ARL-TN-1085 ● Sep 2021

Inverse Reinforcement Learning with High-Level Task Information (Year 1)

by Craig Lennon

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer’s or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TN-1085 ● SEP 2021

Inverse Reinforcement Learning with High-Level Task Information (Year 1)

Craig Lennon
Computational and Information Sciences Directorate,
DEVCOM Army Research Laboratory

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)* September 2021	**2. REPORT TYPE** Technical Note	**3. DATES COVERED (From - To)** 06/10/2020–09/02/2021

4. TITLE AND SUBTITLE Inverse Reinforcement Learning with High-Level Task Information (Year 1)	**5a. CONTRACT NUMBER** 76759-VT-ARL
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Craig Lennon	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) DEVCOM Army Research Laboratory ATTN: FCDD-RLC-IS Aberdeen Proving Ground, MD 21005	**8. PERFORMING ORGANIZATION REPORT NUMBER** ARL-TN-1085
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	**10. SPONSOR/MONITOR'S ACRONYM(S)**
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This report describes research to discover means by which future Army robotic systems could learn new behaviors from human teammates while still verifiably meeting system and mission specifications. Within the first year, research was conducted on learning jointly from human demonstrations and specifications given in linear temporal logic, under conditions of partial observability. The research demonstrated success within a simple grid world environment. Research is ongoing to extend this progress to more complicated environments, such as that of the US Army Combat Capabilities Development Command Army Research Laboratory’s autonomy stack Unity environments.

15. SUBJECT TERMS

autonomy, verification, linear temporal logic, robots, reinforcement learning

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Craig Lennon
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	UU	15	**19b. TELEPHONE NUMBER (Include area code)** (410) 278-9886

Contents

List of Figures

1. Introduction

The goal of the cooperative research described in this report is to discover means by which future Army robotic systems could learn new behaviors from human teammates while still verifiably meeting system and mission specifications. For example, a robotic ground vehicle must be able to prefer terrain with certain qualities (e.g., concealment or terrain type) based on the instructions of its human teammates. In instances in which the system's understanding of terrain features is not easily explainable to humans, demonstration provides a means by which systems can learn to prefer appropriate types of terrain. Demonstrations, however, are unlikely to cover terrain that is suboptimal, and thus fail to distinguish between less preferred terrain and dangerous terrain. The latter could be identified by the system designer or human supervisor and specified as side information to be introduced into the learning process. Additionally, side information could inform the learning process by providing context for demonstrations, for example, by indicating that the system should exhibit one type of behavior within a secured area and a different behavior when within range of enemy artillery.

Within this report, Section 2 describes the state of research in this area prior to the Cooperative Agreement (CA) and what was done within the first year, while Section 3 describes the planned path forward for the CA.

2. Year 1 Research

At the initiation of this CA, research into using human demonstrations to inform rewards (without side information) had been conducted by researchers at the US Army Combat Capabilities Development Command Army Research Laboratory (Wigness et al. 2018). Within the framework of this method (shown in Fig. 1), humans provided example trajectories, which determined the weights corresponding to human-selected terrain features via maximum entropy inverse reinforcement learning (IRL). This combination of weights and reward features then guided the robot's planning through a trajectory planning cost map.

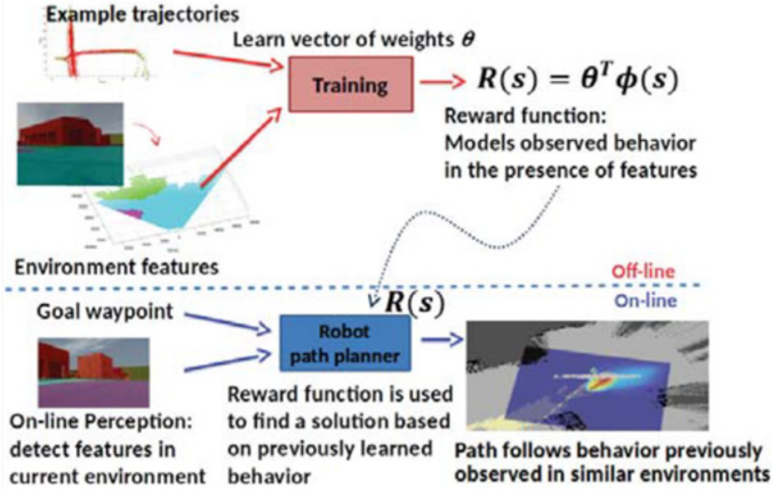
Example trajectories
Learn vector of weights θ
Training
$R(s) = \theta^T \phi(s)$
Reward function:
Models observed behavior
in the presence of features
Environment features
Off-line
On-line
$R(s)$
Goal waypoint
Robot path planner
On-line Perception: detect features in current environment
Reward function is used to find a solution based on previously learned behavior
Path follows behavior previously observed in similar environments

Fig. 1 The learning process followed in Wigness et al. (2018)

This process had been demonstrated both on a physical robot and within the DEVCOM Army Research Laboratory autonomy stack.

Separately, the process of IRL with side information had been demonstrated in Wen et al. (2017). In this research, the learning agent is given a set of specifications, for example, start in the blue cell, proceed to the yellow cell after passing through both green cells and do not contact a red cell, as depicted in Fig. 2. These specifications (referred to as side information) are encoded in linear temporal logic. In addition to this side information, demonstrations are provided to indicate user preferences on how navigation should take place (e.g., following a shortest path or other desired property). Learning is conducted jointly over the demonstrations and the specifications as maximum entropy IRL with a penalty for violating the specifications. The result of the learning is a Markov decision process–based policy that provides for optimal navigation within the demonstration environment. Within any cell of the environment, the policy directs the agent as to which direction it should go to conform to both the side information and demonstrations. Crucially, the policy is environment specific and makes no adaptations to different environments. Once trained on a given grid world, the policy is only applicable to that grid world environment.

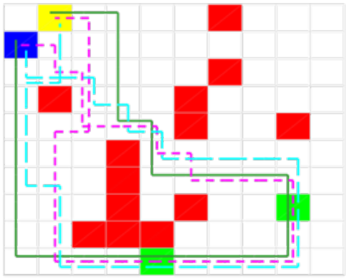

Fig. 2 An example environment for IRL with side information in Wen et al. (2017)

Also noteworthy is the size of the environment, a 10 × 10 grid world, which is substantially smaller that a representative environment from the DEVCOM ARL autonomy stack. For example, the Camp Lejeune environment in the autonomy stack might be decomposed as a 250 × 250 grid cell environment.

With year 1, three parallel, though coordinated, activities took place:

a) Students gained understanding of the DEVCOM ARL autonomy stack, with which they had previously not worked, and applied the methods of Wen et al. (2017) within that environment.

b) We conducted research into how to handle uncertainty, in the form of partial observability, inherent in robotic applications of learning.

c) Research was completed in the use of convolutional neural networks to model rewards for learning of the type in Wen et al. (2017). This research was demonstrated outside of the DEVCOM ARL autonomy stack, within a grid world environment.

With respect to (a), there was considerable time invested in the first year in understanding the DEVCOM ARL autonomy stack and attempting to apply the algorithm from Wen et al. (2017) to the map created during robot navigation. In particular, in the DEVCOM ARL autonomy stack, the robot creates a map as it navigates and labels terrain by type based on a semantic segmentation of that terrain via its perceptual system. This means there is no canonical grid world of the type depicted in Fig. 2. We addressed this challenge by having the robot navigate throughout a region of interest in the Camp Lejeune environment, and then extracting a grid cell representation of the semantically labeled terrain types within that environment.

This led to a second challenge, which is that the labels of terrain types may change as the robot moves through the environment, which is being addressed by research thrust (b). We were able to integrate the algorithms from Wen et al. (2017) into the DEVCOM ARL autonomy stack. Within the Camp Lejeune environment, we could train a robot, via demonstrations, to prefer certain terrain types while avoiding no-go areas specified with side information (Fig. 3). However, behavior exhibited through the demonstrations was not reliable, leading to questions as to whether this lack of reliability was due to uncertainty or terrain type, or due to a more fundamental lack of transferability of learning between different environments or different regions within the same environment.

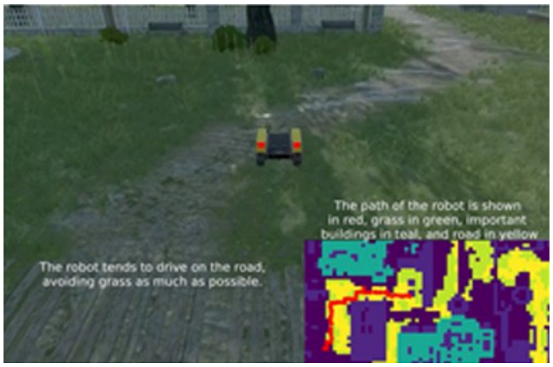
The path of the robot is shown in red, grass in green, important buildings in teal, and road in yellow
The robot tends to drive on the road, avoiding grass as much as possible.

Fig. 3 A robot learns via demonstrations to prefer road to grass while respecting side information directing it to avoid certain areas

The research thrust in (b) consists of reformulating the problem from Wen et al. (2017) as the problem of finding an optimal policy for a partially observable Markov decision process (POMDP), a framework that allows for uncertainty of observations or knowledge. This research resulted in Djeumou et al. (2021), in which the POMDP framework is successfully applied to environments simpler than the DEVCOM ARL autonomy stack Camp Lejeune environment, for example, the maze environment depicted in Fig. 4. This environment is, obviously, substantially smaller than Camp Lejeune, but serves as an accepted benchmark for POMDP because in the maze the agent can only sense its neighboring walls. Thus the agent does not maintain a global map of the environment and has limited memory of what it has sensed within its environment. The idea is that an agent that can plan under such uncertainty could better cope with changing labels in a partially mapped environment, since complete environmental knowledge is not required for formulating a policy. Research efforts to extend the results of Djeumou et al. (2021) to the autonomy stack Camp Lejeune environment are ongoing in year 2.

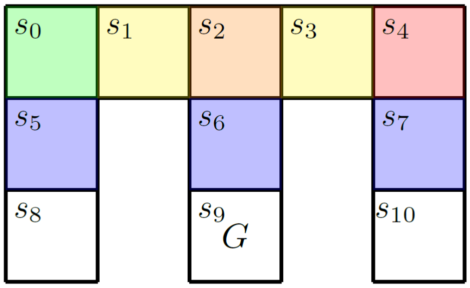

Fig. 4 A maze environment; an agent in it can only sense neighboring walls

With respect to research thrust (c), research was conducted in Memarian et al. (2020) into using a convolutional neural network (CNN) to learn reward features, while a discrete automata structure was learned as a representation for task structure. Such a framework allows an agent to simultaneously learn features upon which rewards are based and task structures. This research was tested in a 14 × 14 cell grid world, which was smaller and simpler than the Camp Lejeune environment. The combination of CNN and automata did outperform alternative IRL methods, but required between 2000–10000 iterations (examples) to jointly learn features and task structure. This is a number prohibitively large for implementation in the DEVCOM ARL autonomy stack. Such an investment in training might be worthwhile if training in one environment was readily transferable to other environments, but at present, this transferability is uncertain. Consequently, we have stopped research into using deep neural networks as representations for reward features and redirected efforts toward understanding how learning transfers between environments, which we will pursue in year 2.

In summary, in year 1, we applied research conducted within a grid world environment to the larger and more complicated Camp Lejeune environment through the DEVCOM ARL autonomy stack. We researched the possibility of using CNNs to learn reward features, but found the number of iterations needed for training to be prohibitively large for the autonomy stack environment. Faced with the challenge of uncertain perception, we discovered an alternative formulation that incorporates uncertainty and demonstrated its value within a simple environment.

3. Planned Year 2 Research

In year 2, our goal is to understand how well IRL with side information transfers between environments and whether it can reliably function as part of the DEVCOM ARL autonomy stack within the Camp Lejeune environment. Toward that end, we will integrate the POMDP oriented research of Djeumou et al. (2021) into the DEVCOM ARL autonomy stack and see how the POMDP formulation of the problem improves reliability of performance. We will also explore transfer between simpler grid world type environments, within which we can modify the structure more readily than we can modify the Camp Lejeune environment. Finally, we will examine transfer of learning when training and testing are done within different regions of the Camp Lejeune environment, and, if behaviors do transfer within the Camp Lejeune environment, we will examine transfer between different Unity environments.

4. Summary and Conclusions

In year 1, we applied research conducted within a grid world environment to the larger and more complicated Camp Lejeune environment through the DEVCOM ARL autonomy stack, we discovered an alternative formulation that incorporates uncertainty and demonstrated its value within a simple environment, and we researched the possibility of using CNNs to learn reward features. We concluded that the number of iterations needed for training CNN to learn reward features was prohibitively large for the autonomy stack environment. We also concluded that transfer of learning was a challenging problem with our current methods, and that in year 2 we need to better understand learning transfer between environments and find modifications of the current method to address this challenge.

5. References

Djeumou F, Cubuktepe F, Lennon C, Topcu U. Task-guided inverse reinforcement learning under partial information. arXiv; 2021 May 28. https://arxiv.org/abs/2105.14073.

Memarian F, Xu Z, Wu B, Wen M, Topcu U. Active task-inference-guided deep inverse reinforcement learning. Proceedings of the IEEE Conference on Decision and Control; 2020.

Wen M, Papusha I, Topcu U. Learning from demonstrations with high-level side information. Proceedings of the 26th International Joint Conference on Artificial Intelligence; 2017.

Wigness M, Rogers J, Navarro-Serment L. Robot navigation from human demonstration: learning control behaviors. Proceedings of the IEEE International Conference on Robotics and Automation. 2018 May;1150–1157.

List of Symbols, Abbreviations, and Acronyms

ARL	Army Research Laboratory
CA	Cooperative Agreement
CNN	convolutional neural network
DEVCOM	US Army Combat Capabilities Development Command
IRL	inverse reinforcement learning
POMDP	partially observable Markov decision process

1 DEFENSE TECHNICAL
(PDF) INFORMATION CTR
DTIC OCA

1 DEVCOM ARL
(PDF) FCDD RLD DCI
TECH LIB

1 DEVCOM ARL
(PDF) FCDD RLC IS
C LENNON