MR. CRABS: Mobile RGB-D Camera-LIDAR Robot for Autonomous Bimanual Manipulation & Sensing

Giuse Pham*[1], Matthew Strong*[2], Alex Qiu[1], Joonwon Kang[1], Monroe Kennedy III[1]

***Abstract*— Near-ground bimanual manipulation is seldom studied for manipulation, as a variety of household items require ground interaction with two arms. Additionally, most robots in this domain require moving their *entire body* to view the relevant scene, making tasks, such as environment reconstruction, much slower. To fill these gaps, we present a fully integrated perception system for near-ground bimanual mobile manipulation. This consists of a novel bimanual holonomic robot equipped with dual Intel RealSense depth cameras and a small LIDAR. Unlike other near-ground and bimanual robots, our robot, entitled *MR. CRABS*, is equipped with a pan-and-tilt bottom Realsense, which we show improves mapping and search tasks. Finally, MR. CRABS can perform a proof-of-concept bimanual task for demonstration collection, alleviating the need for teleoperation.**

I. INTRODUCTION

Near-ground manipulation tasks, such as retrieving objects from under furniture, inspecting floor-level infrastructure, or operating in low-clearance environments, are more common than public perception suggests. Unlike tabletop robots, **near-ground workspaces pose significant perception challenges for mobile robots**. Near-ground workspaces suffer from severe viewpoint constraints: cameras mounted at standard heights produce oblique views with heavy occlusion, and the geometry and area of the ground plane within the frames of near-ground robots limit diverse achievable perspectives [1]. To explore a scene, a robot must move inefficiently.

While recent bimanual mobile platforms have often demonstrated ground-level exploration and manipulation capabilities, they suffer from key shortcomings. [2] and [3] only support tabletop tasks. The non-holonomic AhaRobot [4] supports ground tasks, but the mounted camera cannot move in place. YOR [5], the most similar robot to ours, is both holonomic and supports both table-top and bimanual tasks. However, none of its cameras are *active, in-place*, and it does not have LIDAR. Similar to how humans pan and tilt their head to observe a scene without full physical movement, this motivates an *active camera* for near-ground tasks – a feature to quickly improve understanding of a scene.

We present a system that addresses this challenge through three contributions a). **novel bimanual holonomic mobile manipulation platform** equipped with dual Intel RealSense depth cameras, and a top LIDAR; b). a **simple method for active vision information gain** based on an equipped RGB-D pan-and-tilt camera; c). a **quantitative & qualitative** evaluation of the effects of this camera for mapping.

This research was supported by NSF Graduate Research Fellowship No. DGE-2146755.

[·][1]Department of Mechanical Engineering [·][2]Department of Computer Science, Stanford University, Stanford, CA, USA. Emails: {gpham123, mastro1, aqiu34, jwkang, monroek}@stanford.edu.

*Both authors contributed equally to this work.

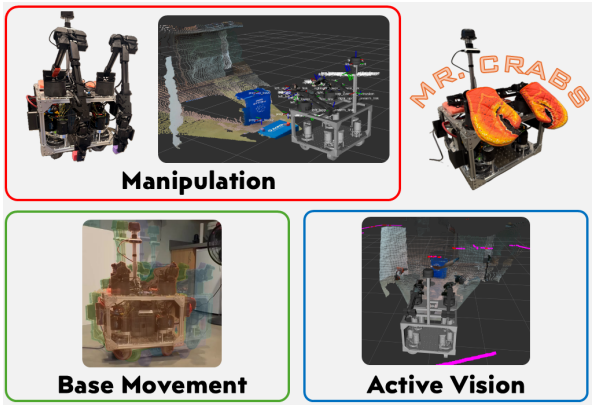

Fig. 1: MR. CRABS: a holonomic mobile TidyBot++ base with dual WidowX-250 arms, a top-mounted LIDAR, and two RealSense depth cameras: one with pan-tilting for active near-ground perception.

II. METHOD

We now present *MR. CRABS*, the bimanual mobile manipulator on a TidyBot++ base. The method is divided into two components: **Hardware Overview** and **Active Perception**.

A. Hardware Overview.

MR. CRABS features a low-cost, modular, smaller mobile base designed from the TidyBot++ hardware package [6]. MR. CRABS is capable of holonomic motion (independently capable of moving in its three DoFs: x, y, θ). Unlike TidyBot++, MR. CRABS uses two Trossen Robotics WidowX-250 arms for bimanual manipulation, given its size and ability to fit on the top of the smaller base. MR. CRABS' perception and mapping capabilities also differ from TidyBot++, relying on two Intel RealSense RGB-D cameras and an RPLiDAR system. The RealSense camera mounted on the base is attached to two Dynamixel motors for panning and tilting motion, and is used to observe the scene closer to the ground, but can be used for objects higher off the ground. The base has an aluminum extrusion to elevate the RPLiDAR sensor for obstacle detection and mapping. It also holds a static RealSense camera to observe the greater scene.

B. Active Perception.

We formulate object search as greedy selection over a discrete action space $\mathcal{A}$ combining base movements $(\Delta\theta, \Delta x, \Delta y)$ with pan-tilt camera poses $(\phi_{\text{pan}}, \phi_{\text{tilt}})$, totaling

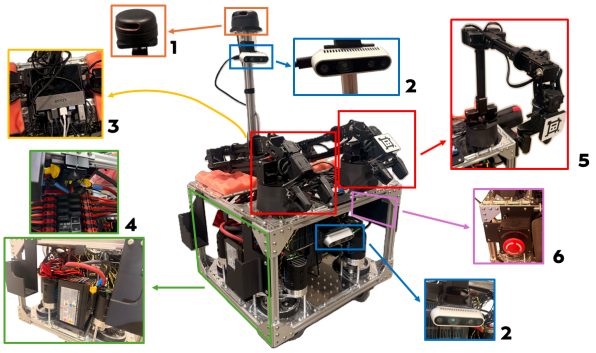

Fig. 2: **Hardware overview of MR. CRABS**: (1) RPLiDAR for mapping and obstacle avoidance, (2) two realsense cameras for navigation and object searching, (3) Intel NUC with USB hubs for sensor input, (4) base movement sub-system, including battery, swerve modules and a breaker to e-stop the base, (5) dual WidowX-250 arms for bimanual manipulation, and (6) an e-stop for the arms.

a set $|\mathcal{A}| = 81$ actions, which consists of every unique combination of 9 total camera poses (3 different pans and tilts) with 9 total combined combinations of base movements. The robot maintains an observed voxel set $\mathcal{V}_t \subset \mathbb{Z}^3$ at resolution $\delta = 5$cm. For each candidate action a, we **project the camera frustum** from the hypothetical post-action pose by sampling $N = 400$ rays up to $d_{\max} = 3$m, yielding a predicted visible voxel set $\mathcal{F}(a)$. The information gain is: $g(a) = |\mathcal{F}(a) \setminus \mathcal{V}_t| \cdot (\rho)$, where the ρ is a quadratic multiplier that upweights actions into unexplored regions, specifically by using the amount of non-viewed voxels. Pan-tilt actions at saturated poses are penalized, while base movements into fresh poses receive a novelty bonus. The robot selects $a^* = \arg\max_a \; g(a)/c(a)$, where $c(a)$ is the estimated execution time, then queries SAM3 [7] for each target object. Naturally, this perception scheme favors camera sweeps with high viewpoint variance over expensive base motions. Future work will consider information gain over *trajectories*, then compute actions over the most informative paths.

III. Experiments And Results

A. Robot Task

As a proof of concept, MR. CRABS autonomously picks up two blocks in each gripper and places them in a trash bin. Using the combined point cloud, we employ SAM3 to compute object masks. PCA on the segmented point cloud extracts the object's major axis, and the gripper is oriented perpendicular to it to grasp across the narrow dimension. We utilize J-PARSE [8] for safe entry in and out of kinematic singularity while retrieving objects at the edge of the workspace. Such tasks can be used as augmented demonstrations for a robot learning policy, where only successful tasks are used. Across a set of 5 trials, we see a strong 80% success rate.

B. Active Search for Multiple Objects

We highlight trials of the ability of the pan-tilt camera motion to assist the base in finding semantically meaningful objects faster. We construct a simple experiment where the robot has to select the highest voxel information gain to find a Stanford bunny and a water bottle on opposite sides of the robot. We run SAM3 on each view (both top and bottom RealSense cameras) until both objects are detected. On 5 trials' average, Mr. CRABS finds both objects in 11.7 seconds with camera and base motion, and compared to 24.3 seconds with base motion.

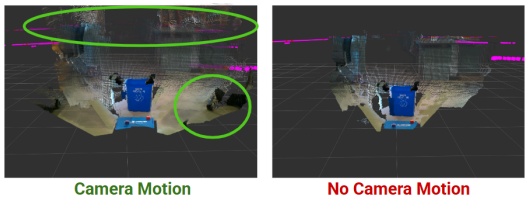

Fig. 3: Point cloud reconstruction with (left) and without (right) pan-tilt camera motion. Green circles highlight additional coverage from active sweeping.

TABLE I: Voxel Coverage Comparison (Sweep vs Fixed Camera)

	Fixed Camera	Sweep Camera	Ratio
Voxels	39,196	196,515	5.0x
Volume	1.06 m^3	5.31 m^3	5.0x
OctoMap nodes	51,453	554,044	10.8x

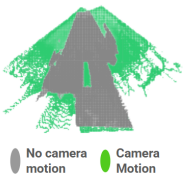

Fig. 4: Voxel coverage with a fixed (gray) vs. panning (green) camera, showing the additional environment coverage gained through camera motion.

1) Straight Line Mapping: To highlight the robot's usefulness in moving and using the Pan-Tilt motion, we perform an isolated experiment to have the robot move approximately over 1 meter in the forward direction with OctoMap [9] and a voxel size of 0.03m. In one ablation, we run with no camera motion; in the other, with consistent camera motion. In Table I, we find that voxel coverage consists of about 5 times more voxels and volume. A top-down view of the voxels is shown in Fig. 4 in ground slices, demonstrating a much broader view of the environment, while moving in only one direction.

IV. Conclusion

In this work, we presented **MR. CRABS**, a bimanual, holonomic mobile manipulator with a sensing suite for more efficient active perception. The design of the hardware is selected in such a way as to allow for rapid exploration of a scene for manipulation, mapping, and search. Future work focuses in two directions: a full-body teleoperation interface and autonomous mapping of an environment, especially to alleviate the burden of a human operator; and **active** perception during manipulation for difficult close-to-floor manipulation tasks, such as occlusion handling and cluttered scene manipulation.

REFERENCES

[1] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Transactions on robotics*, vol. 32, no. 6, pp. 1309–1332, 2017.

[2] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," *arXiv preprint arXiv:2401.02117*, 2024.

[3] A. Shaw, C. Liu, J. Costa, R. Gray, A. Skowronek, K. Diaz, N. Bui, and N. Correll, "Cutting the cord: System architecture for low-cost, gpu-accelerated bimanual mobile manipulation," *arXiv preprint arXiv:2603.09051*, 2026.

[4] H. Cui, Y. Yuan, Y. Zheng, and J. Hao, "Aharobot: A low-cost open-source bimanual mobile manipulator for embodied ai," *arXiv preprint arXiv:2503.10070*, 2025.

[5] M. H. Anjaria, M. E. Erciyes, V. Ghatnekar, N. Navarkar, H. Etukuru, X. Jiang, K. Patel, D. Kabra, N. Wojno, R. A. Prayage *et al.*, "Yor: Your own mobile manipulator for generalizable robotics," *arXiv preprint arXiv:2602.11150*, 2026.

[6] J. Wu, W. Chong, R. Holmberg, A. Prasad, Y. Gao, O. Khatib, S. Song, S. Rusinkiewicz, and J. Bohg, "Tidybot++: An open-source holonomic mobile manipulator for robot learning," *arXiv preprint arXiv:2412.10447*, 2024.

[7] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang *et al.*, "Sam 3: Segment anything with concepts," *arXiv preprint arXiv:2511.16719*, 2025.

[8] S. Guptasarma, M. Strong, H. Zhen, and M. Kennedy III, "J-parse: Jacobian-based projection algorithm for resolving singularities effectively in inverse kinematic control of serial manipulators," *arXiv preprint arXiv:2505.00306*, 2025.

[9] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "OctoMap: An efficient probabilistic 3D mapping framework based on octrees," *Autonomous Robots*, 2013, software available at https://octomap.github.io. [Online]. Available: https://octomap.github.io