



Intelligent gripping: Application study of vision-supported robotic arm gripping and placement

Lulu LI^{1,2} , Hichem SNOUSSI¹, Abel CHEROUAT², and Tian WANG^{3,4} *

¹ UR-LIST3N, University of Technology of Troyes, Troyes, France

² UR-GAMMA3, University of Technology of Troyes, Troyes, France
{lulu.li, hichem.snoussi, abel.cherouat}@utt.fr

³ The School of Artificial Intelligence and SKLSDE, Beihang University, China

⁴ Zhongguancun Laboratory, Beijing 100083, China
wangtian@buaa.edu.cn

Abstract. Computer vision and robotics in assistive technology have made significant progress. However, there are still great challenges, such as how to improve the success and stability of grasping in the development of home assistance and medical devices. In order to more intelligently help technicians in the company factory or people with disabilities in their daily activities, this paper summarizes our latest research results in computer vision and robotics in grasping prediction, target recognition, localization, and robotic arm control. The described method joins object detection and grasping attitude to predict the optimal grasping points and gestures for all targets in the scene. The given method not only achieves globally effective grasping prediction, but additionally explores the potential of these techniques to improve quality of life and work efficiency. In particular, deep learning models are combined to quickly and accurately recognize and localize various complex target objects, providing reliable visual support for grasping and placing tasks, as well as improving the autonomous and precise performance of grasping tasks.

Keywords: Intelligent gripping · Visual support · Robotic arm · Grasping prediction

1 Introduction

With the rapid development and continuous updating iterations of computer vision and robotics, these technologies have made remarkable progress in the application of assistive technology [5, 13, 15]. In recent years, computer vision techniques such as SSD (Single Shot MultiBox Detector) [7], Faster R-CNN [2], Mask R-CNN [6], YOLO (You Only Look Once) [11] target detection algorithms have enabled real-time, high-precision object recognition and classification. The application of these techniques has enabled robots to quickly and accurately recognize and localize various target objects, providing reliable visual support for

* Corresponding author

grasping and manipulation tasks. A number of works evaluate grasp by combining deep learning models such as Graspnet [3], AnyGrasp [4], GG-CNN [11], GQ-CNN [16], Dex-Net 3.0, Dex-Net 4.0 [1], RGBgrasp [12], and so on. These works predict object-independent and pixel-orientated grasping poses, or grasping estimates for single-target objects in the scene, or add suction-based grippers for selective grasping, or even predict only rectangular grasping regions instead of accurate 6-dimensional grasping poses. While these methods can predict the optimal grasping point and achieve a certain grasping success rate. However, the application of these methods is limited in most practical scenarios, i.e., only recognizing and grasping a single object or grasping indiscriminately without autonomy in complex scenarios, which fails to satisfy the needs of assistive technologies in daily life.

To completely change this situation, this paper proposes a generalized vision-supported robot grasping prediction method [8–10]. Which is based on advanced algorithms [3,14] for joint vision and grasping prediction and aims to improve the efficiency and accuracy of grasping in multiple complex environments. Overall, this paper explores in detail the latest research results on the collaboration of computer vision with robotics applied in assistive technologies performing general grasping and placing tasks. Therefore the motivation and contribution of this paper are:

- Create and present effective target grasping prediction methods that can be used as an assistive industrial process.
- 3D dection-driven masking methodology for visually-supported global target grasping prediction in order to make recognizing and grasping targets more comprehensive and autonomous.
- Joint YOLO and Graspnet model for grasping and placing tasks.

2 Research Method and Experiments

2.1 Research Method

Intelligent grasping is the method of target detection as a visual enhancement for grasping evaluation, as shown in the method flowchart in Fig. 1. The whole approach is divided into two main phases:

The first stage is the target detection based on YOLOv7 and point cloud preprocessing. Firstly, we input the RGB channels of the RGBD image for target localization, and obtain the bounding boxes, masks and the grasping order of the targets of the whole scene for all the targets in the scene. The grasping order is divided according to the target categorization, and the grasping order of different categories is artificial, i.e., we can predefine the categories of target objects we want to grasp. Next, we input the depth (point cloud) image and align it with the RGB map to obtain a lossless complete point cloud of the whole scene, and crop the complete point cloud to obtain the final workbench spatial point cloud according to the necessity of data processing. At last the target regions

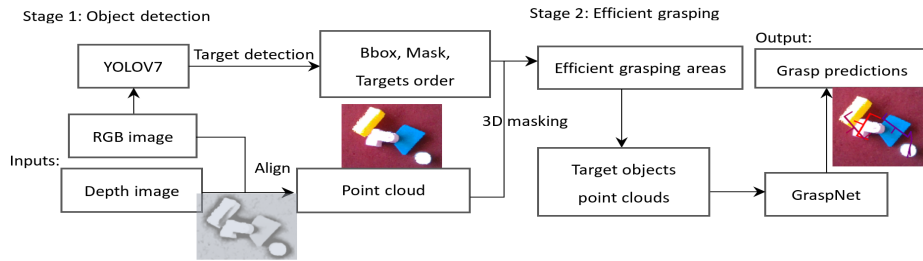


Fig. 1: Framework of effective global target grasping prediction method.

are cropped on the point cloud using our proposed detection-driven 3D masking approach.

The purpose of detection-driven 3D masking is to obtain target-level regions for valid grasping predictions. It utilizes the bounding boxes and segmentation masks of the target recognition to determine the location of the target and prioritizes the order of the grasping predictions by the confidence score. Where the grasping predictions correspond to the target categories and locations. In this way, the point clouds of all target objects, the grasping prediction regions and the corresponding target categories and locations are finally obtained.

The second stage of intelligent grasping is to evaluate the target-level grasping gestures based on the point clouds of the target regions, categories and locations acquired in the first stage. GraspNet is utilized for the estimation of grasping predictions. For each target object, a two-finger grasping clip is predicted for the optimal grasping pose, and then the grasping order of all the targets in the scene is obtained based on the grasping prediction confidence, so as to provide the optimal grasping solution for all target objects at the same time. This is distinguished from traditional methods that predict only one optimal grasping pose for the whole scene; we are detecting and obtaining effective grasping poses for all targets in the scene.

Through the close coordination of these two stages, efficient and accurate intelligent grasping can be realized. This approach not only improves the robot's ability to operate in complex environments, but also provides reliable technical support for the practical application of assistive technologies. Whether in industrial automation, home assistance or medical device development, this intelligent grasping method demonstrates its broad application prospects and great potential.

2.2 Experiments

We performed gripping experiments using a six-axis robotic arm, EPSON C4-A601S, and a depth camera, Realsense D435i. The robotic arm is equipped with a two-finger inflatable gripper jaws that ranges from 18 mm to 35 mm in width. The depth camera is mounted at the wrist of the robotic arm. We set the robotic arm to take pictures at a fixed position to obtain the scene image and point cloud

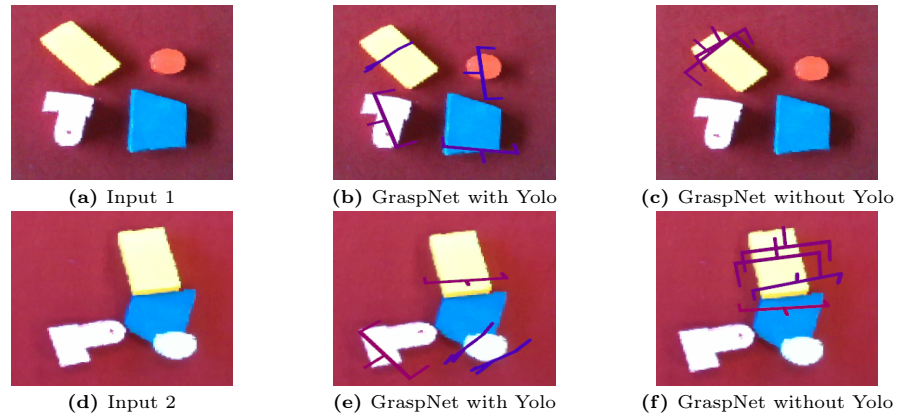


Fig. 2: Comparison of methods with and without visual support.

of the workspace. The target location and its grasping area are first recognized based on the image, then the grasping pose is predicted for all targets in the scene, and finally the targets are grasped based on the categories.

3 Results and Discussion

As the results shown in Fig. 2, (a-c) show the upright scene results, where (a) shows the input image, (b) indicates that the method described in the article estimates the grasping poses of all the targets with visual, and (c) gives the grasping pose estimation for the method without visual support. Similarly (d-f) give the occlusion scene results. The comparison results show that our method is more effective and can estimate the grasping poses of all targets simultaneously, with high precision in fast time, while the method without visual support can only estimate the grasping poses of partial targets, or even the grasping poses beyond the target area.

In conclusion, our described research has achieved excellent results in grasping and placing tasks. In the future, we will integrate the large language model to realize the commanded robot-controlled grasping and further overcome the difficulties to broaden the applications in assistive technology.

4 Conclusion

This research provides a method for efficient grasping predictions for a wide range of assisted robotic grasping placement tasks. The approach skillfully combines computer vision with a robotic arm for dealing with the problem of autonomous, intelligent grasping of object targets. This not only improves the intelligence level of the robot, but also greatly enhances the user's convenience and autonomy. The wide application of this effective grasping prediction method will show its potential and value in more scenarios.

Acknowledgements

This work was supported in part by China Scholarship Fund.

References

1. Dong, M., Zhang, J.: A review of robotic grasp detection technology. *Robotica* pp. 1–40 (2023)
2. Duan, J., Zhuang, L., Zhang, Q., Qin, J., Zhou, Y.: Vision-based robotic grasping using faster r-cnn-grcnn dual-layer detection mechanism. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture* p. 09544054241249217 (2024)
3. Fang, H.S., Gou, M., Wang, C., Lu, C.: Robust grasping across diverse sensor qualities: The graspnet-1billion dataset. *The International Journal of Robotics Research* **42**(12), 1094–1103 (2023)
4. Fang, H.S., Wang, C., Fang, H., Gou, M., Liu, J., Yan, H., Liu, W., Xie, Y., Lu, C.: Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics* (2023)
5. Kaulage, A., Agrawal, S., Jagdale, S., Salunkhe, P., Salunkhe, R.: Yolo-driven robotic system for automated object singulation. In: *2024 International Conference on Inventive Computation Technologies (ICICT)*. pp. 1800–1805 (2024). <https://doi.org/10.1109/ICICT60155.2024.10544827>
6. Kijdech, D., Vongbunyong, S.: Manipulation of a complex object using dual-arm robot with mask r-cnn and grasping strategy. *Journal of Intelligent & Robotic Systems* **110**(3), 103 (2024)
7. Kwolek, B.: Continuous hand gesture recognition for human-robot collaborative assembly. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 2000–2007 (2023)
8. Li, L., Cherouat, A., Snoussi, H., Hu, R., Wang, T.: Detection-driven 3d masking for efficient object grasping. *The International Journal of Advanced Manufacturing Technology* **129**(9), 4695–4703 (2023)
9. Li, L., Cherouat, A., Snoussi, H., Wang, T.: Grasping with occlusion-aware ally method in complex scenes. *IEEE Transactions on Automation Science and Engineering* (2024)
10. Li, L., Cherouat, A., Snoussi, H., Wang, T., Lou, Y., Wu, Y.: Vision-based deep learning for robot grasping application in industry 4.0. In: *Technological Systems, Sustainability and Safety* (2024)
11. Li, Z., Xu, B., Wu, D., Zhao, K., Chen, S., Lu, M., Cong, J.: A yolo-ggcnn based grasping framework for mobile robots in unknown environments. *Expert Systems with Applications* **225**, 119993 (2023)
12. Liu, C., Shi, K., Zhou, K., Wang, H., Zhang, J., Dong, H.: Rgbgrasp: Image-based object grasping by capturing multiple views during robot arm movement with neural radiance fields. *IEEE Robotics and Automation Letters* (2024)
13. Nguyen, N.K., Nguyen, V.K., Pham, V.H., Pham, V.M., Pham, V.N., et al.: Novel design of a robotic arm prototype with complex movements based on surface emg signals to assist disabilities in vietnam. *International Journal of Advanced Computer Science & Applications* **15**(3) (2024)

14. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7464–7475 (2023)
15. Xu, W., Huff, T., Ye, S., Sanchez, J.R., Rose, D., Tung, H., Tong, Y., Hatcher, J., Klein, M., Morales, E., et al.: Virtual reality-based human-robot interaction for remote pick-and-place tasks. In: Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction. pp. 1148–1152 (2024)
16. Zheng, T., Wang, C., Wan, Y., Zhao, S., Zhao, J., Shan, D., Zhu, Y.: Grasping pose estimation for robots based on convolutional neural networks. *Machines* **11**(10), 974 (2023)