On-the-fly Learning and Monitoring of Partially Observed Navigation Plan

(Extended Abstract)

J. Vince Pulido
University of Virginia
151 Engineer's Way
Charlottesville, VA
j.vincepulido@virginia.edu

MaryAnne Fields
U.S. Army Research Laboratory
4727 Deer Creek Loop
Aberdeen Proving Grounds, MD, USA
mary.a.fields22.civ@mail.mil

Laura Barnes
University of Virginia
151 Engineer's Way
Charlottesville, VA
lbarnes@virginia.edu

ABSTRACT

This work focuses on a robot's task of predicting the navigation intent of human teammates using Inverse Reinforcement Learning. The purpose of this study is to introduce the On-the-fly Maximum Margin Planner (OTF-MMP) method which estimates a predictive navigation model in real-time from the observed actions of a human teammate. We include an experiment to test the predictive ability of the method using simulation.

CCS Concepts

•**Computing methodologies** → *Inverse reinforcement learning;*

Keywords

Inverse Reinforcement Learning, Intent Recognition

1. INTRODUCTION

Increasing desire for robotic applications that situate autonomous robots as human partners in cooperative and tightly interactive tasks beckons rapid advancement in several areas [18, 16, 6]. Although recent advances in sensors and perception methods–measuring the current state of the environment–have improved human-robot interactive capabilities, inferring the plans, strategies, and intentions of agents around robots would enable them to have better decision-making abilities [8, 12, 17]. The ability to recognize plans and goals of the other agents may enable robots to reason about what these agents are doing, why they are doing it, and what they will do next. This fundamental cognitive ability is critical for human-robot interaction because teammate coordination presupposes the ability to understand the motivations of the coordinating participants.

For this work, we focus on the task of understanding the path-finding process and predicting navigation actions of human teammates–a highly unpredictable and ambiguous cognitive process. The purpose of this study is to present an approach to learn a predictive navigation model "on-the-fly" from observation of the human teammate's actions–which we call On-The-Fly Maximum Margin Planner (OTF-MMP). To this end, we propose a novel adaptation of the Maximum Margin Planner (MMP) [15] by adjusting how we encode and calculate parameters.

Previous behavior prediction studies have been verified only on relatively short-term horizons [13, 19, 20, 9, 10]. To improve long-term predictability, "goals" have to be postulated assuming that they are following the "shortest path" [4, 23, 2]. Recent interest in autonomous vehicles brought about developments in pedestrian path prediction for tracking and avoidance [5, 23], and navigation assistance [21, 22, 7]. In a recent published work, Karasev et al. [5] presented a method to predict the long-term motion of pedestrians assuming that agents have predetermined feature biases (e.g. pedestrians prefer sidewalks). This study aims to predict long-term path-finding actions of pedestrians in "one-shot" without previous knowledge of historical paths. Furthermore, unlike previous work, this study avoids restricting human agents to a predetermined feature preference.

We ground this work within the framework of Inverse Reinforcement Learning's (IRL) Maximum Margin Planner [15], which learns policies from observing the actions of a teacher [11, 1, 14, 15]. Assuming that the human agents act in an optimal manner, IRL can be used as a tool for the human agent to teach a robot its path-finding principles.

Humans can navigate to a known location in an infinite amount of ways due to the continuous nature of the real-world. In order to keep the inference computation tractable, we model the navigation problem as a discrete Markov Decision Process (MDP) organized as a grid world with rewards for each state (or cell) the agent enters [3]. We assume that the goal of the agent is to seek sequences of optimal actions that maximize the collected reward. Thus, in order to understand an agent's path-finding strategy, we must be able to infer the reward, $R(s)$, at every state $s \in S$.

While Ratliff et. al. [15] introduced batch learning and online (or incremental) learning MMP methods to estimate the reward function, $R(s)$, this study adapts the MMP method to make inferences on $R(s)$ by observing a single incomplete path (i.e. the agent is still en route to the goal location) and estimating reward functions on-the-fly. For more details on the MMP notation and formulation, See [15].

Before the agent's path-finding process begins, we initialize $\mu(s,a)$ and $\mu^*(s,a)$ as a zero vector of size $(m*n)*|A|$, where $|A|$ is the number of actions per state (e.g. $|A| = 4$

Appears in: *Proc. of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017), S. Das, E. Durfee, K. Larson, M. Winikoff (eds.), May 8–12, 2017, São Paulo, Brazil.*

for a = {*north, south, east, west*}), and m and n are the sizes of the x-axis and y-axis of the grid world, respectively. Note that $(m * n)$ corresponds to the size of the state space S. Let w be initialized as a non-zero vector.

To adapt the MMP in an "on-the-fly" fashion, we present three adjustments to the algorithm (a)-(c):

(a) *Encode the observed trajectory* μ:
Let $\mathcal{O} = \{(s_0, a_o), ..., (s_t, a_t)\}$ be a sequential list of observed visited states $s_i \in S$ and the corresponding observed actions $a_i \in A$, for all $i = \{0, \ldots, t\}$. As the agent performs action a_i in state s_i, we update the sparse vector $\mu(s_i, a_i) = 1$.

(b) *Encode optimal planner* μ^*:
Given the current estimate of w, we run the value iteration algorithm to produce an optimal policy, $\pi : S \mapsto A$, representing the optimal action a^* at each state s_i. After each observed transition, we update $\mu^*(s, a^*)$:

$$\mu^*(s,a) = \begin{cases} 1, & (s_i, a_i) = (s_i, a_i^*),\ \forall (s_i, a_i) \in \mathcal{O} \\ 0, & otherwise \end{cases} \tag{1}$$

(c) *Define Loss function*:
We define a loss function that is stricter than that of [15]. Our loss function $L(\mu^*, \mu)$ is a function that counts the number of dissimilar state-action policies.

Using these three elements (a)-(c), we iteratively solve for w as the partial-path OTF-MMP objective function defined as

$$\begin{aligned} \underset{w}{\text{minimize}} \quad & \frac{1}{2}||w||^2 + \beta\zeta^2 \\ \text{s.t.} \quad & w^T F\mu + \zeta \geq w^T F\mu^* + L(\mu^*, \mu) \end{aligned} \tag{2}$$

2. RESULTS

In this section, we analyze the predictive ability of the OTF-MMP when exposed to different levels of an agent's path-finding noise. This experiment simulates noisy scenarios of an agent traversing a map to a known goal. The features are: *Road, Side Walk, Low Vegetation, Medium Vegetation, High Vegetation, Buildings, Water, Distance.*

For this experiment, we considered two strategies: 1. *Road Strategy*: "least-effort" path on roads (i.e. $w_{Road} = 1$, otherwise $w_i = 0$ for all other weights not corresponding with road). 2. *Concealed Strategy*: difficult but hidden route with high vegetation to remain invisible (i.e. $w_{HighVeg}, w_{MedVeg} = 1$, otherwise $w_i = 0$ for all other weights not corresponding with $HighVeg$ and $MedVeg$). We determine the agent's policy using value iteration to determine the optimal action to take a every state s. Thus, the road and concealed strategy would each have one exact optimal policy.

We simulate an agent's paths imposing 5 levels of obedience levels $P = \{100\%, 95\%, 90\%, 85\%, 80\%\}$. Obedience levels are defined as the probability that the agent would follow the optimal strategy at each state. For each obedience level per strategy, we simulate 30 distinct example paths. Figure 1 shows the test terrain overlaid by three representative paths. The *red* line shows the optimal path using road strategy. The *green* line shows the optimal path using the concealed strategy. The *yellow* line shows an example sub-optimal path using concealed strategy which diverged from the optimal strategy due to errors in transitions.

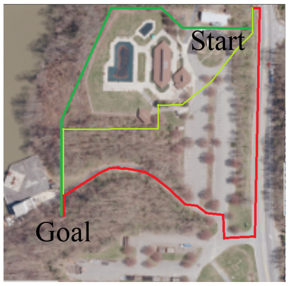
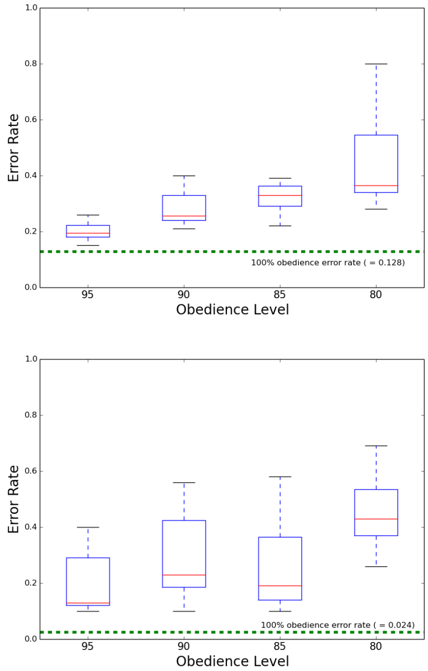


Figure 1: Example trajectories for road strategy (red), and concealed strategy (green–optimal, yellow–sub-optimal).



(a) Error results for **road** strategy.

(b) Error results for **concealed** strategy.

Figure 2

We divide each simulated path into two sets with equal number of transitions: the first half of the path will be used as our training set and the latter half will be used as our test set. We use OTF-MMP to build a model–using the training set–whose predictability will be compared to the test set. We count the number of transitions that were predicted incorrectly and normalized with the number of total transitions observed. Figure 2a and 2b summarizes the results. We see that error rate increases at a faster rate as the obedience level decreases which signifies a deterioration of model reliability. Since the obedience level dictates the proportion of reliable observed action, the OTF-MMP is highly sensitive to the path-finding noise of the human agent.

3. CONCLUSION

In this work, we adapt the maximum margin techniques to the problem of inferring navigation strategies of agents on-the-fly and estimate their optimal policy, π, based on observed actions. Using methods of maximum margin planning, autonomous robots can plan a path that is consistent with its teammate's strategy, evaluate the progress of the mission, and anticipate future team location and vulnerabilities. Ultimately, this method takes the step to provide the robot with information it needs to be a better teammate.

REFERENCES

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

[2] C. L. Baker, R. Saxe, and J. B. Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3):329–349, 2009.

[3] R. Bellman. Dynamic programming and lagrange multipliers. *Proceedings of the National Academy of Sciences of the United States of America*, 42(10):767, 1956.

[4] A. F. Foka and P. E. Trahanias. Predictive autonomous robot navigation. In *Intelligent Robots and Systems, 2002. IEEE/RSJ International Conference on*, volume 1, pages 490–495. IEEE, 2002.

[5] V. Karasev, A. Ayvaci, B. Heisele, and S. Soatto. Intent-aware long-term prediction of pedestrian motion. In *Proceedings of the International Conference on Robotics and Automation (ICRA)(May 2016)*, 2016.

[6] H. Kidokoro, T. Kanda, D. Brscic, and M. Shiomi. Will i bother here?-a robot anticipating its influence on pedestrian walking comfort. In *Human-Robot Interaction (HRI), 2013 8th ACM/IEEE International Conference on*, pages 259–266. IEEE, 2013.

[7] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert. Activity forecasting. In *European Conference on Computer Vision*, pages 201–214. Springer, 2012.

[8] H. S. Koppula and A. Saxena. Anticipating human activities using object affordances for reactive robotic response. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 38(1):14–29, 2016.

[9] M. Kuderer, H. Kretzschmar, C. Sprunk, and W. Burgard. Feature-based prediction of trajectories for socially compliant navigation. In *Robotics: science and systems*. Citeseer, 2012.

[10] M. Monfort, A. Liu, and B. Ziebart. Intent prediction and trajectory forecasting via predictive inverse linear-quadratic regulation. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[11] A. Y. Ng, S. J. Russell, et al. Algorithms for inverse reinforcement learning. pages 663–670, 2000.

[12] B. Ni, G. Wang, and P. Moulin. Rgbd-hudaact: A color-depth video database for human daily activity recognition. In *Consumer Depth Cameras for Computer Vision*, pages 193–208. Springer, 2013.

[13] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th International Conference on Computer Vision*, pages 261–268. IEEE, 2009.

[14] D. Ramachandran and E. Amir. Bayesian inverse reinforcement learning. *Urbana*, 51:61801, 2007.

[15] N. D. Ratliff, J. A. Bagnell, and M. A. Zinkevich. Maximum margin planning. pages 729–736, 2006.

[16] K. W. Strabala, M. K. Lee, A. D. Dragan, J. L. Forlizzi, S. Srinivasa, M. Cakmak, and V. Micelli. Towards seamless human-robot handovers. *Journal of Human-Robot Interaction*, 2(1):112–132, 2013.

[17] J. Sung, C. Ponce, B. Selman, and A. Saxena. Unstructured human activity detection from rgbd images. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 842–849. IEEE, 2012.

[18] G. Trafton, L. Hiatt, A. Harrison, F. Tamborello, S. Khemlani, and A. Schultz. Act-r/e: An embodied cognitive architecture for human-robot interaction. *Journal of Human-Robot Interaction*, 2(1):30–55, 2013.

[19] P. Trautman and A. Krause. Unfreezing the robot: Navigation in dense, interacting crowds. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 797–803. IEEE, 2010.

[20] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg. Who are you with and where are you going? In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1345–1352. IEEE, 2011.

[21] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. pages 1433–1438, 2008.

[22] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey. Human behavior modeling with maximum entropy inverse optimal control. In *AAAI Spring Symposium: Human Behavior Modeling*, page 92, 2009.

[23] B. D. Ziebart, N. Ratliff, G. Gallagher, C. Mertz, K. Peterson, J. A. Bagnell, M. Hebert, A. K. Dey, and S. Srinivasa. Planning-based prediction for pedestrians. In *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, pages 3931–3936. IEEE, 2009.

This research was supported in part by an appointment to Student Research Participation Program at U.S. Army Research Laboratory administered by the Oak Ridge Institute for Science and Education through inter-agency agreement between the U.S. Department of Energy and U.S. ARL.