Towards robustly picking unseen objects from densely packed shelves

Markus Grotz, Soofiyan Atar, Yi Li, Paolo Torrado, Boling Yang,
Nick Walker, Michael Murray, Maya Cakmak and Joshua R. Smith
Paul G. Allen Center for Computer Science & Engineering
University of Washington, USA
Email: {grotz, mcakmak, jrs}@cs.washington.edu

***Abstract*—For industrial warehouses to be fully automated, robots must be able to pick previously unseen objects from densely packed shelves. Dense packing makes objects more difficult to distinguish visually and requires manipulation to be sensitive to the arrangement of objects in the shelf itself. Hence, key challenges that arise are the visual segmentation and tracking of previously unseen objects in cluttered environments, as well as manipulation planning and control to pick densely packed objects from the shelf. We present a complete system that is able to pick unseen objects from a cluttered shelf. Our system consists of components to track shelf inventory, re-identify requested objects and to autonomously pick them. In our experiments, the system is able to pick objects in highly cluttered scenes with a success rate of 66% and 53.8 successful picks per hour. We provide a classification of the pick attempts and their frequency to motivate future research.**

I. Introduction

The human ability to pick objects from densely packed shelves is unparalleled, unmatched by today's industrial pick-and-place setups. Consider, for instance, a bookshelf with meticulously arranged books, or an industrial storage room housing a wide variety of objects on large shelves. The goal in these environments is to maximize space utilization and minimize the time required to find and retrieve objects.
As depicted in Fig. 1, an industrial robotic arm manipulator faces the task of picking several objects from an industrial shelf. This task presents multiple robotic challenges, such as visual perception of previously unseen objects or robust object handling. Perception and robotic manipulation in such cluttered environments are difficult due to occlusion and lack of a priori information about the objects, where often only a label or book title is known. For industrial applications, the robotic system must also handle a diverse range of objects quickly and robustly.

To address these challenges, we propose a system capable of picking objects in cluttered and constrained shelf environments. Our method is evaluated through real-world experiments using industrial shelves and a variety of objects. We also introduce categories for failure classification to identify promising future research directions and challenges.

The rest of this paper is structured as follows: Sec. II provides a literature review. Sec. III describes our method and system architecture. Sec. IV presents our evaluation results, and Sec. V concludes our work.

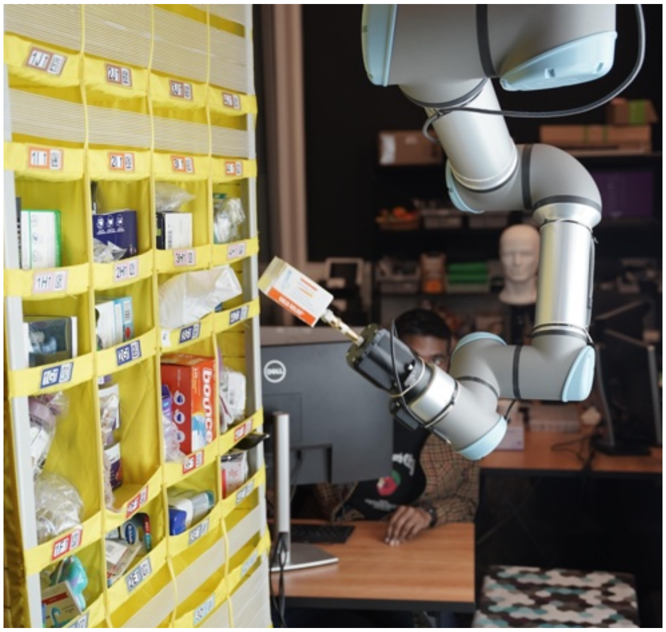

Fig. 1. A UR16e robotic arm manipulator picking a requested object from a densely packed shelf. Before the pick, a virtual inventory model was built as objects were stowed in the bins. Objects can be rearranged while stowing and thus need to be re-identified before manipulation.

II. Related Work

Robustly grasping and manipulating objects is an active research area [11] and includes aspects such picking objects from cluttered table-top scenarios [8] or 6-DoF grasping in clutter [10]. Benchmarks and competitions have played a significant role in evaluating grasping and manipulation [15]. Benchmarking is time-consuming, however, and the sim-to-real gap limits the utility of simulated alternatives. Frameworks have been developed to understand the sim-to-real gap or to make experiments more reproducible [12]. However, these are limited to table-top scenarios. For industrial applications perhaps the most visible research competition is the Amazon Picking Challenge. Here object detection is a key element [17]. The ACRV benchmark [6] introduces reproducible guidelines for object arrangement using a widely available shelf. However, the object set is limited to 42 object and the focus is not on cluttered scenarios. Other works deal with manipulation

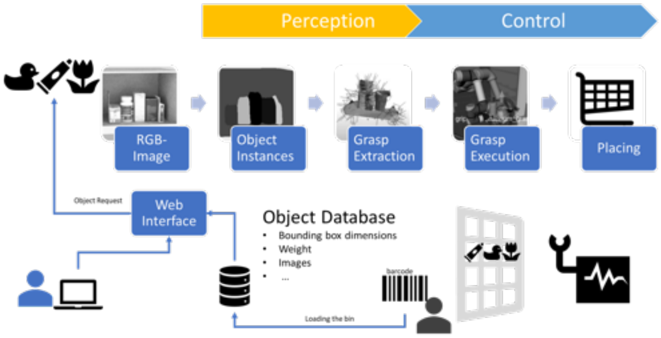

Fig. 2. The system architecture and workflow for picking objects from densely packed shelves.

planning and control for logistics scenarios like restocking a grocery store shelf [4]. Recently, [9] introduced ARMBench, a large-scale, object-centric benchmark dataset for robotic manipulation in warehouses. The major focus of ARMBench is on object segmentation and identification in clutter as well as on defect detection. In contrast to previous work, our focus is on highly cluttered and densely packed industrial shelf environments.

III. System Architecture

Our software architecture consists of perception and control components sequenced by a state machine. Fig. 2 delineates its structure. The state machine follows the canonical workflow of picking items from a shelf. Our system also includes a database and a web interface, which are used to curate the inventory of the bins and to enable a human to send pick requests to the state machine.

A. Object database and web interface

The database contains the stowing history of the shelves as well as information for all available objects. This includes, for example, object names, physical properties, and a unique product identification number.

The web interface coordinates the process by which a human "stower" loads the shelf. When a human scans both an object barcode and a barcode of a bin, it creates appropriate database entries to keep track of the bin in which an object has been placed. The interface calculates and displays metrics such as the bin utilization, which indicates "how full the bins are" to help the stower achieve desired levels of density. Finally, the web interface is also used to request objects to be picked from the bins. Once an operator has selected the objects a request is generated that triggers the state-machine.

Every time an object has been stowed an RGB-D image of the bin is captured and used to extract object feature embeddings for later re-identification.

B. State machine

The state machine processes the list of requested objects and manages the pick process for each of them. The object is re-identified in the bin, a pick pose is extracted and the manipulator is moved accordingly. If the object is picked successfully the manipulator drops the object at a location where it can be processed further.

The state machine also contains a retry mechanism that allows a human operator to ask the robot to execute a picking action on the same object again after the previous pick attempt has failed.

C. Visual perception

The central challenge for the visual perception system is accurately segmenting previously unseen objects in the bin and consistently identifying the same object across multiple images. In large-scale industrial warehouse settings, the ability to handle unseen objects is crucial, as new objects are frequently introduced. However, the definition of *what constitutes an object* can be ambiguous and varies across different environments. For instance, a pack of shampoo should be regarded as a single entity rather than multiple identical objects. Another critical factor is the ability to re-identify objects. Many objects in these environments are textureless, which limits the use of keypoint-based methods for identification [7, 13]. The wide variety of object categories also constrains the use of template-based methods for identification [16]. Moreover, the presence of discrete frames in the sequence hampers the effectiveness of methods assuming continuous frames [2, 5]. To tackle these challenges, we employ a method that jointly performs instance segmentation and re-identification of unseen objects.

The term "re-identification" is akin to tracking, but instead of operating on a continuous video stream, it involves tracking the same object across frames that are not continuous in time. Our method is based on the works of [5] and [1], which pioneered innovative techniques for object segmentation and tracking in complex environments.

The primary innovation of our perception system is its ability to simultaneously generate *queries* for each frame independently while allowing communication between frames efficiently. The property that the queries are generated for each frame independently is crucial for maintaining segmentation accuracy over non-sequential frames. Concurrently, our method allows efficient communication between queries originating from different frames. This dual capability facilitates both intra-frame and inter-frame communication, thereby enhancing the accuracy of instance segmentation and re-identification.

During the stowing process, objects are re-identified using extracted object embeddings queries. These embeddings are essentially unique identifiers for objects, generated based on their visual features. Our network, trained solely on synthetic data, is capable of segmenting and re-identifying objects in real-world image sequences, even in challenging conditions with heavy clustering and large object movements. The result of this process is a segmentation mask for the requested object, as demonstrated in Fig. 3, showing the segmentation results of a bin.

D. Manipulation planning and execution

After the requested object has been identified successfully, the next step is to extract possible pick points or grasp

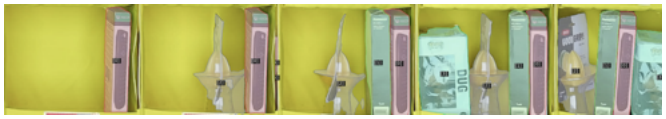

Fig. 3. Segmentation masks and matching results. An inventory model is built by tracking object embeddings over time-discrete frames. From left to right: Five segmentation masks and tracking results of a single bin after a new objects are stowed.

Fig. 4. A subset of the object set used throughout evaluation. The set consists of objects of different shapes, sizes and weights, as well as deformable objects and objects wrapped in plastic bags.

hypotheses. A heuristic computes the pick pose close to the center of the segmentation and tests the reachability with inverse kinematics. A pre-grasp pose is also computed and the end-effector is controlled accordingly. For control and planning of our system we use the widely available MoveIt library [3]. We use the libraries OMPL [14] for planning and TRAC-IK to find inverse kinematic solutions and MoveIt's Cartesian planner plans the path from the pre-grasp pose to the grasp pose. When the grasp pose is reached, the end-effector vacuum is activated and the system executes a pushing motion towards the back of the bin. The pushing motion ensures that the vacuum is sealed and the object is picked securely.

E. Success and Failure Detection

During grasp execution the system uses the robot's force torque sensor and a vacuum level sensor in the suction cup to determine if an object has been picked. The force output controls the execution of the Cartesian planner. Above a certain threshold the plan is deemed executed and the state machine then transitions to the next state. Finally, the state machine moves the end-effector with the picked object to a predefined placement location and releases the object. Failure is determined by the detected pressure on the suction cup. The object is not securely attached to the suction cup if the detected pressure value falls below a threshold value.

IV. Evaluation

A. Workcell setup

Our workcell, shown in Fig. 1, uses a cantilever-mounted Universal Robots UR16e, configured with a Robotiq EPick suction gripper, which has been extended to increase the reachability (c.f. Fig. 1). A frame mounted Azure Kinect RGB-D sensor points towards a warehouse shelving unit. The shelf has several bins each of which can be densely packed with objects.

	Trial 1	Trial 2	Trial 3	Trial 4
Average bin utilization	40 %	40%	41%	41%
# unique requests	16	17	18	17
# pick attempts	22	21	26	23
# of successfully picked items	7	14	9	15
# of fail items	15	7	17	8
# of retries	6	4	6	3
Pick success %	43%	82%	50%	88%
Time (min:sec)	15:12	12:34	13:08	11:19
UPH (units per hour)	27.6	66.8	41.1	79.5

TABLE I
Evaluation results

The shelving unit consists of four sides with differently sized bins, however for the evaluation we use one side of the unit and a subset of 16 bins due to the robot's limited workspace. While adding a gantry or rotating the shelves would alleviate these limitations, this was not necessary for the purposes of our research. The bins are stowed with a huge variety of different objects, which differ in size and shape. Fig. 4 shows a small subset of the objects. Overall, the databases curates more than 1050 objects of various shapes and size and is continuously growing.

B. Experiments

We conducted four trials, each with different kinds of objects. To quantify our approach we report on (a) Pick Success Rate, (b) Units Per Hour (UPH) as outlined in [11]. Moreover, we also report the bin utilization as the density within the bin is an important indicator for the difficulty of grasping objects. The bin utilization u defined for a bin is defined as

$$u(bin) = \frac{1}{\mathcal{V}_{bin}} \sum_{o \in bin} \mathcal{V}_{\mathcal{OBB}}(\mathrm{o}) \,, \tag{1}$$

where $\mathcal{V}_{bin}$ denotes the volume of the bin, and $\mathcal{V}_{\mathcal{OBB}}(\mathrm{o})$ volume of the object oriented bounding box of object o. The bin utilization in Eq. 1 can then be averaged for all bins of the shelf.

For the evaluation 173 objects of different shapes and sizes were stowed across the trials (or an average of 43 objects per trial). Fig. 5 shows the state of the bins after stowing. The bins were loaded with a bin utilization of more than $40\,\%$. The number of objects picked from each bin was limited to maintain the challenge of clutter throughout the trial; removing an object lowers the bin utilization and each subsequent pick in the bin easier. In the first and third trial, objects show a greater displacement while stowing the shelf, i.e. objects were moved within a bin.

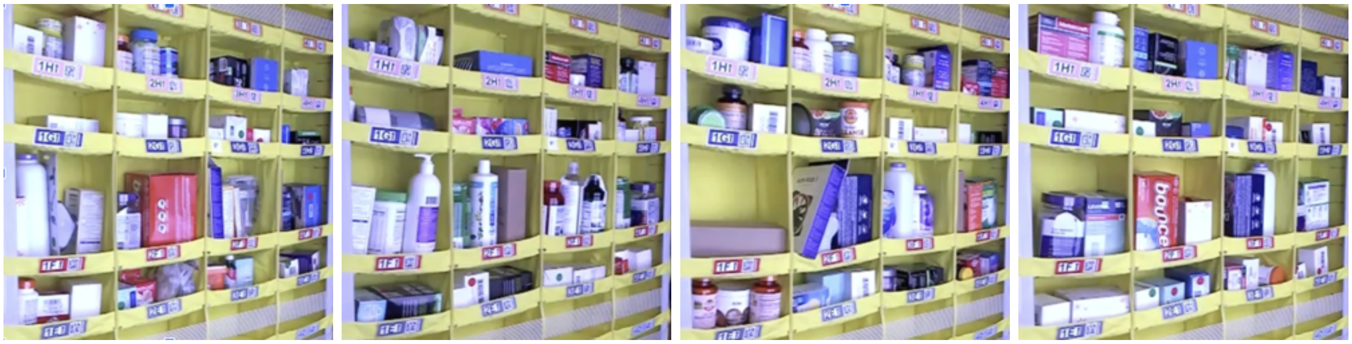

Fig. 5. Overview of the objects that were stowed during the trial runs 1-4 from left to right. Trial 1 and 3 include higher displacements during the stowing.

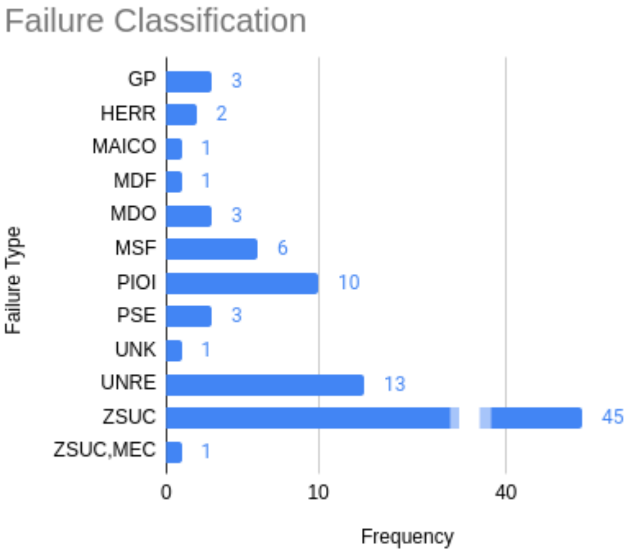

Fig. 6. Classification of the failure types during the trial runs.

Error code	Description
Grasp Planning	
GP	general error
Manipulation	
MAICO	mis-pick - actuator incapable of capturing object
MDF	mis-pick - dynamic failure
MDO	mis-pick - deformable object
UNRE	unreachable object
MSF	mis-pick - suction failure
MEC	extraction collision
Perception	
PIOI	incorrect object identification
PIOL	incorrect object location/pose
POFO	object fully occluded
PSE	segmentation error
Other	
HERR	human error
UNK	unknown
XPASS	object is skipped
Success	
YSUCH	success w/ human help
ZSUC	success

TABLE II
CLASSIFICATION OF PICK ATTEMPT OUTCOMES

C. *Results*

Tab. I lists the evaluation results. Out of the trials, 68 objects were requested. In total the system made 92 pick attempts in 52 minutes and 13 seconds. 45 were successful, resulting in an overall success rate of 66 % and a rate of successful picks of 53.8 mean picks per hour. Performance variation of our system partially comes from different stowing approaches as well as object configurations. This includes, for example, objects being stacked on top of each other. Other sources of variation include stochastic processes in the system like the motion planner.

1) Pick and failure classification: Failures observed when using the system can be roughly categorized as resulting from *grasp planning*, *manipulation*, *perception*, reflecting the different stages of the state machine. Tab. II lists categories that are used for classification of successful / unsuccessful pick attempts. Fig. 6 shows the frequency of the pick classifications. In some rare cases, multiple categories are used. For example if an object is picked successfully (**ZSUC**), but during picking it dropped another object (**MEC**).

V. DISCUSSION AND FUTURE DIRECTIONS

We present a system that picks previously unseen objects out of densely packed shelf environments. We evaluate our system in real-world application with heavy clutter and a constrained environment. Our method is able to pick previously unseen objects with a pick rate of 53.8 units per hour. Our pick attempt classification shows that most of the unsuccessful picks are due to the design of the gripper or the manipulation strategy. To address the limitations of our commercial vacuum gripper, we are investigating innovative gripper designs that include Time of Flight sensors, which will provide live system state for online manipulation, and an array of movable suction cups with higher vacuum flow rates to assist with object grasping. Additionally, we are exploring approaches that involve humans in the loop including human assisted failure correction and learned manipulation strategies from human teleoperated demonstrations.

Acknowledgments

This work has been supported by Amazon under the research project ”Robotic Manipulation in Densely Packed Containers”. We further would like to thank our students Atharva Kashyap, Jack Lowry, Kavi Dey, Liang-Hsi Yen, Shakti Senthil, Suyash Sanjeev, Tenzin Low, and Thomas Kaminski for their valuable work during their research internships.

References

[1] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022.

[2] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, pages 640–658. Springer, 2022.

[3] David Coleman, Ioan Sucan, Sachin Chitta, and Nikolaus Correll. Reducing the barrier to entry of complex robotic software: a moveit! case study. *arXiv preprint arXiv:1404.3785*, 2014.

[4] Marco Costanzo, Simon Stelter, Ciro Natale, Salvatore Pirozzi, Georg Bartels, Alexis Maldonado, and Michael Beetz. Manipulation planning and control for shelf replenishment. *IEEE Robotics and Automation Letters*, 5 (2):1595–1601, 2020. doi: 10.1109/LRA.2020.2969179.

[5] Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. Vita: Video instance segmentation via object token association. *arXiv preprint arXiv:2206.04403*, 2022.

[6] Jürgen Leitner, Adam W Tow, Niko Sünderhauf, Jake E Dean, Joseph W Durham, Matthew Cooper, Markus Eich, Christopher Lehnert, Ruben Mangels, Christopher McCool, et al. The acrv picking benchmark: A robotic shelf picking benchmark to foster reproducible research. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4705–4712. IEEE, 2017.

[7] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004.

[8] Jeffrey Mahler, Jacky Liang, Sherdil Niyaz, Michael Laskey, Richard Doan, Xinyu Liu, Juan Aparicio Ojea, and Ken Goldberg. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. *arXiv preprint arXiv:1703.09312*, 2017.

[9] Chaitanya Mitash, Fan Wang, Shiyang Lu, Vikedo Terhuja, Tyler Garaas, Felipe Polido, and Manikantan Nambi. Armbench: An object-centric benchmark dataset for robotic manipulation. In *ICRA*, 2023.

[10] Adithyavairavan Murali, Arsalan Mousavian, Clemens Eppner, Chris Paxton, and Dieter Fox. 6-dof grasping for target-driven object manipulation in clutter, 2019. URL https://arxiv.org/abs/1912.03628.

[11] Rhys Newbury, Morris Gu, Lachlan Chumbley, Arsalan Mousavian, Clemens Eppner, Jürgen Leitner, Jeannette Bohg, Antonio Morales, Tamim Asfour, Danica Kragic, Dieter Fox, and Akansel Cosgun. Deep learning approaches to grasp synthesis: A review, 2022.

[12] Martin Rudorfer, Markus Suchi, Mohan Sridharan, Markus Vincze, and Aleš Leonardis. Burg-toolkit: Robot grasping experiments in simulation and the real world, 2022. URL https://arxiv.org/abs/2205.14099.

[13] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020.

[14] Ioan A. Şucan, Mark Moll, and Lydia E. Kavraki. The Open Motion Planning Library. *IEEE Robotics & Automation Magazine*, 19(4):72–82, December 2012. doi: 10.1109/MRA.2012.2205651. https://ompl.kavrakilab.org.

[15] Yu Sun, Joe Falco, Maximo A Roa, and Berk Calli. Research challenges and progress in robotic grasping and manipulation competitions. *arXiv preprint arXiv:2108.01483*, 2021.

[16] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6): 2872–2893, 2021.

[17] Andy Zeng, Kuan-Ting Yu, Shuran Song, Daniel Suo, Ed Walker, Alberto Rodriguez, and Jianxiong Xiao. Multi-view self-supervised deep learning for 6d pose estimation in the amazon picking challenge, 2016. URL https://arxiv.org/abs/1609.09475.