# Robot Acting and Interacting Under Partial Observability and Perceptual Uncertainty

Kaiyu Zheng
Department of Computer Science, Brown University

## I. Introduction

Uncertainty is a primary challenge that robots must handle to be competent in many complex, human-centered domains. Imagine a robot that can help an elderly person search for a missing pair of glasses at home. The robot must decide what to do under uncertainty about (1) where the pair of glasses is; (2) what it is observing currently; (3) what observations it will receive after performing an action. Arguably, such uncertainty is constantly present, due to limited knowledge about the environment, noisy on-board sensors, and the unstructured nature of the human world. Consequently, considering uncertainty, namely, *partial observability* (1), and *perceptual uncertainty* (2, 3), is of central importance for robot *acting* in the human world, and it is crucial to my research.

At the same time, humans can provide, conveniently through natural language, a powerful source of knowledge and feedback to the robot (*e.g.*, "I remember seeing my glasses last time in front of the TV."). However, natural language is inherently subjective and ambiguous. Furthermore, to make the most of the human's presence, the robot should ideally be able to continuously and naturally *interact* with humans using natural language to better accomplish given tasks, a demanding yet necessary capability towards future collaborative robots.

To address these challenges, **my research aims to enable robots acting in human environments and interacting with humans in a principled manner.** The current methodology I have taken is based on Partially Observable Markov Decision Processes (POMDPs), which principally model both partial observability and perceptual uncertainty. I view natural language as an additional modality of stochastic perception as well as a type of action the robot can perform, which reduces the barrior of interfacing with humans.

Solving POMDPs for real world problems is computationally prohibitive. The key idea behind our work is in exploiting structures in the human world (*e.g.*, octrees, correlations) and human-robot interaction (*e.g.*, spatial language), which significantly eliminates unrealistic compromises that previous work make (such as constraining to 2D and object independence).

In support of this approach, I present my work that progress from "act" to "interact," (Fig. 1) using the POMDP framework. My study has focused on *object search*, a practically valuable yet generally complex problem that contains the key elements of uncertainty a robot faces. This statement discusses my research arranged into two parts:

- On the end of "act," I propose scalable planning algorithms for large POMDPs. This includes a multi-resolution POMDP planning algorithm for object search in 3D [25] and a hierarchical planning algorithm for POMDPs that model correlational object search [26].
- Progressing towards "interact," I extended Object-Oriented POMDP [16] with a spatial language observation model that let robots understand spatial language with potentially ambiguous spatial prepositions [24]. Last but not least, I discuss ongoing work on the end of "interact," where a robot engages in spoken dialogue with a human while searching for objects [11].

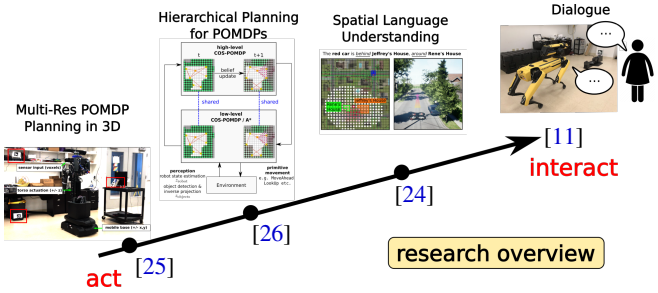


Fig. 1: My long-term research goal is to enable robots to *act* in the human world and *interact* with humans. My current research has focused on handling uncertainty, particularly partial observability and perceptual uncertainty, towards this goal ([25, 26, 24, 11]).

Prior to the above line of work, I dealt with uncertainty in robot perception through working on learning semantic maps [23, 21] and mobile robot navigation [20]. I have also developed a general-purpose POMDP library called `pomdp_py` [22].

## II. Act: Scalable Planning for Large POMDPs

### A. Multi-Resolution POMDP Planning in 3D

Robots deployed in households must find objects on shelves, under tables and in cupboards, naturally a 3D environment. Although past experience or semantic knowledge helps hypothesizing likely search regions [7, 3], object search ultimately depends on the ability to search carefully under limited field of view (FOV) within each region. Due to its computational complexity, previous works have constrained POMDP for object search in 2D regions [2, 5, 17]. The key challenges lie in the intractability of maintaining exact belief due to large state space [13], and the high branching factor for planning due to large observation space. In Zheng et al. [25], we propose a general POMDP formulation for the multi-object search task with 3D state and action spaces, and a realistic observation space in the form of labeled voxels within the viewing frustum from a mounted camera. To address

challenges of computational complexity, we propose a multi-resolution planning algorithm that centers on a novel octree-based belief representation, which captures beliefs at different resolutions simultaneously and allows efficient and exact belief updates. Our simulation results show that, as the problem scales, our approach outperforms exhaustive search as well as POMDP baselines without resolution hierarchy under the same computational requirement. We also show that our method is more robust to sensor uncertainty against the POMDP baselines. Finally, we demonstrate our approach on a torso-actuated mobile robot in a lab environment. The robot finds 3 out of 6 objects placed at different heights in two $10\text{m}^2 \times 2\text{m}$ regions in around 15 minutes. Our work demonstrates that such challenging POMDPs can be solved online efficiently and scalably with practicality for a real robot by extending existing general POMDP solvers with domain-specific structure and belief representation.

### B. Hierarchical Planning for Correlational Object Search

In realistic applications of object search is that robots will need to locate target objects in complex environments while coping with unreliable sensors, especially for small or hard-to-detect objects. In Zheng et al. [26], we introduce COS-POMDP (Correlational Object Search POMDP), a general planning framework for optimal object search leveraging given correlational information. This overcomes a limitation of the above work [25] where objects in the environment are assumed to be independent (for computational reasons). We address scalability by proposing a hierarchical planning algorithm, where a high-level COS-POMDP plans subgoals, each fulfilled by a low-level planner that plans with low-level actions (for example, using the local 3D object search algorithm in the above work [25]); both levels plan online based on a shared and updated COS-POMDP belief state, enabling efficient closed-loop planning. We evaluate the proposed approach in AI2-THOR [8], a realistic simulator of household environments, and we use YOLOv5 [10, 6] as the object detector. Our results show that, when the given correlational information is accurate, COS-POMDP leads to more robust search perfomance for target objects that are hard-to-detect. In particular, for target objects with a true positive detection rate below 40%, COS-POMDP significantly outperforms the POMDP baseline not using correlational information by 42.1% and a greedy, next-best view baseline [19] by 210% in terms of SPL (success weighted by inverse path length) [1], a recently developed metric that reflects both search success and efficiency.

## III. Interact: Language as an Interface

### A. Spatial Language Object-Oriented POMDP

Humans use spatial language to naturally describe object locations and their relations (*e.g.*. "The red car is in *front* of Chase Bank"). However, spatial language is inherently subjective and potentially ambiguous or misleading. In Zheng et al. [24], we consider spatial language as a form of stochastic observation. We propose SLOOP (Spatial Language Object-Oriented POMDP [16]), a new framework for partially observable decision making with a probabilistic observation model for spatial language. We apply SLOOP to object search in city-scale environments given a spatial language description of target locations. We collected a dataset of five city maps from OpenStreetMap [9] as well as spatial language descriptions through Amazon Mechanical Turk (AMT). To understand ambiguous, context-dependent prepositions (*e.g. behind*), we train a model that infers the latent frame of reference (FoR) given an egocentric synthetic image of the referenced landmark and surrounding context. Results show that our method finds objects faster with higher success rate by understanding spatial language compared to a keyword-based baseline used in prior work [16]. We deploy SLOOP for object search in AirSim [12], a realistic drone simulator, where the drone is tasked to find cars in a neighborhood environment.

### B. Dialogue Object Search

We envision robots that can collaborate and communicate seamlessly with humans. It is necessary for such robots to decide both what to say and how to act, while interacting with humans. In this ongoing work [11], we introduce a new task, *dialogue object search*: A robot is tasked to search for a target object (*e.g.*, fork) in a human environment (*e.g.*, kitchen), while engaging in a "video call" with a remote human assistant who has additional but inexact knowledge about the target's location. We conducted a pilot study where we experimented with both speech-based dialogue and text-based dialogue in a home simulation environment based on AI2-THOR [8] and found that participants typically engage in frequent back-and-forth when using speech. We also observed several common behaviors by participants such as describing their observations, beliefs, or movement recommendations.

Our focus on decision-making systems for full interactions under a principled framework differs from recent work in visual-dialogue navigation that predict the next action given dialogue history [14] and assume language input from an oracle during training [27]. There are fundamental challenges related to language grounding and dialogue systems (*e.g.*data collection, evaluation). Our current step is to develop a method that allows a robot to generate sequential intents together with planning physical actions for the search.

## IV. Future Work

For the near future, I plan to invest time mainly in [11] as I see the importance to develop a practical robot system capable of deciding what to say and how to act simultaneously, while combatting fundamental challenges in dialogue. I hope to learn from the success of recent large language models [15, 4]. My goal is to conduct a user study with an effective dialogue object search system in the real world. This could be indicative for my research towards future robots in human environments. In the long run, my research goal is to develop a general framework that unifies decision-making for dialogue and physical action that can be deployed in real-time on a robot for various tasks. Finally, I hope to expand my research to domains where a robot interacts with objects under partial observability, a topic I recently started to work on [18].

## References


[1] Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018.

[2] Nikolay Atanasov, Bharath Sankaran, Jerome Le Ny, George J Pappas, and Kostas Daniilidis. Nonmyopic view planning for active object classification and pose estimation. *IEEE Transactions on Robotics*, 30(5):1078–1090, 2014.

[3] Alper Aydemir, Kristoffer Sjöö, John Folkesson, Andrzej Pronobis, and Patric Jensfelt. Search in the real world: Active visual object search based on spatial relations. In *2011 IEEE International Conference on Robotics and Automation*, pages 2818–2824. IEEE, 2011.

[4] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[5] Michael Danielczuk, Andrey Kurenkov, Ashwin Balakrishna, Matthew Matl, David Wang, Roberto Martín-Martín, Animesh Garg, Silvio Savarese, and Ken Goldberg. Mechanical search: Multi-step retrieval of a target object occluded by clutter. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 1614–1621. IEEE, 2019.

[6] Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, ChristopherSTAN, Liu Changyu, Laughing, tkianai, Adam Hogan, lorenzomammana, yxNONG, AlexWang1900, Laurentiu Diaconu, Marc, wanghaoyang0106, ml5ah, Doug, Francisco Ingham, Frederik, Guilhen, Hatovix, Jake Poznanski, Jiacong Fang, Lijun Yu, changyu98, Mingyu Wang, Naman Gupta, Osama Akhtar, PetrDvoracek, and Prashant Rai. ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements, October 2020. URL https://doi.org/10.5281/zenodo.4154370.

[7] Thomas Kollar and Nicholas Roy. Utilizing object-object and object-scene context when planning to find things. In *IEEE International Conference on Robotics and Automation*, pages 2168–2173. IEEE, 2009.

[8] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

[9] OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org , 2017.

[10] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[11] Monica Roy, Kaiyu Zheng, Liu, Jason, and Stefanie Tellex. Dialogue Object Search. In *RSS Workshop on Robotics for People (R4P): Perspectives on Interaction, Learning and Safety*, 2021. URL https://arxiv.org/pdf/2107.10653.pdf. Extended Abstract.

[12] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics*, 2017.

[13] David Silver and Joel Veness. Monte-Carlo planning in large POMDPs. In *Advances in neural information processing systems*, 2010.

[14] Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406. PMLR, 2020.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[16] Arthur Wandzel, Yoonseon Oh, Michael Fishman, Nishanth Kumar, and Stefanie Tellex. Multi-Object Search using Object-Oriented POMDPs. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019.

[17] Yuchen Xiao, Sammie Katt, Andreas ten Pas, Shengjian Chen, and Christopher Amato. Online planning for target object search in clutter under partial observability. In *Proceedings of the International Conference on Robotics and Automation*, 2019.

[18] Shangqun Yu, Sreehari Rammohan, Kaiyu Zheng, and George Konidaris. Hierarchical reinforcement learning of locomotion policies in response to approaching objects: A preliminary study. In *The 5th Multidisciplinary Conference on Reinforcement Learning and Decision Making (RLDM)*, 2022.

[19] Zhen Zeng, Adrian Röfer, and Odest Chadwicke Jenkins. Semantic linking maps for active visual object search. pages 1984–1990. IEEE, 2020.

[20] Kaiyu Zheng. *ROS Navigation Tuning Guide*, pages 197–226. Springer International Publishing, Cham, 2021. ISBN 978-3-030-75472-3.

[21] Kaiyu Zheng and Andrzej Pronobis. From pixels to buildings: End-to-end probabilistic deep networks for large-scale semantic mapping. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3511–3518, 2019.

[22] Kaiyu Zheng and Stefanie Tellex. pomdp_py: A framework to build and solve POMDP problems. In *ICAPS 2020 Workshop on Planning and Robotics (PlanRob)*, 2020. Github link: "https://github.com/h2r/pomdp-py".

[23] Kaiyu Zheng, Andrzej Pronobis, and Rajesh P. N. Rao. Learning Graph-Structured Sum-Product Networks for probabilistic semantic maps. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, New Orleans, LA, USA, February 2018.

[24] Kaiyu Zheng, Deniz Bayazit, Rebecca Mathew, Ellie

Pavlick, and Stefanie Tellex. Spatial language understanding for object search in partially observed cityscale environments. In *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2021.

[25] Kaiyu Zheng, Yoonchang Sung, George Konidaris, and Stefanie Tellex. Multi-resolution POMDP planning for multi-object search in 3D. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* **IROS RoboCup Best Paper Award**. IEEE, 2021.

[26] Kaiyu Zheng, Rohan Chitnis, Yoonchang Sung, George Konidaris, and Stefanie Tellex. Towards optimal correlational object search. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2022.

[27] Yi Zhu, Yue Weng, Fengda Zhu, Xiaodan Liang, Qixiang Ye, Yutong Lu, and Jianbin Jiao. Self-motivated communication agent for real-world vision-dialog navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1594–1603, 2021.