Inverse Reinforcement Learning of Interaction Dynamics from Demonstrations

Mostafa Hussein, Momotaz Begum, Marek Petrik
{mhussein, mbegum, mpetrik}@cs.unh.edu

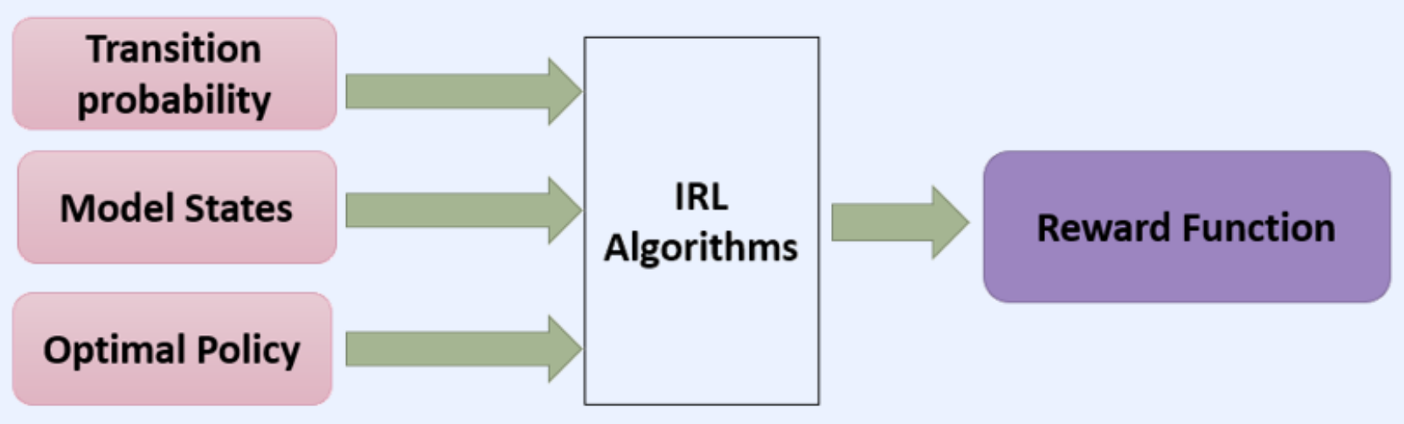
University of New Hampshire

CARL Cognitive Assistive Robotics Lab

Introduction

- In learning from demonstrations, high-level sequential tasks are often modeled as POMDPs. It is important to know the POMDP reward function to learn the policy of the demonstrator.
- IRL for POMDPs is an ill-posed problem. We propose an alternative approach for learning the reward function of a POMDP model through reducing the POMDP to an MDP.
- We perform extensive experiments to show that the reward learned using our proposed framework generates policies that are better than policies generated by the state-of-the-art POMDP-IRL algorithms.

Background



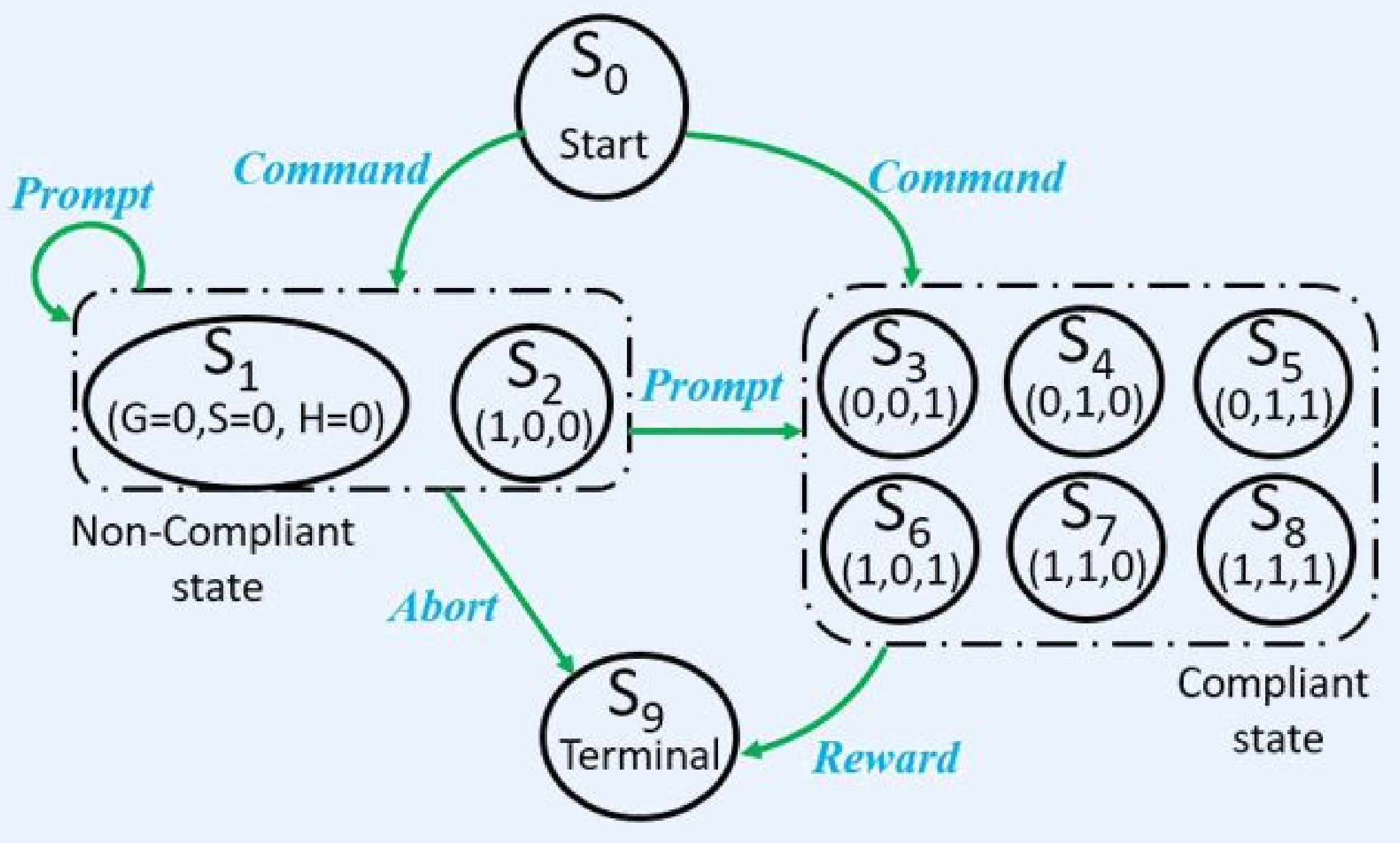
Inverse reinforcement learning model

- Existing inverse reinforcement learning (IRL) methods for POMDPs are computationally very expensive and the problem is not well understood. In comparison, IRL algorithms for MDP are well defined and computationally efficient.

Proposed Approaches

The core idea is to reduce the POMDP to an MDP and extract the reward function using an efficient IRL algorithm for MDPs. Then the reward function is employed to generate policies from the original POMDP.

A- **Naive Reduction:** The main idea of naive reduction is to map each possible observation to one MDP state, thereby eliminating the uncertainty with state estimation.

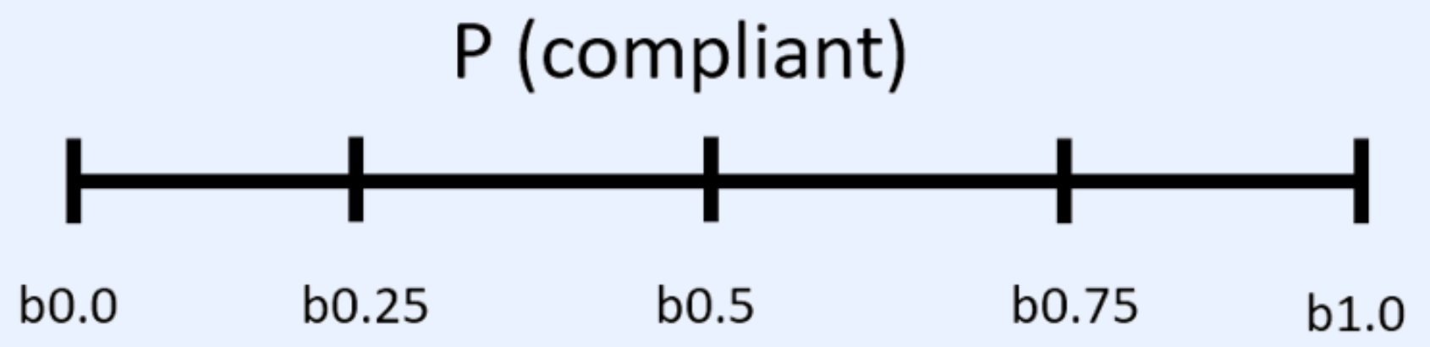


Naive MDP model: 0 and 1 indicate the presence and absence of a specific observation, respectively.

B- **Discretization:** Discretization is where we discretized the POMDP belief state to a pre-defined number n of belief segments. Each segment represents one state of an MDP.

The transition function ($\hat{T} = P(b''|b,a)$) is calculated as follows:

$$P(b''|b,a) = \sum_{o' \epsilon O} P(o'|a,b) \cdot \mathbf{1}_{b_n \in \text{argmin}_{\tilde{b}} \|b''-b'\|_1}$$



The Discretized belief space

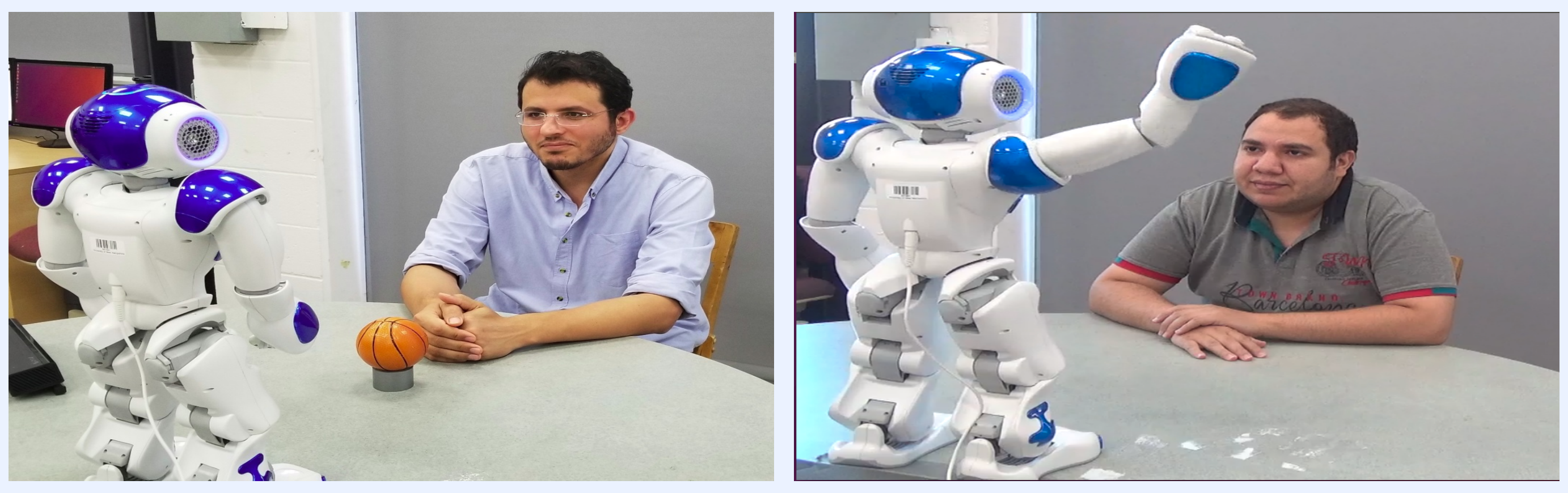
Learning Behavioral Intervention

- We learned two applied behavior analysis (ABA)-based educational interventions from demonstrations: social greetings (SG) and object-naming (ON)
- ABA is proven to be effective to teach children with autism.
- ABA interventions follow a rigid structure: Command–> prompt (if no response) –> reward (if positive response)–> End session

We learned this structure (i.e. the policy of the demonstrator) through learning the reward function using our proposed framework.

Experiments

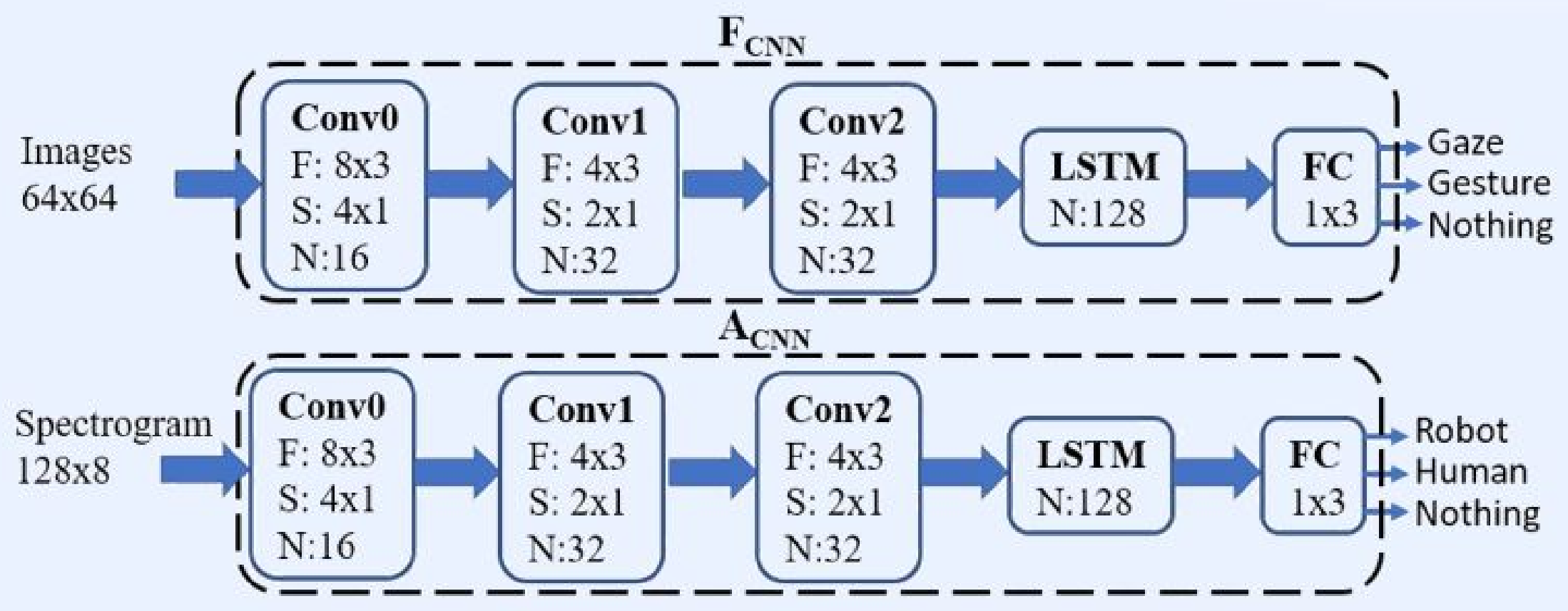
- During the user studies, we restricted the maximum number of prompt before executing a terminal action to one for the SG intervention and to five for the ON.
- Six students without autism participated in the study. Each participant completed 18 interactions with the tele-operated robot.



User interaction with the robot

Perception using CNN

- We used a CNN-based framework to identify the presences of any or more of the audio-visual cues denoting compliance non-compliance of the participants.
- There are two separate CNNs in this framework: F_{CNN}, is trained to detect gaze and hand gesture and, A_{CNN}, is trained to process verbal response.
- We trained both networks using 139 videos from our training dataset and evaluated the model's accuracy on a set of 50 videos. The accuracy was 98.4% for the A_{CNN} and 92.6% for the F_{CNN}.



The CNN-based framework for observation processing. F: filter dimension, S: Stride, N: the number of filter

Results

Social greeting: Accuracy of different reward functions

Obs.	DP	Witness	H.C.R	Discretized	Simplified
1	87.5%	87.5%	87.5%	87.5%	100%
2	87.5%	91.6%	95.8%	100%	100%
Acc.	**87.5%**	**89.6%**	**91.6%**	**93.75%**	**100%**

Object naming: Accuracy of different reward functions

Obs.	DP	Witness	H.C.R	Discretized	Simplified
1	50%	50%	50%	50%	100%
2	100%	75%	100%	100%	100%
3	100%	75%	100%	100%	100%
4	100%	75%	100%	100%	75%
5	50%	50%	75%	75%	75%
Acc.	**80%**	**65%**	**85%**	**85%**	**90%**

Conclusion

- The proposed framework offers a simple yet elegant way to use POMDP models to learn high-level sequential tasks from demonstration and outperforms the policies generated using existing POMDP-IRL algorithms.
- Through a series of experiments with two real-world HRI tasks, we show that the POMDP policies generated using the generated reward functions accurately mimic a demonstrator's policies.

Contact Information

- **Email:** mhussein@cs.unh.edu
- **WebSite:** http://carl.cs.unh.edu