Modular Inverse Reinforcement Learning on Human Motion

Shun Zhang, Matthew Tong, Mary Hayhoe, Dana Ballard
Department of Computer Science, Center for Perceptual Systems
University of Texas at Austin

Introduction

- Humans are able to learn and carry out very complex tasks involving many different objectives, while most reinforcement learning algorithms suffer from the curse of dimensionality.
- One promising possibility is that the complex task can be broken down into **sub-tasks** that are each learned separately.
- This **decomposition** allows the behavior in the complex task to be chosen based on the value of a weighted sum of individual sub-tasks.
- Our experimental analysis shows that the modular reinforcement can be a good model of predicting **human subjects'** sub-task priorities in a way that explains their behavioral choices.
- We use the **Modular Inverse Reinforcement Learning** approach to analyze human subjects' behaviors.

Modular Inverse Reinforcement Learning

- **Factored MDP.** Define $S = S_1 \times S_2 \times \cdots \times S_m$, where S_i is a state component. We expect the state components are less correlated. The transition function can be defined as a Dynamic Bayesian Network between two time frames of state components.
- **Q-Decomposition.** We further decompose the Q function. Assume that the global Q function is a weighted sum of all Q_i, where Q_i is the Q function for S_i, the i-th state component.

$$Q(s, a) = \sum_i w_i Q_i(s_i, a)$$

where w_i is a weight scalar. $w_i \geq 0, \sum_i w_i = 1$. Now the original MDP is **modularized** into sub-MDPs.

$$MDP_i =< S_i, A, T, R_i, \gamma >$$

- **Modular Inverse Reinforcement Learning.** The objective is to recover the weights given observed samples, $(s^{(i)}, a^{(i)})$, and sub-MDPs. We maximize the likelihood of observing such samples,

$$\max_w \prod_t \frac{e^{\eta Q(s^{(t)}, a^{(t)})}}{\sum_b e^{\eta Q(s^{(t)}, b)}}$$

where $s^{(t)}$ is the state at time t, and $a^{(t)}$ is the action at time t, which are both from samples. $Q(s, a) = \sum_i w_i Q_i(s_i, a)$, as defined before. η is a hyperparameter that determines the consistency of human's behavior.

Multi-objective Domain

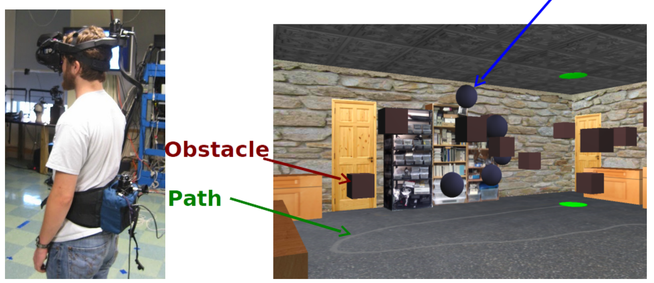
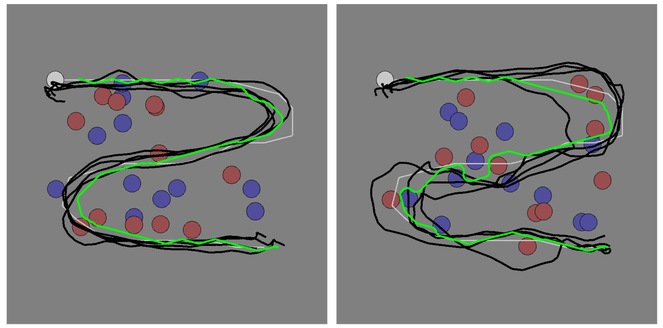


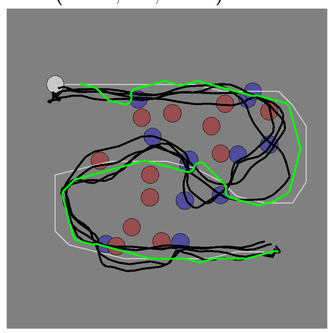
Figure : (Left) A human subject with a head mounted display (HMD) and trackers for the eye, head, and body. (Right) The environment the human can see through the HMD. The red cubes represent obstacles. The blue balls represent targets. There is also a gray path on the ground that the human subject can follow.

Experiments

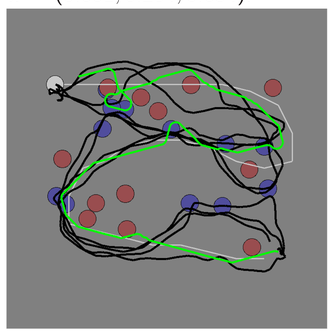


(a) Path module only, $w = (0.039, 0.0, 0.960)$.

(b) Obstacle + Path, $w = (0.081, 0.264, 0.654)$.



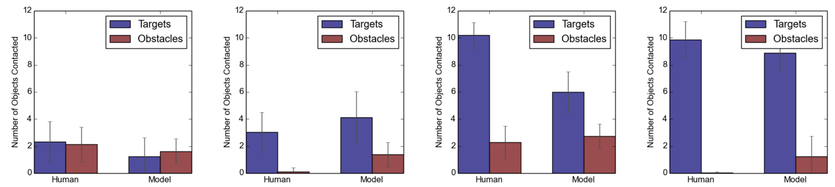
(c) Target + Path, $w = (0.254, 0.089, 0.655)$.



(d) Target + Obstacle + Path, $w = (0.215, 0.414, 0.369)$.

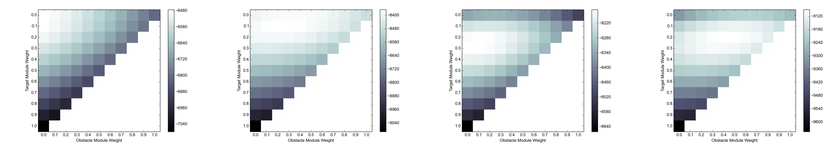
Figure : The trajectories of humans and the agent in the four tasks. Targets are blue and obstacles are red. The black lines are trajectories of human subjects, and the green lines are trajectories of the learning agent by using the optimum weights, w, derived from modular inverse reinforcement learning. Weights for each task are given as (target, obstacle, path).

Experiments



(a) Path module only. (b) Obstacle + Path. (c) Target + Path. (d) All modules.

Figure : Number of targets hit and number of obstacles hit of the human subjects and the agent.



(a) Path module only. (b) Obstacle + Path. (c) Target + Path. (d) All modules.

Figure : Heatmaps of the log of the likelihood for different weights for the four tasks, respectively. The white zones indicate higher probabilities. The vertical axis is the weight of the target module. The horizontal axis is the weight of the obstacle module.

Conclusion

- We analyzed human behavior using inverse modular reinforcement learning.
- The experimental results show that modular reinforcement learning can explain human behavior well, even though the performance of the agent is currently inferior to human subjects'.

Following Work

- Learning weights (or rewards) and discounters of sub-MDPs simultaneously.
- Testing in gridworld domains, with hundreds of sub-MDPs. We compared Modular IRL with Bayesian IRL in this condition.
- Evaluation by angular differences in policies, and likelihood of trajectories. This is compared with other baseline agents.

More details in our paper to appear in a future conference!

Acknowledgment

- This work is supported by NIH EY05729.