Inverse Modular Reinforcement Learning on Human Motion

Shun Zhang
Department of Computer Science
University of Texas at Austin
Austin, TX 78712
menie482@cs.utexas.edu

Matthew Tong
Center for Perceptual Systems
University of Texas at Austin
Austin, TX 78712
mhtong@gmail.com

Mary Hayhoe
Center for Perceptual Systems
University of Texas at Austin
Austin, TX 78712
hayhoe@utexas.edu

Dana Ballard
Department of Computer Science
University of Texas at Austin
Austin, TX 78712
dana@cs.utexas.edu

Abstract

Keywords: enter key words here

Acknowledgements

1 Introduction

Human is able to learn complicated behaviors much faster than machines can do.

We analyze human's behavior of accomplishing various tasks. With the best of our knowledge, there is no work using inverse reinforcement learning to understand human motional behavior in the literature.

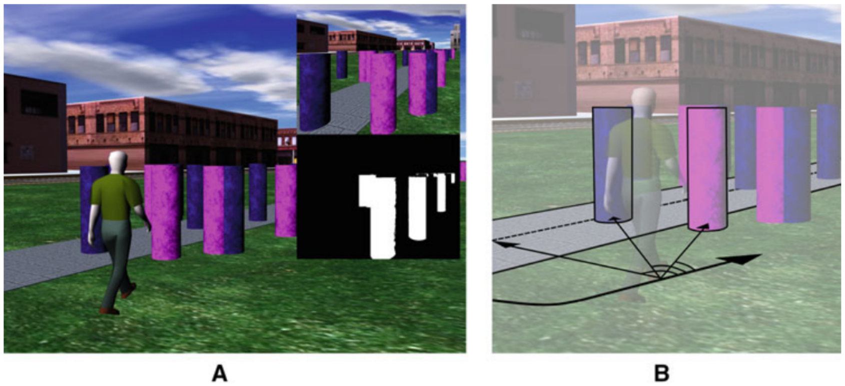

Figure 1: The task of collecting targets, avoiding obstacles and following the path.

Consider the task illustrated in Figure 1 A). The avatar is asked to do three sub-tasks simultaneously — 1) following the path, indicated by the gray line on the ground, 2) getting targets, the blue cylinders, and 3) avoiding obstacles, the pink cylinders [2].

From the reinforcement learning perspective, this task can be decomposed to be three sub-tasks as described above. In Figure 1 B), if the agent knows the distance and angle to an object, he is expected to know the optimal action to avoid or pursue it.

A human has not done this kind of task before can achieve a good performance easily, shown in the experiments described later.

2 Modular Reinforcement Learning

Value functions is decomposed [1].

If a human knows the policies of the sub-tasks, or sub-MDPs, he can accomplish a complicated behavior by combining the sub-MDPs. That is,

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

where Q_i is the Q value of the i-th sub-MDP, w_i is the weight of the i-th sub-MDP. $w_i \geq 0, \sum_i w_i = 1$.

Different weights can yield different performance. Let w_1, w_2, w_3 be weights for the task of target collection, obstacle avoidance, and path following, respectively. Let w be the vector of $(w_i)_1^n$. An agent with $w = (1, 0, 0)$ only collect targets, and one with $w = (0, 0.5, 0.5)$ may avoid the obstacles and follow the path.

3 Experiments

We have four tasks that volunteers are asked to do respectively. Task 1, following the path only, and ignoring other objects. Task 2, following the path, while avoid the obstacles. Task 3, following the path, while attain targets when possible. Task 4, following the path, collecting the targets and avoiding obstacles simutanously.

From a different perspective, can we find a weight vector to best interpret human's behavior? In Figure 2, there are two scenarios. Same as Figure 1, the red circles are obstacles. The blue circles are targets. The gray line is the path. The black lines are trajectories of human . The left figure shows the case where human follows the path and ignores the targets and obstacles. The right figure shows the case that human does three sub-tasks simultaneously. The human data are collected by the Center for Perceptual Systems in University of Texas at Austin.

We make some constraints on our learning agent to make it behavior like humans. It is allowed to go straight ahead, turn left with a small step, or turn right with a small step.

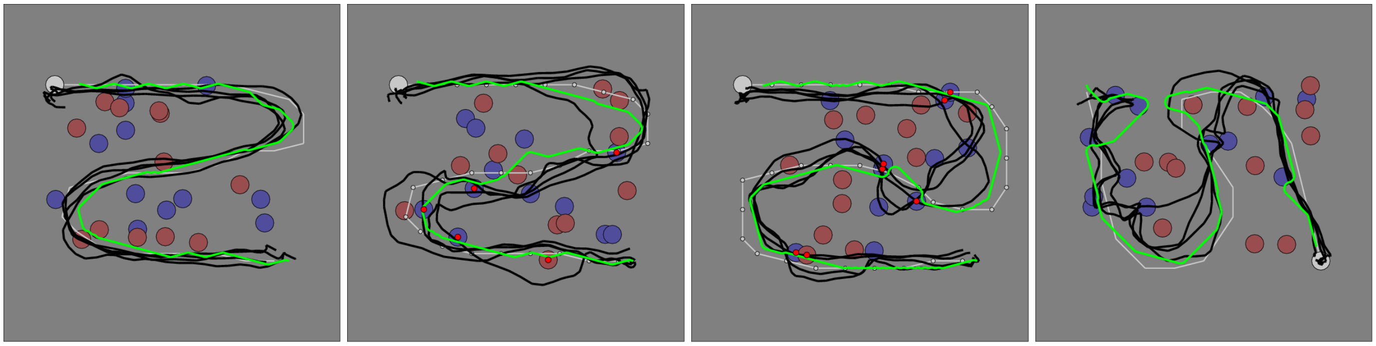

Figure 2:

Now, we assume a learning agent that only knows the three sub-MDPs and the human data. It looks at the human behavior, and finds the weights that can interpret such behavior. Using such weights, the trajectories of our agents are drawn in the green lines. We can tell that in the left figure, the agent puts a large weight on path-following. In the right figure, it puts weights on all sub-MDPs.

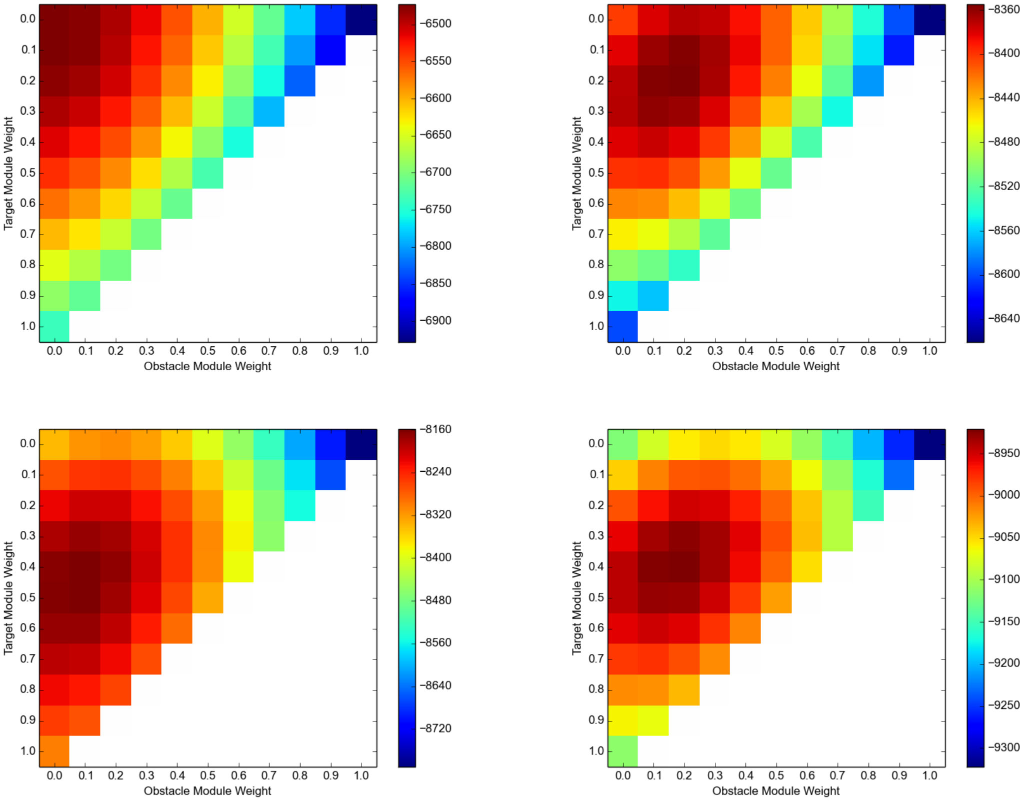

Figure 3: Heatmaps of the objective values of different weights. The red zones indicate higher values.

Average by Task	Num Targs Hit	Num Obst Hit
1	2.34	2.13
2	3.03	0.13
3	10.19	2.28
4	9.88	0.03

Table 1: Number of targets hit and number of obstacles hit for human's behavior.

In Figure 3, we show the object values for different weights.

4 Conclusion and Future Work

Weighted sum of Q function is one way to combine multiple sub-MDPs. We also propose other ways, for example, scheduling between different modules, with only one active at one time. However, when a human tries to collect targets while avoiding obstacles, these two modules should be both on. A scheduling approach may yield frequent oscination between target and obstacle modules.

References

[1] Daphne Koller and Ronald Parr. Computing factored value functions for policies in structured mdps. In *IJCAI*, volume 99, pages 1332–1339, 1999.

[2] Constantin A Rothkopf and Dana H Ballard. Modular inverse reinforcement learning for visuomotor behavior. *Biological cybernetics*, 107(4):477–490, 2013.