# Towards Meaningful Human-Robot Collaboration on Object Placement

Advik Iyer Guha, Stefanie Tellex
Humans To Robots Laboratory
Brown University
{aiguha, stefie10}@cs.brown.edu

## I. Introduction

Countless collaborative tasks that we envision for humans and robots *need* robotic object placement. We imagine robots that can help humans at home (cooking meals, laying the table, moving around furniture), and ones that can help at work (handing a mechanic specific tools as needed, aiding a doctor in surgery, carrying boxes and inventory in warehouses). In almost every such task, a robot must not only be able to pick up the object of interest, but also be able to put the object back down, and do so in the *right place*.

Collaborating with a human to infer this right place is a challenging and important problem. Three capabilities are required for a robot to perform this task. First, the robot must be able to understand how humans reference space in the physical world. Humans reference space with expressions that combine speech and gesture, similar to how they refer to objects (Eldon et al. [2]). Unlike objects however, space is a *continuous* variable; this both changes how humans construct referring expressions, and makes the inference task more complex and computationally expensive. Second, the robot must be able to understand and respond to its human collaborator in real-time. Third, the robot must be able to provide continuous feedback on its understanding to the human (termed *social feedback* by Wu et al. [7]), since this can significantly speed up the collaboration process (Clark and Krych [1]).

We propose a planning framework for collaborative object placement in tabletop environments that has these three capabilities. First, it maintains a discretized spatial model of the placement environment to understand space-referring gesture and speech. Second, it uses a `POMDP` that receives multimodal observations and chooses multimodal actions to interact with the human and continuously estimate their placement objective. Third, it continually quantifies its own understanding of the objective, and visualizes this for the human.

To evaluate our proposed framework, we conducted a pilot study where participants collaborate with our Rethink Robotics Baxter Robot to place an object on a table. Our findings are discussed below.

## II. Related Work

Kaelbling et al. [3] advocates for the modelling of sequential planning problems in partially observable domains as Partially Observable Markov Decision Processes. Within a `POMDP`, the agent maintains, through observations and actions, a *belief state* or probability distribution over possible states, in lieu of concrete knowledge of the current state. In Wu et al. [7], a `POMDP` is used to model the task of object pickup and hand-off in a tabletop environment. Though the intended approach and environment are similar, their focus is on a conceptually and computationally different problem than object placement. Still, these two works will be a foundation for ours.

In Matuszek et al. [6], a corpus of unscripted language and gesture object-referring expressions is used to train models to identify objects. However, these models are static in time. In Eldon et al. [2], object-referral expressions *are* treated as dynamic (occurring over time). However, this model is more rule-based than collaborative, since robot actions are predefined to occur at certain observation thresholds. Our intended `POMDP` framework allows us to perform higher-level reasoning over time, thus moving towards a more meaningful collaboration experience.

## III. Technical Approach

First, we discretize the tabletop environment into a two-dimensional grid $\mathcal{G}$ to make the spatial inference more computationally tractable. Then, we define the collaboration task as a `POMDP` with the robot as the agent, and use an approximate solving technique called Belief Sparse Sampling (Kearns et al. [4]). Finally, we quantify the robot's understanding using the internally maintained `POMDP` belief state, and convey this visually to the human.

The framework has been implemented using the Brown-UMBC Reinforcement Learning and Planning Library (BURLAP, [5]).

### A. POMDP Definition

The proposed `POMDP` framework is formally defined by the seven-tuple, $(S, A, T, R, \Omega, O, \gamma)$.

**States** $S$ is the set of states in this domain, each of which is a tuple $(\chi, \lambda)$:

- $\chi$ is the object being held by the robot
- $\lambda$ is a grid cell in $\mathcal{G}$ where the human collaborator wants $\chi$ to be placed (partially-observable)

**Actions** $A$ is the set of actions that can be taken by the robot. For each cell $c$ of grid $\mathcal{G}$, we define the following actions:

- `PLACE`, which places the object $\chi$ in cell $c$
- `HOVER`, which hovers the held object over cell $c$, and asks the user a question

We also define a strictly information-gathering action:

- `PROMPT`, which simply waits and asks for further input from the human

**Transitions** $T$ is the transition function governing the probability of moving to a new state, given a current state and a particular action. We model $\lambda$ has having a slow uniform decay over all grid cells of the table grid $\mathcal{G}$. This makes our planning framework robust in situations where the human changes their mind.

Specifically, we assume that the human collaborator has a high probability $k$ of referring to the same goal location $\lambda$ during the task, and therefore a probability $(1-k)$ of changing their mind. We further assume that switching to any other location is equally probable. These assumptions lead to the following transition model, for $\lambda$:

$$p(\lambda_{t+1}|\lambda_t, a_t) = \begin{cases} k & \text{if } \lambda_t == \lambda_{t+1} \\ \frac{1-k}{n-1} & \text{otherwise} \end{cases} \tag{1}$$

where $k > 0.99$, $a_t$ is `WAIT` or `HOVER`, and $n$ is the number of states in our belief, i.e, the dimensions of our grid abstraction, $\mathcal{G}$.

**Reward Function** $R$ is the reward function, governing the reward received given the current state and the action taken. A correct `PLACE` returns a reward of $+5$. An incorrect `PLACE` incurs a large negative reward of $-20$. The `HOVER` and `PROMPT` actions incur small negative rewards, to encourage the framework to choose these actions in the short term to gather information, but to converge quickly.

**Observations** $\Omega$ is the set of observations, each of which is a tuple $o = (l, g)$, representing the human collaborator's language, and the human collaborator's gestures.

The observation $l$ is the natural, unaltered transcription of the user's speech. The gesture observation $g$ is a set of four vectors representing the world $(x, y, z)$ coordinates of the human collaborator's left and right shoulders and wrists. We use this information to calculate the targets of their pointing gestures.

**Observation Function** $O$ is the observation function, governing the probability of witnessing a particular observation, based on the action taken and the resulting state in which said observation occurred. Given the variables defined so far, the function is:

$$O = p((l_{t+1}, g_{t+1})|s_{t+1}, a_t) \tag{2}$$

We assume (similar to [7]), the conditional independence of the language and gesture components of our observations, simplifying the observation function to be:

$$O = p(l_{t+1}|s_{t+1}, a_t).p(g_{t+1}|s_{t+1}, a_t) \tag{3}$$

*1) Language Model:* Our language model can understand two types of speech from the human:

- Simple Affirmative/Negative Expressions
- Relative Location Referring Expressions

Affirmative/Negative Expressions are expressions like *"yes, put it there"*. These expressions support simple question-answer dialogue between robot and human. Relative Location Referring Expressions are expressions that refer to a location in *relation* to other known locations in the environment like *"to the left of your hand, Baxter"*. These expressions support descriptive dialogue between human and robot.

*2) Gesture Model:* Our gesture model can understand pointing gestures from the human. These are defined as the vector $v$ collinear with the human collaborator's raised forearm, with the point of intersection $p$ of this vector with the tabletop considered the intended target, (similar to [2]).

For the observation function, the human collaborator's pointing target $p$ is considered to be sampled from a bivariate Gaussian (normal) distribution centered on the human's true goal placement location. In terms of our `POMDP` variables, this is:

$$p(g_{t+1}|s_{t+1}, a_t) \propto \mathcal{N}_{2,(\mu,\Sigma)}[(p.x, p.y)] \tag{4}$$

### *B. Backchannel Feedback*

We designed two visualizations of the `POMDP` belief state to show the human collaborator, as forms of backchannel feedback.

**Heat Map** The heat map visualizes the continuously updating belief state of the `POMDP` as *heat* on the tabletop grid $\mathcal{G}$; high heat corresponds to high belief that the human wants the object in that grid cell.

Figure 1 shows three frames of the heatmap belief state visualizer, at different stages of a placement task.

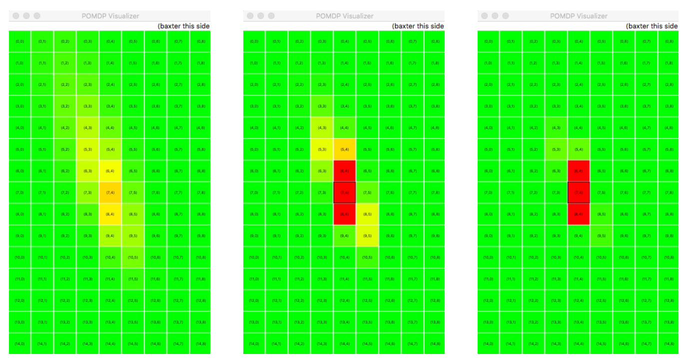

Fig. 1. Heatmap Progression

**Robot Emotions** We also use facial emotion to give feedback to the human collaborator. Specifically, we use varying degrees of confusion to indicate when the human should provide more information, and when the robot is becoming confident enough to act. We calculate our degree of confusion using the entropy of the continuously updating `POMDP` belief state. The standard equation for entropy, $H$, is:

$$H(b) = -\sum_{i}^{n} P(b(i)). \log P(b(i)) \tag{5}$$

where $n$ is the total number of states in the belief state, i.e., grid cells in $\mathcal{G}$.

Fig. 2. Robotic Emotion

## IV. Evaluation

We conducted a pilot study with three subjects and three configurations of our framework. Our objective was to test our hypotheses that (a) a robot supporting language and gesture would outperform one supporting only gesture, and (b) a robot offering backchannel feedback would outperform one that does not, in the task of object placement.

### A. Framework Configurations

The set of configurations of the framework used in various trials are available in Table I.

TABLE I
Framework Configurations

| Code | Configuration |
|---|---|
| `G` | Gesture only |
| `GYR` | `G` + Yes/No Speech + Referring Speech |
| `GYR-HF` | `GYR` + Heatmap + Facial Expressions |

### B. Procedure

**Trial** Each trial was performed with a Rethink Robotics Baxter Robot set up in a tabletop environment. A small test object was placed in Baxter's right hand. A piece of tape was placed on the table to indicate the target placement location to each subject (completely unknown to Baxter).

**Subject** Each subject was instructed to stand in front of the table and collaborate with Baxter on placing the held object in the specified target location. They were informed that Baxter might ask them yes/no questions. They were given a microphone to record their speech, and were informed that a Microsoft Kinect was tracking their movements. They were asked to communicate in the most natural way possible, given the capabilities of the particular configuration being used.

**Experiment** Each experiment involved a subject running at least four trials for each configuration of the framework in the table I, giving a total of at least 12 trials per subject.

**Metrics** *Error*, representing the distance between the target and actual placement; *Time*, representing the time to completion of the `POMDP`; and *Steps*, representing the number of actions planned and taken in the `POMDP`.

### C. Pilot Study Results

Table II gives the mean-value metrics for the configurations used. We note here that the average placement distance error drops by almost $50\%$ from configuration `G` to configuration `GYR` (*hyp. (a)*) and drops by a further $20\%$ from configuration `GYR` to configuration `GYR-HF` (*hyp. (b)*).

The results of our pilot study have given us tangible insights into our collaborative framework. They form a good foundation on which to design our full user study, and thereby gain statistically significant evaluations of our work.

TABLE II
Pilot Study Configuration-Wise Metrics

| Configuration | Error (cm) | Time (s) | Steps (#) |
|---|---|---|---|
| `G` | 14.19 | 38.25 | 23.375 |
| `GYR` | 7.51 | 53.38 | 23.75 |
| `GYR-HF` | 5.94 | 47.38 | 21.88 |

## V. Conclusion

To meaningfully collaborate with humans, robots must understand how humans collaborate with each other. Collaboration is founded on communication, which starts with language and gesture, but extends far beyond into implicit and explicit feedback, and shared knowledge and experience. This research proposes an approach to the task of collaborative object placement on tabletops: a framework that models discrete space in the environment, performs high-level reasoning for estimation of the human's objective using a `POMDP`, utilizes speech and gesture for direct interaction, and provides continuous feedback to its human collaborator. Our results, though preliminary, suggest that each of these components contribute to a more accurate and meaningful collaboration between human and robot.


## Acknowledgments

This extended abstract is based on research completed under Stefanie Tellex in the Humans To Robots Laboratory at Brown University.


## References


[1] Herbert H. Clark and Meredyth A. Krych. Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, (50):62–81, 2004.

[2] Miles Eldon, David Whitney, and Stefanie Tellex. Intrepreting multimodal referring expressions in real time. *International Conference on Robotics and Automation*, 2016.

[3] Leslie P. Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2), 1998.

[4] Michael Kearns, Yishai Mansour, and Andrew Y. Ng. A sparse sampling algorithm for near-optimal planning in large markov decision processes. *Machine Learning*, 49: 193–208, 2002.

[5] James MacGlashan. *Brown UMBC Reinforcement Learning and Planning Library*, 2015. burlap.cs.brown.edu.

[6] Cynthia Matuszek, Liefeng Bo, Luke Zettlemoyer, and Dieter Fox. Learning from unscripted deictic gesture and language for human-robot interactions. 2014.

[7] Emily Wu, Yuxin Han, David Whitney, John Oberlin, James MacGlashan, and Stefanie Tellex. Robotic social feedback for object specification. *AAAI: Artificial Intelligence for Human-Robot Interaction*, 2015.