Apprenticeship Bootstrapping: Inverse Reinforcement Learning in a Multi-Skill UAV-UGV Coordination Task

Robotics Track

Hung The Nguyen
School of Engineering & IT, UNSW-Canberra
Canberra, Australia
hung.nguyen@student.adfa.edu.au

Matthew Garratt
School of Engineering & IT, UNSW-Canberra
Canberra, Australia
m.garratt@adfa.edu.au

Lam Thu Bui
Department of Information Technology, Le Quy Don Technical University
Hanoi, Vietnam
lam.bui07@gmail.com

Hussein Abbass
School of Engineering & IT, UNSW-Canberra
Canberra, Australia
h.abbass@adfa.edu.au

ABSTRACT

Apprenticeship learning enables learning from human demonstrations performed on tasks. However, acquiring demonstrations in complex tasks where a human expert is not available can be a challenge. In this paper, we propose a new learning algorithm, called Apprenticeship bootstrapping via Inverse Reinforcement Learning using Deep Q-learning (ABS via IRL-DQN), to learn a complex task through using demonstrations performed on primitive sub-tasks. The algorithm is evaluated on an aerial and ground coordination scenario, where an Unmanned Aerial Vehicle (UAV) is required to maintain three Unmanned Ground Vehicles (UGVs) within a field of view of the UAV 's camera (FoV). The results show that performance of our proposed algorithm is comparable to that of a human, and competitive to the original IRL using expert demonstrations performed on the composite task.

KEYWORDS

Inverse Reinforcement Learning; Apprenticeship Learning; Deep Q-learning; UAVs; UGVs; Ground-Air Interaction

ACM Reference Format:
Hung The Nguyen, Matthew Garratt, Lam Thu Bui, and Hussein Abbass. 2018. Apprenticeship Bootstrapping: Inverse Reinforcement Learning in a Multi-Skill UAV-UGV Coordination Task. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. André, S. Koenig (eds.), Stockholm, Sweden, July 2018,* IFAAMAS, 4 pages.

1 INTRODUCTION

In a recent survey [5] about apprenticeship learning, the main challenges come from the problem of how to transfer human skills to agents or robots through demonstrations. Some recent research focuses on solving the difficulties in collecting expert demonstrations. For example, Knox et al. [6] introduces a shaping technique in which an agent is interactively trained through positive or negative signals from an expert. In another approach to the problem of low-level controls, Faust et al. [3] aims to find the suitable preference-balancing task that supports human to control a quadrotor in real-world conditions such as strong wind. In these research, human skill-levels are assumed to be available. However, when designing a new task for an autonomous system, particularly in complex tasks, there is no guarantee that a human expert is available to to create a dataset for the apprenticeship learning.

Abbeel et.al [2] proposed apprenticeship learning via inverse reinforcement learning (AL via IRL) to produce an approximate policy that is close enough to the observed one. AL via IRL was successfully applied for developing a controller of helicopter aerobatics [1]. Recently, the Deep Q-network (DQN) [7] algorithm demonstrated that DQN agents are able to successfully learn from complex state spaces, with the performance of DQN achieving a professional skill-level when tested on 49 Atari games.

In this paper, we proposes a new apprenticeship learning algorithm, called "Apprenticeship Bootstrapping via Inverse Reinforcement Learning Using Deep Q-learning" (ABS-IRL-DQN), to learn a composite task using human demonstrations on sub-tasks.

2 THE PROPOSED ALGORITHM

The main idea of ABS-IRL-DQN is that a complex task is decomposed into sub-tasks that require less skilled humans, so that we can bootstrap the higher skills from these building blocks. The sub-tasks represent a decomposition of the action space. Not all actions are needed for a sub-task. It may also involve a decomposition of the state space since sub-tasks are associated with simpler contexts that represent partial representations of the original context.

Demonstrations on these sub-tasks are used to approximate the reward signal needed for a subset of the state vector. These incomplete approximations are then fused through an expectation function to approximate the overall reward function for Deep Q-learning to discover a policy over the composite task. The primary assumption we make is that there exists a human who can perform the sub-tasks. Each sub-task encodes sub-skills for the composite task. However, the fusion of these sub-skills is left to the RL agent to learn how to switch and combine them.

Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. André, S. Koenig (eds.), July 2018, Stockholm, Sweden.

It is important to emphasize that in our case, the sub-tasks are orthogonal. Therefore, taking the sum of the feature expectations vector is equivalent to taking the union of the state space. Abbeel et.al [2] explained that averaging the feature expectation for policies is equivalent to calculating the feature expectation of a distribution over policies. However, in our case each sub-task is defined with a different state space and actions; in other words, the sub-spaces are non-commensurable. When we average the policies' feature expectations, the non-commensurable state spaces are morphed into a composite state representing the overall state space needed for conducting the overall task.

3 EXPERIMENTAL RESULTS

Our proposed algorithm is evaluated on a ground-air tracking task, where an unmanned aerial vehicle (UAV) attempts to maintain a mobile group of unmanned ground vehicles (UGVs) within its camera range. This scenario can be seen in our previous research[4].

We used a human to generate the ground truth for training and testing using four maneuvers: Fixed-Altitude, Climb, Descend, and Combined maneuver. The first three maneuvers define the low-level skills required to perform the fourth maneuver; these are: the UAV needs to either track UGVs by moving forward, climbing and descending. In the fourth maneuver, the UGVs move in more complex maneuvers where all three forms of behaviour (lateral tracking, climbing and descending) are used simultaneously.

Two experiments are conducted to assess our proposed algorithm. The first experiment uses the original IRL [2] for the human demonstrations collected in the fourth maneuver. Meanwhile, in the second experiment, ABS-IRL-DQN is used to learn from human demonstrations on the first three maneuvers. To evaluate performance, we calculate the distance between the UAV Center of mass (cx, cy) and the center of UGVs' mass $(UGVx, UGVy)$, as well as the difference between the actual radius ra_i and the ideal radius ra_a. Meanwhile, ra_a and ra_i is the actual and ideal radius of the UGVs' operating circle within the FoV, respectively.

Table 1: Average and Standard Deviations of Errors in All Testing Experiments in the Combined Experiments. Results highlighted in boldface are different from the human's results and the differences are statistically significant at $\alpha = 0.05$.

Experiment ID	Distance Errors	Radius Errors
Human Performance	21.2 ± 13.1	6.1 ± 5.4
Original IRL	33.7 ± 22.5	12.6 ± 9
ABS-IRL-DQN	23.3 ± 13.2	**12.6 ± 6.5**

Table 1 shows that both ABS-IRL-DQN and original IRL agents are able to perform nearly equivalent to the human subject. This is despite the fact that the reward functions used in ABS-IRL-DQN are based on an observed state space that is a subset of the overall state-space used in the original IRL. ABS-IRL-DQN seemed to have favored the distance metric more than the original IRL did.

To better understand the phenotypic differences between the human performance and agents, we visualize the behaviour of the UAV in Figure 1. As expected, the IRL trajectory is qualitatively similar to human trajectory. The ABS-IRL-DQN trajectory, however, has smoother oscillation but with larger magnitude. This smoother oscillation is desirable as too much oscillations generate inefficient flights; possibly outside the performance envelop of the UAV. However, this comes with a cost, where distance error increases. The more the UAV attempts to have a smoother trajectory, the less it is able to quickly adjust to the UGVs and the greater the distance error. The averaging of the feature expectations vector favors the smoother trajectory. It should be possible to control the trade-off between smoothness and distance errors by changing the fusion function.

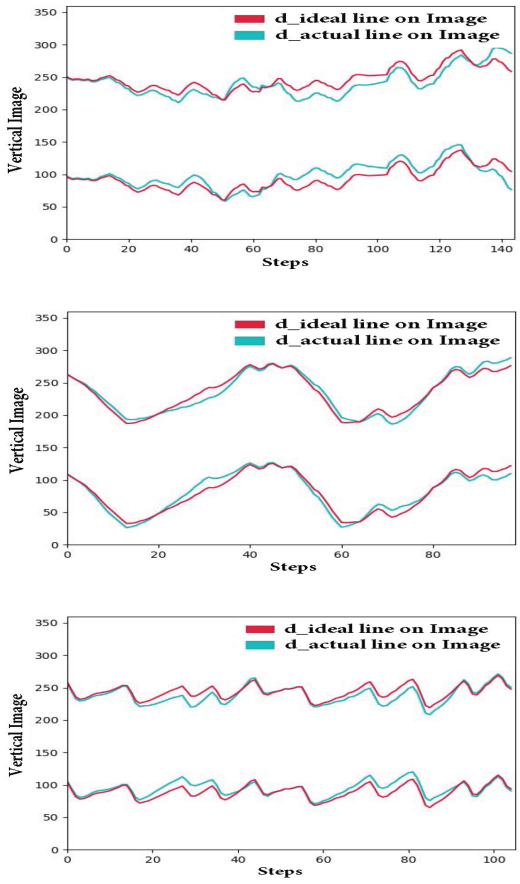

Figure 1: The Ideal and Actual UGVs Circle Trajectories on Vertical Image in the Combined Maneuver. Top-Left: Human Control; Top-Right: ABS-IRL-DQN; Bottom: Original IRL. It is important to note that UGVs do not have identical dynamics in the three scenarios because of stochastic noise.

4 CONCLUSION

The paper shows that the ABS via IRL-DQN agent perform in a comparable manner to humans and are competitive to the agent trained on data collected from humans performing the more complex task.

ACKNOWLEDGEMENT

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA2386-17-1-4054.

REFERENCES

[1] Pieter Abbeel, Adam Coates, and Andrew Y Ng. 2010. Autonomous helicopter aerobatics through apprenticeship learning. *The International Journal of Robotics Research* 29, 13 (2010), 1608–1639.

[2] Pieter Abbeel and Andrew Y Ng. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*. ACM, 1.

[3] Aleksandra Faust, Nick Malone, and Lydia Tapia. 2015. Preference-balancing motion planning under stochastic disturbances. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, 3555–3562.

[4] Nguyen Hung, Garratt Matthew, Bui Lam, and Abbass Hussein. 2017. Supervised Deep Actor Network for Imitation Learning in a Ground-Air UAV-UGVs Coordination Task. In *IEEE Symposium Series on Computational Intelligence (IEEE SSCI 2017)*.

[5] Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. 2017. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)* 50, 2 (2017), 21.

[6] W Bradley Knox and Peter Stone. 2009. Interactively shaping agents via human reinforcement: The TAMER framework. In *Proceedings of the fifth international conference on Knowledge capture*. ACM, 9–16.

[7] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529.