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

- One promising possibility is that the complex task can be broken down into sub-tasks that are each learned separately.

Modular Inverse Reinforcement Learning

- **Reinforcement Learning.**
- **Modular Renforcement Learning.** We assume that the global Q function is a weighted sum of all Q_i, where Q_i is the Q function for i-th module.

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

where w_i is the weight of the i-th sub-MDP. $w_i \geq 0, \sum_i w_i = 1$. s_i denotes the decomposition of s in the i-th module.

- **Modular Inverse Reinforcement Learning.** We maximize the objective function

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

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

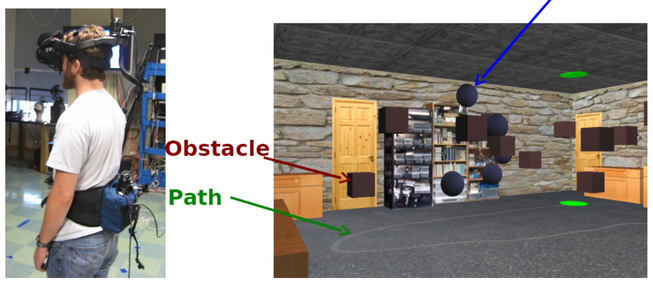
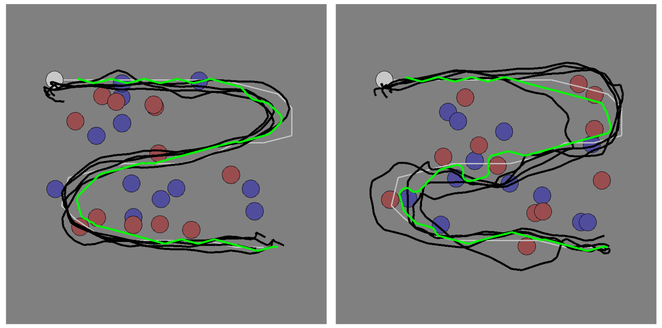
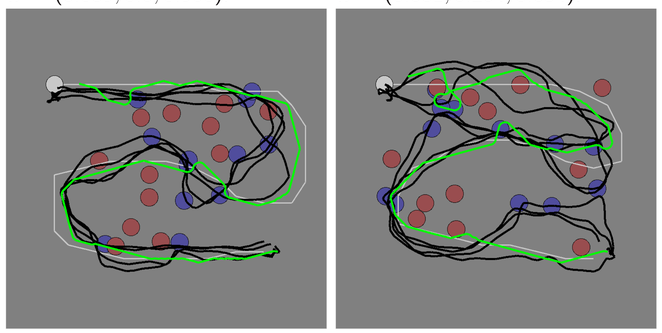


Figure : (Left) A human subject with a head mounted display (HMD) and trackers for the eye, head, and body. (Right) The environment the human can see through the HMD. The red cubes represent obstacles. The blue balls represent targets. There is also a gray path on the ground that the human subject can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

Experiments



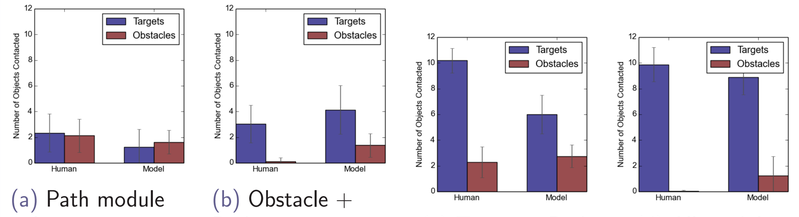
(a) Path module only, $w = (0.039, 0.0, 0.960)$.
(b) Obstacle + Path, $w = (0.081, 0.264, 0.654)$.



(c) Target + Path, $w = (0.254, 0.089, 0.655)$.
(d) Target + Obstacle + Path, $w = (0.215, 0.414, 0.369)$.

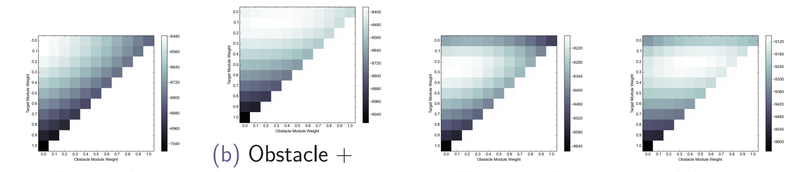
Figure : The trajectories of humans and the agent in the four tasks. Targets are blue and obstacles are red. The black lines are trajectories of human subjects, and the green lines are trajectories of the learning agent by using the optimum weights, w, derived from modular inverse reinforcement learning. 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 following.
(b) Obstacle + Path.
(c) Target + Path.
(d) All modules.

Figure : Heatmaps of the log of the values of Equation **??** for different weights for the four tasks, respectively. The white zones indicate higher probabilities. The weights of all three modules sum to 1, so we only show the weights on the target and the obstacle modules.

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 simutanously.
- Testing in gridworld domains, with hundreds of sub-MDPs. We compared modular IRL and Bayesian IRL in this condition.
- Evaluation by angular differences in policies, and likelihood of trajectories. This is compared with other baseline agents.

Watch our follow-up paper in a future conference!

Acknowledgement

- Funding.