In: 29th International Conference on Automation and Computing (ICAC), 28-30 Aug 2024, Sunderland, United Kingdom
DOI: 10.1109/icac61394.2024.10718793

The Comprehensive Review of Vision-based Grasp Estimation and Challenges

1st Thanavin Mansakul
dept. School of Aerospace, Transport and Manufacturing
Cranfield University
Cranfield, UK
t.mansakul.621@cranfield.ac.uk

2nd Gilbert Tang
dept. School of Aerospace, Transport and Manufacturing
Cranfield University
Cranfield, UK
g.tang@cranfield.ac.uk

3rd Phil Webb
dept. School of Aerospace, Transport and Manufacturing
Cranfield University
Cranfield, UK
p.f.webb@cranfield.ac.uk

***Abstract*—Robotic grasping has emerged as a fundamental skill and a vital task for a robotic manipulator in various sectors over recent decades. Although a preprogramming method is now a general application, the challenges to handling complicated and unstructured scenarios remain. Machine vision, therefore, has become a focus of interest from many researchers as a primary perception to provide flexible manipulation in unknown and uncertain environments rather than control working space. This research presents a comprehensive review of vision-based grasp detection for a parallel gripper, analyzing potential techniques, existing challenges, and future directions. It delves into fundamental concepts of grasp detection and estimation, including traditional and learning-based methods. Additionally, the study explores essential benchmark datasets and metrics. This paper not only offers opportunities to develop grasp detection methodologies but also applications in the real world, such as fruit picking in agriculture, pick-and-pack items in supermarkets and logistics, and pick-and-sort objects in manufacturing. This will enable substantial changes and impacts of the robotic manipulator in the modern world.**

Keywords—Vision-based grasp estimation, object detection, object pose estimation, grasp detection, parallel gripper, manipulator

I. Introduction

In recent years, vision-based grasp detection has attracted research attention. Particularly, a vital role of machine vision provides visual perception to a robotic system inspired by a human eye. A camera is basically a robot's eye, allowing the robot to sense and interact with the environment. In the grasping task, the robot arm needs to know object information such as location related to the robot base, and shape. The robot then generates the best grasping point on the object based on the object model and surroundings. Next, the robot configuration relative to that point is set and ready to execute. Although it seems a simple process, tremendous challenges are inside each stage.

In the beginning, a big question is how the robot knows what and where the object exactly is in a digital image also converting into world coordinates. Afterwards, how to obtain the object information is another difficult point because the image only provides 2D data, pixels and RGB; however, the object contains 6D data or more, apparently 3 positions and 3 orientations in a coordinate system. The system needs to overcome these significant challenges, especially the development of algorithms and methodologies that can provide accurate object and surrounding information relevant to the robot as a crucial prerequisite in effective grasp detection.

Several noteworthy reviews concerning grasp detection and estimation have inspired this work to introduce a key understanding and to update current methodologies and challenges that could help to identify the research gaps and visions for advancement. To explain, the review on vision-based robotic grasping for parallel grippers [1] provides insights and outlines key methodologies, involving object localization, object pose estimation, and grasp estimation. Next, the survey on robotic grasping, spanning from classical to modern approaches [2], presents a comprehensive overview of analytical grasping methods until the latest data-driven grasping techniques. Then, the study on robotic grasp detection for parallel grippers [3] delivers two-dimensional (2D) plane methods and six degrees of freedom (6DOF) methods, revealing the evolving trends and opportunities. Additionally, the comprehensive investigation of three-dimensional (3D) vision-based robot manipulation [4] emphasizes the significance of 3D vision and the challenges in real-world scenarios. Lastly, the review of learning-based robotic grasping [5] outlines the current accomplishments and persisting challenges associated with automated grasping for unfamiliar objects, together with benchmark analyses. That could simplify insight grasp detection and up-to-date remarkable works, leading to further impactful development.

What's more, grasp detection requires further improvements to increase performance and practical ability such as higher accuracy, success rate, and speed. For this reason, this paper is a one-stop source, contributing key information and remaining challenges in this field to offer suggestions for the next evolution. In the initial section, important components of grasp detection are explained, including object detection, object pose estimation and grasp estimation. Next, challenges in grasping are described to emphasise the remaining issues that should be solved. Finally, future research trends are provided to show further elaboration and potential directions.

II. Important components in grasp estimation

A. Object detection

Object detection serves as a primary module in grasp detection. Firstly, definitions should be described. According to Blank et al. [6], object classification categorizes objects within an image into a certain class, and object localization defines the location of objects within the image frame mostly through the bounding boxes.

Meanwhile, object detection is concerned with both object localization and object classification by locating the bounding boxes and corresponding classes in the objects, therefore critical two metrics for object detection are accuracy from

classification and localization, and speed [7]. These explanations could help to clearly understand the significance of each element to determine their suitability for applications.

In the vision-based grasp detection system, the first stage is object detection, involving 2D and 3D detection, and segmentation. The primary input data consists of RGB, RGB-D, or video. The source of these inputs can range from standard webcams to high-definition cameras, depending on the purpose and budget. Generally, RGB-D cameras are favoured choices owing to their capabilities and affordability such as Kinect, Intel RealSense, and ZED. Noteworthy, ZED is now a better option due to its high precision and accuracy [8]. These devices capture images and measure the distance between the camera and the objects. Thereafter, the captured image could pass through image processing techniques to improve image quality and eliminate noises.

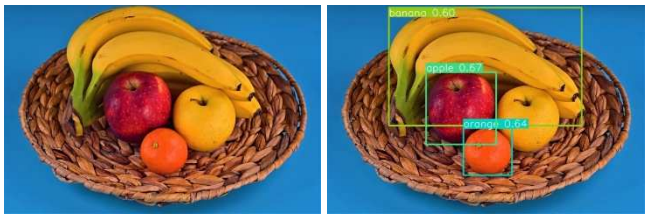

Fig. 1. An example of object detection

After the acquisition of the input image, the process of object localization is employed to extract objects from the background. Du et al. [1] classify object localization into three main groups: 1) Object localization without classification; 2) Object detection; and 3) Object instance segmentation. These groups are further divided into subgroups, such as 2D and 3D methods, classical and learning-based approaches, as well as one-stage and two-stage methods. Several techniques come with advantages and disadvantages, depending on the operating task. For example, the one-stage object detection is known for its speed, but it is less accurate compared to the two-stage methods. When prioritizing speed and lightweight computation, one-stage methods should be the preferred choice. Nevertheless, both accuracy and speed hold significance as they contribute to the achievement of grasp success rate in the context of grasp detection.

Regarding the difference between 2D and 3D methods, 3D object detection is expected to be widely adopted in robotic grasping, but it does not fit the actual object model, resulting in not enough information to perform robotic grasp [1]. They could provide the centre of the estimated 3D bounding box and the boundary for collision detection. Surprisingly, 2D methods are now of outstanding use in grasp detection, particularly for 2D planar grasp and instance segmentation analysis together with depth images for 6D grasp. Both methods still need to be improved in terms of accuracy and speed to ensure alignment with the goal.

Looking into the renowned object detection methods, although the traditional detection methods have become old-fashioned since 2010, their concepts are still important. While learning-based methods are now trends because of high performance [7], so they could be useful now and in the future. Firstly, the single-stage detection methods are designed to localize bounding boxes or segmentation masks, and then to predict class scores for objects within the input image. Accomplishing these tasks is usually incorporated with regression. The examples are now the most effective techniques based on the results of the MS-COCO dataset [9].

As to the current state-of-the-art, YOLOv10 [10] has emerged as a cutting-edge object detection method, showcasing superior accuracy and speed. It uses consistent dual assignments instead of Non-maximum Suppression (NMS) to reduce latency and holistic efficiency-accuracy-driven model design to optimize several components to improve accuracy and efficiency. In addition, Large-Kernel Convolution and Partial Self-Attention (PSA) are introduced to enhance accuracy. Next, the Real-Time Detection Transformer (RT-DETR) [11] is an NMS-free framework and employs a hybrid encoder separating intra-scale interaction and cross-scale fusion to reduce computation. Plus, the uncertainty-minimal query selection is presented to increase accuracy by integrating the uncertainty into the loss function. This method is very competitive to the YOLO series, with only a few differences in accuracy and latency. Last but not least, Single shot multi-box detector (SSD) [12] is a straightforward approach, offering multi-scale bounding boxes through multiple feature maps. Principally, it is run by a feed-forward convolutional network followed by non-maximum suppression, leading to advanced capabilities in object detection precision and speed. However, dense and small objects are still challenging for one-stage detection as well as improving accuracy and speed. The development could suggest analyzing the structure as an explainable process in machine learning and advancing new strategies.

On the other hand, the two-stage methods contain a two-step process. Beginning with the identification of regions of interest (ROI) represents the foreground objects against the background. Subsequently, these regions are classified in the second stage to determine the class labels. They enhance accuracy in comparison to single-stage detection, whereas they tend to be more complicated and relatively slow.

First of all, Faster R-CNN [13] builds upon the advancements of Fast R-CNN and SPPnet by sharing convolutional features in a region proposal network (RPN) with a downstream detection network to reduce processing time. Then, Mask R-CNN [14] extends the capabilities of Faster R-CNN by incorporating an additional branch for predicting an object mask along with the bounding box recognition. Although this method increases accuracy, it does not optimize for speed. Lastly, Cascade R-CNN [15] introduces a multi-stage framework including one stage for the region proposal network (RPN) and three subsequent stages for detection regression to avoid overfitting during training and inference-time mismatch. It shows an excellent average precision (AP). The two-stage methods have not much received attention for a while, but they have been applied and modified to particular applications that show a pleasant outcome. As several remarkable techniques have been described, the selection of a suitable method depends on a range of factors, including the characteristics of the objects, environments, and the capabilities of the equipment. In general, each method is associated with advantages and disadvantages. Accordingly, an understanding of algorithms, requirements, and applications could help to select a well-suited method.

B. Object pose estimation

As far as object information is highly concerned, object pose estimation plays an important role in grasp detection to estimate physical appearance data and then provide it to

grasp estimation for the best decision-making. Essentially, the 6D pose, including 3D translation and 3D rotation, is a need for robotic manipulation.

Du et al. [1] conclude that the 6D pose estimation process can be classified into three primary categories: correspondence-based, template-based, and voting-based techniques. The correspondence-based methods are compatible with rich textured objects or complex geometric shapes, while the template-based methods are suitable for sparse textures or simple geometrical objects. In case of occlusion, limited view, or dealing with objects at the category level, voting-based methods are well-performed.

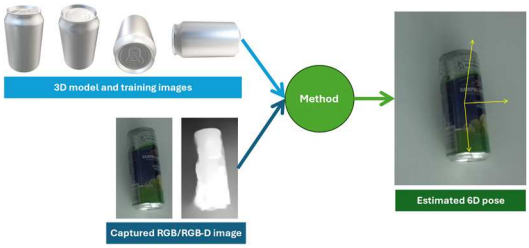

Fig. 2. 6D pose estimation

In the influence of machine learning, Marullo et al. [16] highlight the potential of learning-based methods, especially regression-based approaches. They show outstanding performance in terms of real-time speed and accuracy in scenarios involving texture-less objects, symmetrical objects, occlusions, and cluttered scenes. As a result, the adoption of learning algorithms represents a contemporary advancement, distinctive from traditional algorithms, though dependent on training datasets and substantial computational resources. Nonetheless, each technique has its benefits and drawbacks depending on the tasks such as model-based objects, model-free objects, seen objects, or unseen objects. The benchmark on any mission is needed to investigate performance and rank among other methods.

Regarding the evaluation of pose estimation, the performance of proposed pose estimation methods is often assessed using the Benchmark for 6D Object Pose Estimation (BOP) [17], which constitutes a significant community resource. BOP covers essential training and benchmark datasets, including Linemod (LM), Linemod-Occluded (LM-O), T-LESS, ITODD, and YCB-Video, among others. They also organize the BOP challenge every year to award the best result in each task. The ranking of models is there and some methods are open source that could be useful to implement and further develop, especially robotic field. For example, the GDRNPP [18] stands as the current state-of-the-art, showing impressive performance in both accuracy and execution time while utilizing only RGB images. That is the advantage of RGB over RGB-D input. Consequently, using RGB images now becomes a conventional implementation by employing a sequence of 2D detection followed by 6D pose estimation.

On the other hand, He et al. [19] emphasize a key aspect of 6D pose estimation in real-world robotic grasping applications. They advise that only relying on RGB data could result in incomplete geometry data. Likewise, the depth information and point cloud quality occasionally have troubles such as data sparsity and the absence of texture, leading to down performance. Therefore, the fusion of RGB images and point clouds emerges as a more effective strategy to compensate for their downsides. In regard to improving the accuracy of the input images, traditional point cloud refinement techniques, such as Iterative Closest Point (ICP) or Model-to-Model Cloud Registration (MCN), are often used, but they are suboptimal and time-consuming. If accuracy matters more than time, refinement methods should be utilized.

Many studies have affirmed the significance of incorporating a pose estimation stage within grasp estimation. Ahmad [20] asserts that the integration of pose estimation overcomes many difficulties, including variations in datasets, lighting conditions, clutter, occlusions, and the diverse sizes and shapes of objects. Without the pose estimation, the object's informational representation remains confined to a flat surface, being prone to grasp failure. Furthermore, the Amazon Picking Challenges (APC) [21] show that the fusion of 6D object poses and multi-view images through deep neural networks has proven influential in important constraints such as cluttered environments, self-occlusion, and limited data from depth sensors.

Despite that, this approach requires extensive training data and prior knowledge of 3D models. The universal and ideal method is challenging to receive better accuracy and speed together with robustness in several kinds of objects and situations. Either traditional or learning methods are meaningful and the combination of them could be beneficial.

C. Grasp estimation

Grasp estimation concerns determining the configuration of the gripper relative to the target within the camera frame, as well as assessing the grasp quality [1], [4]. Grasp estimation can be categorized into two parts: 2D planar grasp and 6DOF grasp.

In 2D planar grasp, the two crucial concepts are contact points and oriented rectangles. In the case of contact points, analytical methods or learning-based approaches are mainly employed to generate candidate samples. Analytical methods are used to attain a stable grasp by incorporating the kinematics and dynamics models of the system [2]. The geometric models and physical properties of the object are considered essential. Analytical methods aim to seek grasp stability and quantify grasp quality [4]. Mostly, form-closure and force-closure are frequently utilized for evaluating the quality of a grasp, in addition to Grasp Wrench Space (GWS) [2]. As an explanation, the grasp success rate serves as a key metric in empirical experiments, defining success when the object is successfully placed at the destination. If the object is dropped at any point, it will be considered a failure. This underscores the significance of analytical methods in the assessment process.

For learning approaches, they are primarily divided into two mains: learning to regress and learning to evaluate. The learning to regress approach directly computes the grasp pose for objects, leveraging object detection derived from images or point clouds. In contrast, the learning to evaluate method assesses candidate grasp by employing a scoring system.

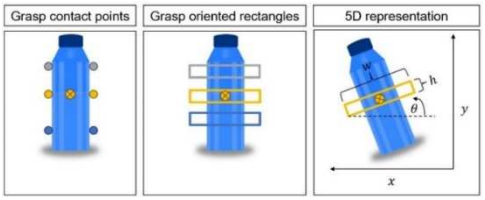

Fig. 3. 2D planar grasp

Jumping to oriented rectangles, it encapsulates gripper configuration, incorporating parameters such as width, height, and orientation on the target objects. In this way, they facilitate fast computation and precise grasp determination. For example, Lenz et al. [22] accomplish remarkable results in accuracy and success rate by introducing a deep learning method with efficient computation and a five-dimensional representation for the rectangle, encompassing position (x, y), orientation (θ), width (w), and height (h) as shown in Fig. 3. The evaluation process involves generating multiple candidates, selecting the one with the highest rank, typically around the centroid area for enhanced stability.

Fig. 4. 6DOF grasp

In 6DOF grasp, it is capable to perform grasping from various directions in 3D space. It requires 6D information about the object and then sets the 6D gripper pose related to the object. The process contains two primary components: perception and planning [3]. The perception involves estimating the object's location and identifying suitable grasping points. While the planning focuses on the movement of the manipulator and the gripper configuration in relation to the target. Generally, the input of the 6DOF grasp can be partitioned into a partial point cloud and complete shape [1].

For the partial point cloud approaches, a section of the object is captured, and then grasp candidates are constructed based on this partial observation. After that, the grasp quality of each candidate is computed to opt for the most effective grasp. Alternatively, grasp positions can be transferred from the existing dataset by identifying correspondences between the observed partial view and the complete model of the objects in a database.

For the complete shape, the objects are recognized with precise 3D models in the database. The retrieving a 6DOF grasp pose could match existing grasps. If a suitable grasp pose is not available in the knowledge base, an analytic method is employed as an optional. This approach is heavily dependent on segmentation, viewpoint, and matching strategies. The hidden parts are not visible to the camera, significantly impacting the complete accuracy. However, in a situation where the complete shape is ideally obtained and accurately matched with the database, this method can offer highly exceptional results. Once the object information is received, gripper configuration will be the next step.

Concerning degrees of freedom, the grasp representations can span from 2 degrees of freedom (2DOF) to 7 degrees of freedom (7DOF), subject to the parameters of the gripper and object representation [4], as illustrated in TABLE I. For description, the variables x, y, and z denote the coordinates of the grasping point; θ represents orientation; h and w denote the height and width of the gripper, respectively; R signifies orientation as a rotational matrix; and T denotes position as a transitional matrix. In real-world scenarios, most current research focuses on the exploration of 6DOF and 7DOF for flexibility and realistic usability. Importantly, the object representation can indicate the grasping representation, hence extracting object information becomes a crucial module.

TABLE I. GRASPING REPRESENTATION [23]

Object representation	*Parameters*	*Number of degrees of freedom*
2D	x, y, z or x, y, θ	3
	x, y, z, θ	4
	x, y, z, θ, h, w	5
3D	R, T	6
	R, T, w	7

In the matter of datasets and metrics, the evaluation for contact points, oriented rectangles, and 6DOF grasp is given. According to Cong et al. [4], many datasets are available for different methods. For contact point representations, the Stanford Grasping Dataset is a common baseline. This dataset, generated through simulation, focuses on single objects without cluttered scenes. However, it lacks the complexity of cluttered environments.

In the case of oriented rectangle representations, the Cornell Grasp Dataset, built on real-world scenarios, is commonly used. This dataset mainly features single objects. A well-known metric for evaluation in this dataset is accuracy, determined by whether the angle between the predicted value and the ground truth is within 30 degrees, in which case the prediction is counted as correct. Similarly, the Jacquard dataset, constructed based on simulation for single object grasp detection, considers a correct prediction if the predicted rectangle intersects with the ground truth by over 25 per cent. According to Meta AI researcher [24], There are the best models on Cornell Grasp Dataset and Jacquard dataset: (1) Grasp_det_seg_cnn [25], (2) GR-ConvNet [26], (3) ResNet50 multi-grasp predictor [27], and (4) Efficient-Grasping [28]. They are all learning-based methods and rely heavily on training datasets.

For 6DOF representations, the GraspNet-1Billion dataset, generated from real cluttered environments, is often employed for evaluation purposes. This dataset offers a more comprehensive and challenging setting for testing grasp estimation methods. Evaluating a proposed method using this dataset is highly constructive. It allows for a comparative analysis with other research works to show improvements and establish new state-of-the-art methodologies in the field of grasp estimation. The best models for this benchmark is Graspnet-1billion [29] and Anygrasp [30].

Nevertheless, the traditional approaches should be in consideration owing to less computation, and math and science background. Recent traditional methods, as provided by Zapata-Impata et al. [31], [32] employ a geometry-based method using point clouds. This approach does not require prior knowledge and emphasises grasp stability. It has achieved state-of-the-art with a high success rate and speed. Typically, other traditional methods focus on a 2D planar grasp by identifying the centre of mass of the target and attempting to grasp it from that point.

III. Challenges in Grasping

In grasp estimation, grasping a target along with stability, accuracy, and robustness under various real-world scenarios become significant challenges. A crucial aspect of grasp stability involves mechanical analysis, which considers factors such as force, friction, and the characteristics of the object. By using machine vision, determining attributes like weight, material, and rigidity, is very challenging and even more likely to exceed human capabilities to receive all that information.

Accuracy in grasp estimation largely depends on a vision system and planning. To start with the vision, the diversity in camera technologies introduces a variety of specifications, including variations in focal length, resolution, frame rate, depth range, and field of view. Higher-quality cameras generally come at increased costs. The camera needs to be operated under the use conditions following manufacturer guidelines in order to obtain optimal results. A further complication arises when positioning the camera on the robot's end effector. The camera frame can be movable together with the robotic arm, the transformation needs to be accurate calculation. Also, the camera should be appropriately aligned with the target. Otherwise, the robot may lack awareness of precise the target's location, especially in cluttered environments where the target may be obscured by other objects. For this reason, searching for the target and identifying the best viewpoint in such scenarios is important.

With the aim of accomplishing planning, obstacle avoidance and safety are critical considerations as well. Collisions can lead to failed grasp attempts. In scenarios where robots interact with humans, it is obligatory to detect humans and surrounding objects to devise safe operational paths, avoiding injury and damage. As well as trajectory, several trajectories could be feasible, selecting the optimal path is essential to minimize time and energy expenditure. Therefore, effective motion planning in grasp execution represents a significant challenge, requiring careful consideration of multiple factors to ensure efficient and safe robotic operations.

As for robustness, devices and methodologies are equally influenced. For instance, depth perception often fails under dim lighting conditions or when encountering surfaces that are highly reflective or transparent. Such conditions result in an inability to accurately determine the distance and coordinates of the target, leading to failed grasping attempts. In the same way, methodological challenges appear when certain parameters are either missing or faulty, resulting in ungraspable. Developing a robust system capable of adapting to variable conditions and mitigating uncertainties is a significant challenge, requiring sophisticated system design and control to ensure reliable performance in diverse scenarios.

Fig. 5. Shopping pick and pack challenge

In the European Robotic League (ERL) 2023 [33], the smart city challenges now require advanced robot capabilities along with human-robot interaction. The challenges include real-world problem tasks such as delivering coffee shop orders, through the door, taking the elevator, shopping pick and pack, and assisting a person in their home.

To give an example, in the shopping pick and pack episode, a mobile manipulator is required to receive an order from an operator. Then, the robot should autonomously pick the object from the shelf and place it in the delivery area. This episode entails tons of challenges, including path planning and motion planning to navigate the robot accurately through waypoints while avoiding obstacles such as ground furniture, shelves, and adjacent objects. One of the difficult tasks is to find the best viewpoint in the random placement of objects, and several shelves and shelf levels.

This mission has three levels of objects, ranging from easy to difficult, organized based on the complexity of their shape and location. The robot needs to recognise the order by detecting and classifying the target object among other items and environments. Once the target is identified, the robot should determine the grasping position to manipulate the object accurately. The stability of the grasp is significant too, as improper handling could lead to the object being dropped en route, causing a failed delivery.

IV. Future Research Trends

In the upcoming research directions, they are going to deal with those challenges. Rather than relying on preprogrammed methods, robots need perceptual capabilities to sense their surroundings and interact with them effectively. The enhancement of machine vision is requisite, as current performance levels can be further optimized to extensive information, covering high-definition RGB and accurate depth data.

Moreover, the integration of multimodal sensing has emerged as a prominent solution. This involves incorporating additional senses, such as force or tactile sensors on the gripper, to open new senses to the robot. The fusion of inputs from various sensors augments the information available concerning the target, enabling effective operations, including the identification of suitable grasping positions and forces, object characteristics, precise dimensions, and coordinates relevant to the robot's context, among others.

Various industries demand the deployment of manipulators in real-world scenarios such as pick and place, pick and sort, or pick and assembly, often involving multiple objects within cluttered environments. In other words, it can be stated that the current focus is on object-centric grasp synthesis [2]. Besides, effective methods characterized by both high accuracy and real-time speed are in demand, specifically in object detection, object pose estimation, and grasp estimation. Either the research could articulate the development of each module or all stages as an end-to-end system. The system design and integration, capable of seamlessly connecting the entire process, is included in future research.

As to methodology, grasp estimation comprises two primary approaches: traditional methods, which contribute significantly to grasp stability, and learning-based methods, which have recently demonstrated impressive performance in grasping tasks due to advancements in databases and

learning techniques [2]. The integration of these two domains holds great potential for developing effective and robust grasping methodologies, particularly in scenarios involving unknown environments and uncertainties. Such a fusion could leverage the strengths of both approaches, so that could offer opportunities for enhancing the efficacy of the robotic grasping system.

V. Conclusion

This comprehensive review provides insights grasp detection system, highlighting fundamentals and advancements, plus the evaluation by benchmark datasets. Accuracy, grasp success rate, and execution time are key metrics in this field. Additionally, ongoing challenges and outlined future research directions are presented. The research and development could be advanced in each stage or the entire system. A variety of combinations of methods are possible, YOLO series can be a detector module followed by the point pair features (PPF) module and then a 6DOF grasp estimation module as an instance. Several challenges are waiting for solutions, including handling a variety of objects, managing multiple items in cluttered environments, planning for obstacle avoidance and human safety, and dealing with uncertainties in real-world scenarios. Improvements in these areas will significantly benefit various applications that employ manipulators and mobile manipulators, such as in service industries, manufacturing, and hazardous missions to safely work with humans and reduce human involvement in repetitive and dangerous tasks.

References

[1] G. Du, K. Wang, S. Lian, and K. Zhao, "Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review," *Artif Intell Rev*, vol. 54, no. 3, pp. 1677–1734, Mar. 2021, doi: 10.1007/s10462-020-09888-5.

[2] H. Zhang, J. Tang, S. Sun, and X. Lan, "Robotic grasping from classical to modern: A survey," *arXiv preprint arXiv:2202.03631*, 2022.

[3] Z. Yin, Y. Li, J. Cai, and H. Lu, "Robotic Grasp Detection for Parallel Grippers: A Review," in *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*, IEEE, 2022, pp. 1184–1187.

[4] Y. Cong, R. Chen, B. Ma, H. Liu, D. Hou, and C. Yang, "A comprehensive study of 3-D vision-based robot manipulation," *IEEE Trans Cybern*, vol. 53, no. 3, pp. 1682–1698, 2021.

[5] Z. Xie, X. Liang, and C. Roberto, "Learning-based robotic grasping: A review," *Frontiers in Robotics and AI*, vol. 10. Frontiers Media S.A., 2023. doi: 10.3389/frobt.2023.1038658.

[6] A. Blank *et al.*, "6DoF pose-estimation pipeline for texture-less industrial components in bin picking applications," in *2019 European Conference on Mobile Robots (ECMR)*, IEEE, 2019, pp. 1–7.

[7] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object Detection in 20 Years: A Survey," May 2019, [Online]. Available: http://arxiv.org/abs/1905.05055

[8] V. Tadic *et al.*, "Perspectives of Realsense and ZED depth sensors for robotic vision applications," *Machines*, vol. 10, no. 3, p. 183, 2022.

[9] T.-Y. Lin et al., "Microsoft coco: Common objects in context," in Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740–755.

[10] A. Wang *et al.*, "YOLOv10: Real-Time End-to-End Object Detection," *arXiv preprint arXiv:2405.14458*, 2024.

[11] Y. Zhao *et al.*, "Detrs beat yolos on real-time object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16965–16974.

[12] W. Liu et al., "Ssd: Single shot multibox detector," in Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, Springer, 2016, pp. 21–37.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Adv Neural Inf Process Syst*, vol. 28, 2015.

[14] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[15] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154–6162.

[16] G. Marullo, L. Tanzi, P. Piazzolla, and E. Vezzetti, "6D object position estimation from 2D images: a literature review," *Multimed Tools Appl*, 2022, doi: 10.1007/s11042-022-14213-z.

[17] BOP, "BOP: Benchmark for 6D Object Pose Estimation," BOP. Accessed: Nov. 09, 2023. [Online]. Available: https://bop.felk.cvut.cz/home/

[18] M. Sundermeyer *et al.*, "BOP Challenge 2022 on Detection, Segmentation and Pose Estimation of Specific Rigid Objects," Feb. 2023, [Online]. Available: http://arxiv.org/abs/2302.13075

[19] Y. He, H. Huang, H. Fan, Q. Chen, and J. Sun, "Ffb6d: A full flow bidirectional fusion network for 6d pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3003–3013.

[20] S. Ahmad, "Robotic assembly, using RGBD-based object pose estimate & grasp detection," 2020.

[21] A. Zeng *et al.*, "Multi-view self-supervised deep learning for 6D pose estimation in the Amazon Picking Challenge," in *Proceedings - IEEE International Conference on Robotics and Automation*, Institute of Electrical and Electronics Engineers Inc., Jul. 2017, pp. 1386–1393. doi: 10.1109/ICRA.2017.7989165.

[22] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *International Journal of Robotics Research*, vol. 34, no. 4–5, pp. 705–724, Apr. 2015, doi: 10.1177/0278364914549607.

[23] BOP, "BOP: Benchmark for 6D Object Pose Estimation," BOP. Accessed: Nov. 09, 2023. [Online]. Available: https://bop.felk.cvut.cz/home/

[24] Meta AI Research, "Robotic Grasping," Papers With Code. Accessed: Nov. 14, 2023. [Online]. Available: https://paperswithcode.com/task/robotic-grasping

[25] S. Ainetter and F. Fraundorfer, "End-to-end trainable deep neural network for robotic grasp detection and semantic segmentation from rgb," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2021, pp. 13452–13458.

[26] S. Kumra, S. Joshi, and F. Sahin, "Antipodal robotic grasping using generative residual convolutional neural network," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2020, pp. 9626–9633.

[27] F.-J. Chu, R. Xu, and P. A. Vela, "Real-world multiobject, multigrasp detection," *IEEE Robot Autom Lett*, vol. 3, no. 4, pp. 3355–3362, 2018.

[28] H. Cao, G. Chen, Z. Li, J. Lin, and A. Knoll, "Lightweight convolutional neural network with Gaussian-based grasping representation for robotic grasping detection," *arXiv preprint arXiv:2101.10226*, 2021.

[29] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11444–11453.

[30] H.-S. Fang *et al.*, "Anygrasp: Robust and efficient grasp perception in spatial and temporal domains," *IEEE Transactions on Robotics*, 2023.

[31] B. S. Zapata-Impata, C. Mateo Agulló, P. Gil, and J. Pomares, "Using geometry to detect grasping points on 3D unknown point cloud," 2017.

[32] B. S. Zapata-Impata, P. Gil, J. Pomares, and F. Torres, "Fast geometry-based computation of grasping points on three-dimensional point clouds," *Int J Adv Robot Syst*, vol. 16, no. 1, Jan. 2019, doi: 10.1177/1729881419831846.

[33] euRobotics, "ERL MK Smart City Challenge," euRobotics. Accessed: Nov. 12, 2023. [Online]. Available: https://eu-robotics.net/2023-09-erl-mk-smart-city-challenge/

The comprehensive review of vision-based grasp estimation and challenges

Mansakul, Thanavin

2024-08-28

Mansakul T, Tang G, Webb P. (2024) The comprehensive review of vision-based grasp estimation and challenges. In: 29th International Conference on Automation and Computing (ICAC), 28-30 August 2024, Sunderland, United Kingdom

https://doi.org/10.1109/icac61394.2024.10718793