide an approximation that we demonstrate accurately and quickly interacts with human users in the desired way. **JGO: Miles take a stab at a better title.**

Brown University

1 Introduction

2 Related Work

3 Overview

3.1 Robotic Setup

JGO: Describe the warehouse (if there is one), the objects on the table, the user's position. Describe the robot and the microphone.

3.2 Pick and Place

JGO: I have the lock on this subsection.

3.3 Integrating Speech and Gestures With Single Agent POMDP

3.4 Evaluation

4 Single Agent POMDP

Defined by:
A set of states $\mathscr{X}$
A set of actions $\mathscr{A}$
The conditional transition probabilities between states $\mathscr{T}$
The reward function $\mathscr{R} : \mathscr{S} \times \mathscr{A} \rightarrow R$
A set of observations Z
A set of observation probabilities $\mathscr{O}$
$\gamma \in [0,1]$ a discount factor

$\mathscr{X}$ is the object the robot thinks the person wants.

$\mathscr{A} = (M,P)$ where $M = \{$'waiting', 'pointing', 'sweeping', 'grouping', 'touching', 'handing'$\}$ and P represents a set of objects being acted upon (limited to one for 'pointing', 'touching', and 'handing' and zero for 'waiting').

Z consists of the observed speech (s), left and right arm gestures (l and r), and also the user belief in the robots belief (u).The user's belief is assumed to be ob-

served variable because it is assumed to depend only on the robots actions, which are known to the robot. At every time step, the belief degrades slightly back towards uniform if no non-waiting action is observed. (Note that this can be computed at any time given the entire history of robot actions)

We assume that a person is likely to continue referring to the same object, but at each timestep has a large probability, c, of not transitioning to a different object:

$$\mathscr{T}(x_t|x_{t-1},a_{t-1}) = \begin{cases} c & \text{if } x_t = x_{t-1} \\ \frac{1-c}{|X|-1} & \text{otherwise} \end{cases} \quad (1)$$

To calculate $\mathscr{O}(o|x_t,a)$, where $o = \{l,r,s,u\}$, that l, r, s, and u are conditionally independent when conditioned on the state.

4.1 Gesture

We model pointing gestures as a vector through three dimensional space. First, we calculate a gesture vector using the skeleton pose returned by NITE. For arms, we compute a vector from the elbow to the wrist, then project this vector so that the origin is at the wrist. For head pose, we compute a vector based on the body orientation. Next, we calculate the angle between the gesture vector and the vector from the gesture origin to the center of each object, and then use the PDF of a Gaussian ($\mathscr{N}$) with variance (σ) to determine the weight that should be assigned to that object. We define a function $\mathrm{A}(o,p_1,p_2)$ as the angle between the two points, p_1 and p_2 with the given origin, o. Then

$$p(l|x_t) \propto \mathscr{N}(\mu_l = 0, \sigma_l)[A(l_o, l_v, x_t)] \quad (2)$$

$$p(r|x_t) \propto \mathscr{N}(\mu_r = 0, \sigma_r)[A(r_o, r_v, x_t)] \quad (3)$$

If the person's arm is more than a certain angle away from the table, we assume they are referring to none of the objects, and perform an update. As a result, these gestures do not effect the robot's estimate of the objects being referenced.

4.2 Speech

We model speech with a unigram model, namely we take each word in a given speech input and calcuate the probability that, given the state, that word would have

been spoken.

$$p(s|x_t) = \prod_{w \in s} p(w|x_t) \tag{4}$$

4.3 User Estimate

We assume that the user estimate of the robots belief, u, is an observed variable. The user belief u is calculated from the history of robot actions (with slight degradation over time, similar to the transition function) and then incorporated into the robots state in $\mathcal{O}(o|x_t, a)$. Since u is a distribution, $p(u|x_t, a) = u(x_t)$, namely, the users belief that the robot is in state x_t.

5 Experiments

6 Discussion

7 Conclusion

Tellex et al. [1]

References

[1] S. Tellex, T. Kollar, S. Dickerson, M.R. Walter, A. Banerjee, S. Teller, and N. Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proc. AAAI*, 2011.