Interactive Delayed Observation POMDPs for Exploration of Uncertain Environments

Shohei Wakayama and Nisar Ahmed*

I. Introduction

Remote exploration will require combining robotic autonomy with human understanding and supervision. Due to size, weight, computation, communication, and power constraints, it is difficult for remote robot explorers to solve complex decision-making problems onboard and in real time, while also accurately perceiving and reasoning about their environments. It can also be very difficult for human users (particularly those who are not robotics experts) to fully trust and take advantage of remote robotic autonomy if the basis on which autonomy acts is unclear [1]. Against this backdrop, research is underway to improve human-autonomous robotic exploration efficiency and quality by treating human users – presumably, task domain experts who possess better high-level perception capabilities than robots – as secondary sensors, which can provide valuable semantic information to improve robotic decision-making in uncertain environments [5]. In these works, semantic soft data (e.g. "There are some interesting deposits around that rock') are given to the robot via a user-friendly natural language and/or map-based sketch interfaces. Probabilistic models can then be developed to interpret and fuse semantic soft data with the conventional 'hard' sensor data observed by the robot [8], [9], and efficient active semantic sensing strategies based on partially observable Markov decision process (POMDP) planning can be used to query human sensors to mitigate state uncertainties during online plan generation [2], [3], [4].

Note that unlike [11], in which supervisors provide reward labels (with some delay), we consider how supervisors act as sensors that can also provide (imperfect and noisy) information about environment dynamics, obstacle and target of opportunity existence/location, etc., which can be strategically queried and fused with other data via Bayesian reasoning methods. Since small amounts of semantic soft data can yield drastic changes in beliefs, huge gains in decision-making performance can be obtained for POMDP-based active sensing, along with the ability of supervisors to inspect robot state beliefs that drive behavior [3]. However, existing methods for active semantic sensing do not account for significant time delays that arise from extreme distances as well as human factors. Ref. [6] describes how offline POMDP policy approximations such as point-based value iteration (PBVI) can be adapted via state augmentation to account for beliefs in delayed observations. However, this approach leads to a rapid increase in the dimension and complexity of the state belief space, and thus becomes impractical to implement for problems with non-trivial state, action and observation spaces. For example, suppose the mobile robot in Fig.1 is exploring a lunar cave, and relays imagery data to a supervisor stationed off the lunar surface to determine whether there are high value science targets in the area that should be localized for followup study. Since the whole cave cannot be explored and imagery cannot be processed onboard due to limited computation power, the supervisor sends back semantic observations based on the received imagery, after discussion with science experts. The robot can still decide where to navigate on its own before receiving answers, but limited onboard sensing capabilities mean that information from human semantic sensing is also required to detect certain hazards (e.g. negative obstacles, impassable terrain, etc.) along with possible targets of opportunity. As the state and semantic observation spaces can be quite large and dynamic for exploration of uncertain environments [4], the curses of dimensionality and history also make it difficult to rely on offline POMDP solvers.

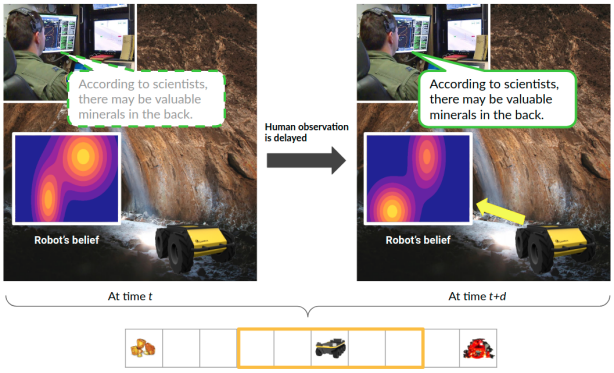

Fig. 1. Operational concept: mobile robot sends data to ask if there is anything of interest in the environment; the supervisor sends delayed semantic observations to the robot; based on the semantic information, the robot updates target/obstacle location beliefs and chooses an action.

In this work, we examine how Partially Observable Monte Carlo Planning (POMCP) [7], an anytime fashion sampling-based online POMDP solver, can be applied to the problem with delayed semantic human observations. Since POMCP has not yet been widely investigated in time delayed observation settings, we present initial simulation results on a simple benchmark problem. Analysis suggests that online planning offers a promising approach to dealing with semantic observation data characterized by time delays and variable accuracy, and that decision-making performance can be more sensitive to human sensor accuracy than time delay in certain cases.

*Smead Aerospace Engineering Sciences Department, 429 UCB, University of Colorado Boulder, Boulder, CO 80309, USA. E-mail: [shohei.wakayama;nisar.ahmed] @ colorado.edu. Shohei Wakayama is supported by Masason Foundation.

II. Technical Approach

Fig. 1 shows a simplified 1D version of the mission concept, where a valuable science 'target' (left) and large obstacle (right) are each at unknown locations along a traversable pathway (grid). The mobile robot (middle) must locate and reach the target cell as quickly as possible, without hitting the obstacle. At each step, the robot obtains image data for cells immediately around it (yellow). A remotely stationed human will receive and analyze the data, and then report relevant semantic science and hazard information to guide planning, according to a predefined semantic observation codebook, e.g. 'Neither target nor obstacle are in view'. For simplicity here, no other sensors are available, and the transmit time is smaller than the (fixed) time $TD \in \mathbb{Z}^{0+}$ the human spends processing each image, creating a one-sided observation receipt delay.

The POMDP tuple ($\mathcal{S}$, $\mathcal{A}$, $\mathcal{T}$, $\mathcal{R}$, Ω, $\mathcal{O}$, γ) for this scenario is as follows. The state $s_t = [s_{r,t}, s_{tar}, s_{obs}]^T \in \mathcal{S}$, where $s_{r,t}$ is the robot's (known) position at time t, s_{tar} is the unknown target position, and $s_{obs} \neq s_{tar}$ is the unknown obstacle position. The action space $\mathcal{A}$ includes $\{Left, Right, Stay\}$; the robot can move to the intended direction with 90% probability, and with 5% probability it ends up executing one of other two actions. For discount $\gamma = 0.9$, a large positive reward $\mathcal{R}$ results when $s_{r,t} = s_{tar}$ and $a = Stay$; a large negative reward results when $s_{r,t} = s_{obs}$ and $a = Stay$. A small positive/negative reward results when moving into the target/obstacle cell; a small loitering penalty results for other state and action pairs. The semantic data codebook has $|\mathcal{O}| = 17$ possible observations o_t defined by $p(o_t|s_{r,t}, s_{tar}, s_{obs})$ (not listed here due to limited space). This characterizes the human sensor: given $s_{r,t}, s_{tar}, s_{obs}$, a 'correct' semantic report is provided with probability HA, and an 'incorrect' report occurs with probability $(1-HA)/16$. While this toy scenario greatly limits $|\mathcal{S}|$ and $|\mathcal{A}|$, the large $|\mathcal{O}|$ makes augmentation-based offline policy approximations [6] quite expensive even for small TD.

To address this, online POMCP policy approximation is used to find an optimal decision at each decision-making instance. The main idea behind POMCP is similar to the sampling-based Monte Carlo Tree Search (MCTS) online planning algorithm for Markov decision processes, which uses four main steps (tree search, tree expansion, simulation, and backpropagation) to estimate the state-action value function $Q(s, a)$ starting from the current state up to some specified tree depth, before choosing an optimal action a which maximizes Q. In POMCP, $Q(s, a)$ is replaced by $Q(h, a)$, where h represents a history of past actions and observations, and simulations of actions and observations to come up with local policy approximations starting from the robot's initial belief over the state s_t, given all available/received observations up to a given time. Following the implementation of the POMCP algorithm from [7], we manually tuned the search depth to 20 and exploration parameter to 2 (though POMCP hyperparameter tuning can also be performed online [12]).

We use simulation studies to assess POMCP's performance in this problem in terms of the number of steps required for the robot to find the target and total accumulated reward. Monte Carlo trials of 100 runs per test condition were used to investigate the extent to which exploration was affected by different values of TD (2,4) and HA (1.0,0.9,0.6,0.3). Note that POMCP implementation for the $TD = 0$ case is used as a comparison against POMCP implementation with $TD \neq 0$, to provide an 'ideal' upper performance bound. We also reran the $TD = 2$ cases with a variety of randomly generated strong and weak initial state/obstacle location distributions to assess sensitivity to the quality of prior state beliefs.

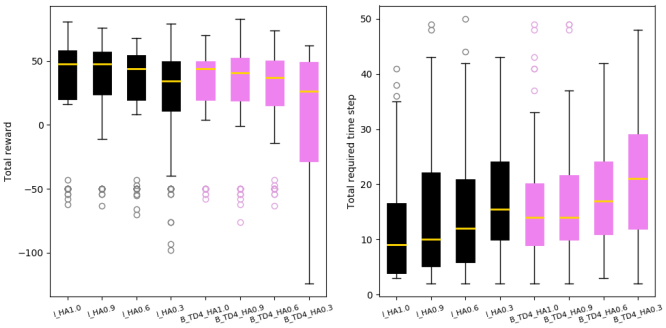

Fig. 2. Comparison between ideal situation (TD=0) and situation with delayed observations (TD=4). In this figure, I, B, TD, and HA represent Ideal, Baseline, Time Delay, and Human Accuracy level, respectively.

We highlight the main results [1]. Fig. 2 shows total reward and time to capture for 'ideal' $TD = 0$ POMCP (black box plots) and $TD = 4$ POMCP (pink Box plots), with increasingly worse human sensor accuracy and uniform priors. When HA is significantly different, performance is also significantly different *within* the $TD = 0$ and $TD = 4$ groups (e.g. $p = 0.004$, when compared the results of total reward in the case of $I\_HA = 1.0$ and $I\_HA = 0.3$). Also, between the $TD = 0$ and $TD = 4$ runs, significant differences are only seen when the human accuracy is high (i.e. $HA = 0.9$ and $HA = 1.0$), although differences are only due to the fixed TD. No significant performance differences were found for POMCP between $TD = 0$ and $TD = 4$ given the same HA values. POMCP thus appears on average to be robust to the delays considered for this scenario (also the case across different prior types). Also, HA plays a more significant role here regardless of TD. This is not too surprising since the human provides all data here, but hints at the need to develop methods that can filter 'outlier' human reports to ensure best performance. We also ran tests to assess whether POMCP's performance could be improved by augmenting delayed observations to the state within the tree search to 'anticipate' incoming information, as in [6]. We modified 'baseline' POMCP to forward sample delayed observations $o_{t-TD+1:t}$, based on augmented state beliefs at time t conditioned on $o_{1:t-TD}$. This did not provide a statistically significant performance gain over baseline POMCP, most likely due to the increased difficulty of sampling relevant joint delayed report, target and obstacle state configurations for large $|\mathcal{O}|$.

[1] in all cases, significance at $p = 0.05$ assessed via Mann-Whitney U-tests on total reward and time to reach target data

REFERENCES

[1] B. W. Israelsen, and N. Ahmed, ""Dave...I can assure you...that it's going to be all right..." A Definition, Case for, and Survey of Algorithmic Assurances in Human-Autonomy Trust Relationships,", *ACM Computing Surveys*, v. 51, no. 6, 2019.

[2] L. Burks, I. Loefgren, and N. Ahmed, "Optimal continuous state POMDP planning with semantic observations: a variational approach,", *IEEE Transactions on Robotics*, v. 35, no. 6, pp. 1488-1507, 2019.

[3] L. Burks, I. Loefgren, L. Barbier, J. Muesing, J. McGinley, S. Vunnam and N. Ahmed, "Closed-loop bayesian semantic data fusion for collaborative human-autonomy target search," *2018 International Conference on Information Fusion (Fusion 2018)*, Cambridge, UK., 2018.

[4] L. Burks and N. Ahmed, "Collaborative semantic data fusion with dynamically observable decision processes," *2019 International Conference on Information Fusion (Fusion 2019)*, Ottawa, Canada., 2019.

[5] T. Kaupp, A. Makarenko, S. Kumar, B. Upcroft and S. Williams, "Operators as information sources in sensor networks," *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Edmonton, Alta., 2005.

[6] P. Varakantham, and J. Marecki, "Delayed Observation Planning in Partially Observable Domains," *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS2012)*, Valencia, Spain, 2012.

[7] D. Silver and V. Joel, "Monte Carlo planning in large POMDPs," Advances in Neural Information Processing Systems 23, 2010.

[8] N. Sweet and N. Ahmed, "Structured synthesis and compression of semantic human sensor models for Bayesian estimation." In 2016 American Control Conference (ACC) 2016 Jul 6 (pp. 5479-5485). IEEE.

[9] N.R. Ahmed, E. M. Sample, M. Campbell, " Bayesian multicategorical soft data fusion for human–robot collaboration." IEEE Transactions on Robotics. 2012 Sep 12;29(1):189-206.

[10] M. Lewis, H. Wang, P. Velagapudi, P. Scerri, K. Sycara. "Using humans as sensors in robotic search." In 2009 12th International Conference on Information Fusion 2009 Jul 6 (pp. 1249-1256). IEEE.

[11] S. Jamieson, J. How, Y. Girdhar. "Active reward learning for co-robotic vision based exploration in bandwidth limited environments." In 2020 IEEE International Conference on Robotics and Automation (ICRA) 2020 IEEE.

[12] S. Wakayama, N. Ahmed. "Auto-Tuning Online POMDPs for Multi-Object Search in Uncertain Environments." In 2020 AIAA SciTech Forum 2020 AIAA.