Anirudh Valiveru
28 September 2023
SuperUROP Proposal

Learning Dexterous Robot Grasps in Unstructured Environments

Motivation

The field of robotics has transformed over the past few decades from a niche field to a fundamental part of daily life. Whether in the form of self-driving cars or factory assembly-line robots, intelligent embodied agents have transformed the way that humans interact with each other and automate tedious tasks. The culmination of robotic research, however, lies in the development of everyday assistant robots, which are able to understand and execute complex multi-step tasks within unstructured and novel environments. Progress in this space has accelerated with the advent of data-driven learning-based methods for control and perception, and *unstructured dexterous manipulation* is especially critical to making the grand vision of generalizable robotics a reality.

While natural language processing and computer vision have benefited greatly from the deep learning revolution due to copious amounts of publicly available datasets, roboticists do not have this luxury. Model-based reinforcement learning for dexterous tasks using human video demonstrations is a promising direction because it allows us to draw correspondences between robot and human limbs, bootstrapping task learning and making scalable real-world grasping possible. More specifically, this project plans to solve this problem with the constraints of limited on-board perception and unknown environments to address key challenges in *mobile manipulation*. We hope to use recent advancements in 3D perception to learn accurate world models that help the robot build a rich semantic and dynamic understanding of its surroundings.

Related Work

Several previously developed methods for training end-to-end visuomotor policies, such as Google's recent RT-2 (Robot Transformer 2) serve as a proof-of-concept that an end-to-end learning approach is a viable strategy for robust and generalizable robot control [1]. Training a policy end-to-end allows the network to learn emergent capabilities that may not be possible in a more modular approach, such as semantic reasoning of actions given visual inputs. These emergent capabilities can be especially useful in mobile dexterous settings, since the robot will need to learn unintuitive policies, features, and embeddings to perceive and plan in its environment.

This project would use an approach similar to *Structured World Models from Human Videos* or *Affordances from Human Videos as a Versatiles Representation for Robotics* to learn the robot's environment using human video exploration data [2, 3]. Leveraging human-collected videos in place of robot teleoperation data is critical for researchers to scale robotic applications in diverse environments. This work focuses on learning *affordances*, which can be interpreted as contextual understanding of objects in the robot's environment, and information about how an object moves

is encoded within the affordance representation. Learning more structured object representations however, such as URDF *(Universal Robot Description Format)* files with joint specifications, will allow robots to have a more concrete understanding of a scene's kinematics using classical physics-based methods, along with a more structured way to plan for motion within the environment.

One major bottleneck in this approach, however, is the relative lack of relevant data necessary to train policies that allow robots to manipulate arbitrary objects, especially when compared to natural language processing or computer vision. We hope to leverage off-the-shelf segmentation with *Segment Anything* and a single-shot 3D reconstruction model such as *One-2345* or *Multiview Compressive Coding* to construct an estimate of an object's shape as a point cloud representation [4, 5, 6]. This will then be tracked over several video frames to learn an object's kinematic structure.

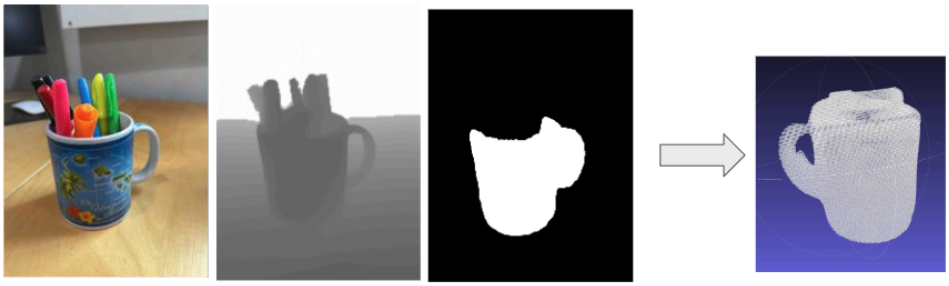

Figure 1: *MCC* (Multiview Compressive Coding) reconstruction of "Mug"

With an accurate environment model, the usage of end-to-end visuomotor systems using RL to train mobile manipulators to navigate, pick, and place objects is quite promising. The recent Georgia Tech paper *Adaptive Skill Coordination for Robotic Mobile Manipulation* utilized end-to-end RL to execute these tasks accurately on small objects located in open boxes [7]. By learning structured models that are easy for robots to leverage, this project hopes to explore similar methods to enable dexterous manipulation tasks if time permits, such as opening doors and watering plants.

Preliminary Methodology

The primary objective in the first stage of this project is to explore various scene representation strategies that will allow a robot to leverage human video demonstrations most effectively. Recently, 3D model reconstruction techniques have proven to be quite promising, and point-tracking may allow a mobile robot like Spot to follow a human around an environment as they demonstrate how various objects interact. We simultaneously plan to look into methods that learn kinematic scene representations such as URDF, that will be useful in later project stages.

After this is done, we plan to train a model that accurately predicts an object's URDF file as a representation of its geometry and kinematic behavior, as opposed to a static point-cloud. This is similar to the *affordance learning* approach used in works mentioned earlier in the **Related Works** section, but with the added benefits of also being reproducible in simulation during training.

Finally, we hope to train a model end-to-end that is able to learn a manipulation task, such as opening a door, simply by watching a human a few times and then performing the task themselves. This will be a long-term undertaking whose details will become more clear as our research progresses.

In conclusion, this research direction is a promising step towards solving unstructured robot grasping while bypassing the primary bottlenecks of teleoperated data and an expensive and controlled camera setup. If successful, this project will be a significant contribution to the goal of enabling robot grasping of any object anywhere.

References

[1] Brohan, Anthony, et al. "Rt-2: Vision-language-action models transfer web knowledge to robotic control." *arXiv preprint arXiv:2307.15818* (2023).

[2] Mendonca, Russell, Shikhar Bahl, and Deepak Pathak. "Structured World Models from Human Videos." arXiv preprint arXiv:2308.10901 (2023).

[3] Bahl, Shikhar, et al. "Affordances from Human Videos as a Versatile Representation for Robotics." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[4] Kirillov, Alexander, et al. "Segment anything." arXiv preprint arXiv:2304.02643 (2023).

[5] Liu, Minghua, et al. "One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization." arXiv preprint arXiv:2306.16928 (2023).

[6] Wu, Chao-Yuan, et al. "Multiview compressive coding for 3D reconstruction." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[7] Yokoyama, Naoki, et al. "Adaptive Skill Coordination for Robotic Mobile Manipulation." arXiv preprint arXiv:2304.00410 (2023).