Looking Good: Visually Informative Motion Generation for Mobile Manipulation

Sophie Lueth[1], Snehal Jauhri[1], and Georgia Chalvatzaki[1,2,3]
[1] Computer Science Department, Technische Universität Darmstadt, Germany
[2] Hessian.AI, Darmstadt, Germany
[3] Center for Mind, Brain and Behavior, Uni. Marburg and JLU Giessen, Germany
sophie.lueth@stud.tu-darmstadt.de
snehal.jauhri@tu-darmstadt.de
georgia.chalvatzaki@tu-darmstadt.de

***Abstract*—Mobile Manipulation (MM) systems incorporate the benefits of mobility and dexterity, thanks to the enlarged space in which they can move and interact with their environment. Mobile manipulator robots can also continuously perceive their environment when equipped with an embodied camera. However, extracting relevant visual information in unstructured environments such as households remains a challenge. In this work, we propose an active perception pipeline for mobile manipulators to generate motions that are informative toward manipulation tasks such as grasping. Our proposed approach moves the robot's mobile base in a scene in a way that incorporates both visual information gain and task success in a time and energy-efficient manner.**

I. Motivation

In the near future, embodied AI agents such as mobile manipulators are expected to operate autonomously in everyday environments such as households. However, performing tasks in these environments is challenging due to the unstructured nature of the real world. The problem becomes harder when considering robots that gather information from an embodied camera. In such a case, the robot's motion generation must consider both visual information gain and the manipulation objective. This abstract proposes an effective and efficient approach for visually informative motion generation for mobile manipulation.

We consider the problem of grasping objects in cluttered scenes. While significant progress has been made in 6-DoF grasp detection [1, 7, 4], grasping in cluttered scenes can still be challenging due to partial observability. Occlusions and partial visibility of the target object can hinder grasp success. Thus, actively perceiving and accumulating volumetric information about the target object can be an effective strategy [2]. We propose an active perception pipeline that makes sense for grasping by mobile manipulators.

II. Related Work

For active perception, many methods have been proposed for motion generation using a Next-Best-View (NBV) strategy [5, 11, 9]. Typically, the robot's objective is reconstruction, i.e., obtaining complete volumetric information about the scene or target object. Most NBV methods choose a viewpoint based on a notion of information gain (IG). The objective is to minimize uncertainty about the scene/object and discover unobserved regions. Delmerico et al. [3] provide a good overview and comparison of different IG formulations for NBV.

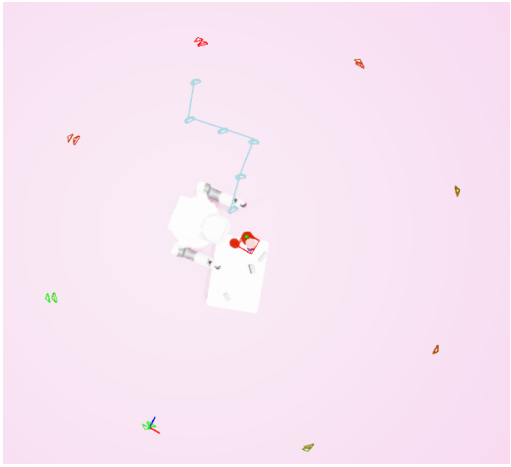
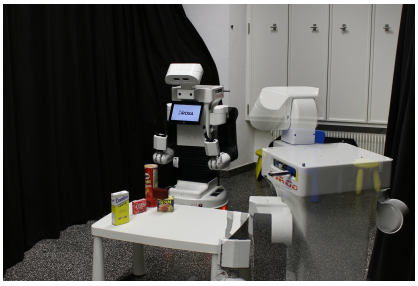


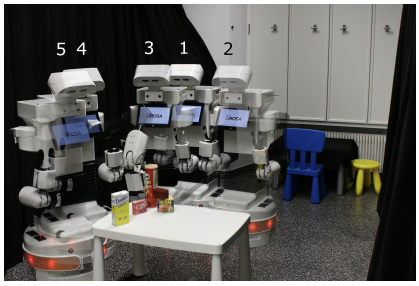
Fig. 1: A top view of an example motion generated by a mobile manipulator to perceive and grasp a target object. The target object is highlighted in the red box. The circle around the scene shows the candidate viewing angles, and green and red correspond to more informative and less informative viewing angles. The light blue line shows the actual viewing trajectory followed by the robot.

For most manipulation tasks, scene reconstruction is not necessary to perform the task effectively, and obtaining full volumetric information is sub-optimal and inefficient. Especially in Mobile Manipulation, since the cost of moving in the scene is higher as compared to static manipulation, time and energy-efficient active perception solutions are needed. For grasping, we need a balance: Collecting enough information to infer a good grasp and doing so efficiently, i.e. focusing on details relevant to grasp execution.

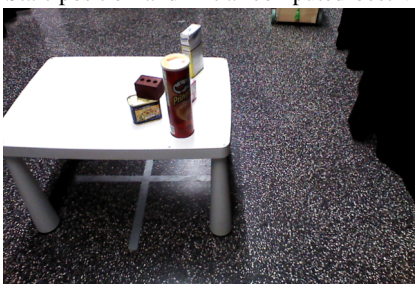
Recently, some methods have adopted grasp quality metrics to choose the next robot viewpoint that minimizes uncertainty in the grasp pose estimate [8, 2]. Breyer et al. [2], who base their work on [5, 3], exploit the fact that only negligible performance



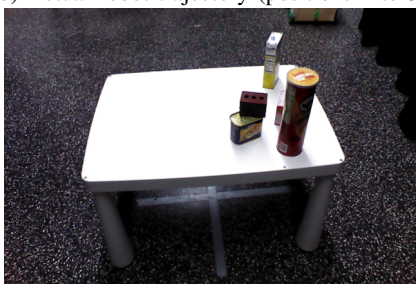
(a) Start position and initial computed best view



(b) Actual robot trajectory (positions 1 to 5)



(c) Robot view from start position (position 1)



(d) Robot view from position 2

Fig. 2: Sub-optimal motion generation of a mobile manipulator robot when applying existing Next-Best-View strategies. The target object is the red 'Jello' box. The robot starts in a position where the object is occluded (c). The initial computed NBV from this start position is shown in (a). This NBV position is far away and sub-optimal in terms of grasp execution. Moreover, whilst moving towards it, the robot captures further RGBD information from position 2 and then *changes* direction to move towards a different NBV at position 4 and execute the grasp on position 5. The information collected at viewpoint 2 is not relevant for the grasp. The robot would be more efficient to move to its right from the start, collecting information at closer viewpoints 3 and 4 and executing the grasp.

differences can be detected in the different formulations of IG computed on an incomplete volumetric representation. They choose the rear side voxel IG formulation and compute IG per proposed next view, equally distributed on a half-sphere above the object, via ray-casting. The active perception is considered complete when a stable grasp quality has been perceived.

While such methods show promise for static manipulation/grasping tasks, there are several drawbacks when considering their application toward mobile manipulation. Firstly, the reachability of the grasps is especially important for mobile manipulators to execute the grasp. Secondly, we take the view that the overall viewing trajectory is important since visual information is gathered over whole trajectories, and we wish to avoid wasteful movement of the whole robot as exemplified in Figure 2). The characteristic of existing methods to only consider the NBV with the most information gain can lead to unnecessarily long motions of the mobile manipulator.

III. Proposed Approach

We consider scenarios where a mobile manipulator robot is placed in a household environment and tasked with picking up a target object placed on a table with clutter. The robot uses a head-mounted RGBD camera to perceive the scene. An RGB object detection network is used to detect a rough bounding box of the target object. Moreover, the depth information is used to build a volumetric grid representation around the target object. The robot's objective is to detect high-quality grasps in the target volume through the movement of its base (SE(2)), adjusting its head elevation angle (θ), and increasing/decreasing the height of its torso (z). Thus the robot has 5 degrees of freedom to change its viewpoint. To detect grasps, a grasping network is used that can process volumetric information [1, 7].

As in [2], a ray-casting procedure is used to predict the information gain from any potential view (Rear-Side Voxel IG from Delmerico et al. [3]). However, we propose to generate trajectories towards the object and consider trajectory-level information gain. This provides two benefits. Firstly, this ensures the robot moves to satisfy the grasping objective. In this way, perception and grasping are combined into one smooth behavior instead of two separate discrete behaviors, which would be more time-consuming. Secondly, the trajectory-level information gain avoids the problem of the constant switching of NBVs and sub-optimal movement, as shown in Figure 2(b).

Besides grasp quality, the robot should also consider grasp reachability to be effective. For example, if a grasp is only reachable by moving to the other side of the table, it is sub-optimal. Thus we propose the incorporation of grasp reachability metrics into the information gain and stopping criterion. Reachability of a mobile manipulator can be computed using reachability inversion [10] or learned reachability metrics [6].

We believe this holistic approach toward visual information gain can be a very effective solution for mobile manipulation in real household scenarios.

References

[1] Michel Breyer, Jen Jen Chung, Lionel Ott, Roland Siegwart, and Juan I. Nieto. Volumetric grasping network: Real-time 6 DOF grasp detection in clutter. In *CoRL*, volume 155 of *Proceedings of Machine Learning Research*, pages 1602–1611. PMLR, 2020.

[2] Michel Breyer, Lionel Ott, Roland Siegwart, and Jen Jen Chung. Closed-loop next-best-view planning for target-driven grasping, 2022. URL https://arxiv.org/abs/2207.10543.

[3] Jeffrey Delmerico, Stefan Isler, Reza Sabzevari, and Davide Scaramuzza. A comparison of volumetric information gain metrics for active 3d object reconstruction. *Autonomous Robots*, 42(2), 2018. URL https://link.springer.com/article/10.1007/s10514-017-9634-0#citeas.

[4] Haojie Huang, Dian Wang, Xupeng Zhu, Robin Walters, and Robert Platt. Edge grasp network: A graph-based se(3)-invariant approach to grasp detection. *ICRA*, 2023.

[5] Stefan Isler, Reza Sabzevari, Jeffrey A. Delmerico, and Davide Scaramuzza. An information gain formulation for active volumetric 3d reconstruction. *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3477–3484, 2016.

[6] Snehal Jauhri, Jan Peters, and Georgia Chalvatzaki. Robot learning of mobile manipulation with reachability behavior priors. *IEEE Robotics and Automation Letters*, 7(3):8399–8406, 2022.

[7] Zhenyu Jiang, Yifeng Zhu, Maxwell Svetlik, Kuan Fang, and Yuke Zhu. Synergies between affordance and geometry: 6-dof grasp detection via implicit representations. In *Robotics: Science and Systems*, 2021.

[8] Douglas Morrison, Peter Corke, and Jürgen Leitner. Multi-view picking: Next-best-view reaching for improved grasping in clutter. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8762–8768, 2019. doi: 10.1109/ICRA.2019.8793805.

[9] Menaka Naazare, Francisco Garcia Rosas, and Dirk Schulz. Online next-best-view planner for 3d-exploration and inspection with a mobile manipulator robot. *IEEE Robotics and Automation Letters*, 7(2):3779–3786, apr 2022. doi: 10.1109/lra.2022.3146558. URL https://doi.org/10.1109%2Flra.2022.3146558.

[10] Nikolaus Vahrenkamp, Tamim Asfour, and Rüdiger Dillmann. Robot placement based on reachability inversion. In *ICRA*, 2013.

[11] David Watkins-Valls, Peter K Allen, Henrique Maia, Madhavan Seshadri, Jonathan Sanabria, Nicholas Waytowich, and Jacob Varley. Mobile manipulation leveraging multiple views, 2021. URL https://arxiv.org/abs/2110.00717.