A Comparative Analysis of Object Detection Methods for Robotic Grasping

Nikita Kolin
Intelligent Robotics Department, Kazan Federal University, 35 Kremlevskaya St., Kazan, 420008, Russian Federation

Elvira Chebotareva
Intelligent Robotics Department, Kazan Federal University, 35 Kremlevskaya St., Kazan, 420008, Russian Federation
Email: nikitakolin7@gmail.com, Elvira.Chebotareva@kpfu.ru

Abstract

The objects grasping is one of the fundamental robotic problems. Accurate and efficient real-time object detection is crucial for successful grasping in robots equipped with monocular vision. Deep machine learning has made significant progress in solving problems of object detection and image segmentation. At the same time, classical computer vision methods do not lose their relevance and can also be used for these tasks. In this research, we conduct a comparative analysis of the effectiveness the YOLOv8-seg neural network model versions for solving the image segmentation problem with classical segmentation methods. The obtained results allowed us to formulate some recommendations on the choice of a particular method for object detection depending on the surrounding environment conditions.

Keywords: Robotic grasping, Robotic arm, Object detection, Object segmentation

1. Introduction

Robotic grasping is one of the fundamental tasks of robotics, relevant to all its branches [1], including industrial [2], service [3] and social robotics [4]. Grasping can be particularly important in collaborative robotics, when a robot and a person use common tools or perform joint assembly [5].

Obtaining information about the shape of the captured objects is an important stage in the implementation of grasping [6]. The shape of objects can be determined using various sensors, including cameras and lidars. At the same time, solutions that allow detect grasping objects using monocular cameras are also of practical interest. Deep machine learning has made significant progress in solving problems related to object detection [7] and image segmentation [8]. However, classical computer vision methods do not lose their relevance and can also be used for these tasks. In this paper, we present the results of our experiments aimed at comparing classical object detection methods with object detection and segmentation using the YOLOv8-seg models family. The aim of our research is to evaluate the accuracy and effectiveness of various methods for detecting object contours in robotic grasping.

2. Related Work

In the past few years, there has been a lot of research reported on robotic grasping problems. Some approaches to solving the problem of robotic grasping are presented in the review [9]. The overview of different approaches to computer vision-based robotic grasping presented in [10] and [11]. The effectiveness of deep machine learning in object detection on RGB-D images for robotic grasping is well demonstrated in the work [12]. At the same time, detecting objects in RGB images seems to be a much more difficult task, since such images do not contain depth information. Examples of works exploring the challenges of robotic grasping using monocular vision and RGB images include works [13] and [14].

Often, the effectiveness of grasping area detection methods is evaluated using one of the popular datasets, such as Cornel dataset [15] and Jacquard dataset [16]. The use of publicly available datasets allows comparing the effectiveness of different approaches. According to the data from the “Paper with Code” platform [17] the accuracy of various grasping methods based on machine learning, estimated with the help of Cornel Dataset was made from 88% in 2016 for Multi-Modal Grasp Predictor [18] to 98.2% in 2021 for Ainetter and Fraundorfer CNN-based architecture [19] and 98.9% for Efficient Grasping model by Cao et. al [20]. At the same time, the accuracy estimates on the Jacquard dataset for the Ainetter and Fraundorfer model [19] was 92.95%, and for the Efficient Grasping model by Cao et. al [20] was 95.6%.

As it can be seen, estimates of the accuracy of different methods may vary from the quality of the test dataset. In addition, although public datasets provide an overview of the effectiveness of a given method, these datasets do not always take into account the peculiarities of objects and

environments encountered in real cases. Therefore, datasets that are close to the real conditions of a particular task are always of special interest. In this research work we will make a comparative analysis of various methods of objects detection on own dataset, adapted to the real object grasping task. As an example of a model that provides acceptable accuracy along with the high inference speed and usability we use the YOLOv8-seg model for object segmentation. The YOLOv8-seg model is based on variant of the U-Net architecture [21]. The YOLO models family based on hybrid convolutional neural network architecture [22], [23] and often used in robotics.

3. Materials and Methods

3.1. *Object detection problem statement*

We model a situation in which the working cell of a robotic manipulator is equipped with a single monocular camera fixed above the working area. During the experiment, the camera does not change its position. This camera transmits image of the table area, on which three objects can be placed at random: pliers, screwdrivers and PIR sensor. We assume that the objects are homogeneous, so we can assume that the center of mass of each object coincides, or is close, with its geometric center. Therefore, the optimal gripping points must also be near the geometric center of each object.

Thus, in the context of the task of detecting objects for their robotic grasping, we need to determine the object class (pliers, screwdriver, or PIR sensor), the contours of the object, the geometric center of the object (the center of mass of the contour) and the orientation of the object. Note that in order to calculate the center and orientation of an object in this case it is enough to define its contour.

3.2 Datasets

For our experiments, we have prepared a set of 100 images with resolution 640×480 pixels containing three different target objects: pliers, screwdriver and PIR sensor. The choice of objects is determined by the context of collaborative tasks in which it is planned to use this robot in the future. In addition, these objects have different shapes and colors, which can help identify potential issues with specific object detection methods. The images were taken under various lighting conditions and against different backgrounds.

However, to test the hypotheses of our study, we also prepared 20 images with backgrounds that differed from those in the training dataset images. A total of 165 annotated images were prepared for model training, with 132 included in the training dataset, 17 images in the validation dataset, and 16 in the test dataset. To improve the quality of training, 236 images derived from the original 132 training dataset images were added to the training dataset through rotation and perspective distortion.

3.3 Object detection methods

For our experiments, we used the YOLOv8n-seg and YOLOv8s-seg models pre-trained on the COCO dataset by Microsoft. YOLOv8n-seg and YOLOv8s-seg models trained on a set of 368 images of size 640×480.

To extract the object's contours we used two classical methods. The first method is based on the conversion to HSV space and channel thresholding. And the second method uses the Canny edge detector, as well as auxiliary erosion and dilation transformations. The classification of the extracted contours is based on a single feature – the aspect ratio of the bounding box.

3.2. *Evaluation*

Each method of object detection and segmentation was evaluated using two commonly used metrics: the Jaccard index (1) – also known as the IoU (Intersection over Union) metric, and the Dice coefficient (2) – also known as the F1-score metric. These metrics are based on the formulas (1) и (2). In formulas (1) and (2) A is the set of pixels belonging to the real object in the image, B is the set of pixels of the result of segmentation.

$$IoU = \frac{A \cap B}{A \cup B} \quad (1)$$

$$F1 = 2\frac{|A \cap B|}{|A| + |B|} \quad (2)$$

For each method, we also measured the inference time per frame in the test video and calculated the inference speed as the number of processed frames per second (FPS).

4. Experiment Results and Discussion

At the first stage of the experiment, we evaluated the average values of IoU and F1-score for 100 images from our dataset. Then, at the second stage of our experiment, we conducted the evaluation using 20 images with significantly different backgrounds compared to the backgrounds of the images in the training dataset. The results of these stages are presented in Table 1 and Table 2.

Table 1. The object detection methods accuracy obtained on images with conditions similar to the training dataset.

Method	IoU	F1-score
Thresholding	0.4112	0.4658
Canny edge detector	0.1720	0.2257
YOLOv8n-seg	0.8248	0.8811
YOLOv8s-seg	0.8612	0.9185

Table 2. The object detection methods accuracy obtained on images with conditions differing from the training dataset.

Method	IoU	F1-score
Thresholding	0.8047	0.8771
Canny edge detector	0.0252	0.0469
YOLOv8n-seg	0.3186	0.3477
YOLOv8s-seg	0.2299	0.2535

Thus, in the case where the images background was present in the training dataset, the best result was obtained by YOLOv8s-seg and YOLOv8n-seg models. However, the YOLOv8n-seg model showed an expectedly lower result compared to YOLOv8s-seg model. The thresholding-based method showed significantly lower results than the YOLOv8-seg models, but exceeded the Canny edge detection method. Examples of object detection and segmentation using models YOLOv8n-seg and YOLOv8s-seg are presented in Fig. 1 and Fig. 2.

Fig. 1. The example of successful recognition of all objects using the YOLOv8n-seg model.

However, in cases where the background on which the objects are located differed from the backgrounds in the training images, the method based on threshold processing of channels in the HSV color space showed better results compared to both the YOLOv8s-seg and YOLOv8n-seg models, as well as the method based on Canny edge detector. Fig. 3 presents an example of object detection and segmentation on an image with a background different from the training dataset. In this instance, the YOLOv8s-seg and YOLOv8n-seg models failed to detect object contours on the complex background not included in the training dataset. The method based on Canny edge detector also did not give positive results.

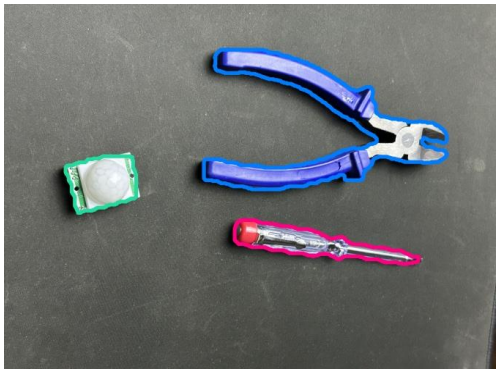


Fig. 2. The example of successful recognition of all objects using the YOLOv8s-seg model.

Fig. 3. The example of object contour detection using a classic threshold-based method in cases where YOLOv8n-seg and YOLOv8s-seg models failed.

Table 3 presents the results of FPS evaluation on the test video for all the considered methods. The best inference speed result belongs to the method based on the Canny edge detector, followed by the threshold-based method in second place, and the YOLOv8s-seg model demonstrated the worst result.

Table 3. Results of FPS evaluation for the considered methods.

Method	FPS
Thresholding	21.8
Canny edge detector	32.6
YOLOv8n-seg	4.2
YOLOv8s-seg	2.1

5. Conclusion

The obtained results allowed us to formulate some recommendations on the choice of a particular method for object detection depending on the surrounding environment conditions. In cases where the scene is not overloaded with a large number of objects, the environmental conditions remain stable, and the objects significantly differ from the background and have relatively homogeneous texture, classical methods can be quite successfully applied to extract the contours of the target objects. Moreover, in contrast to machine learning methods, these methods do not require the preliminary collection of a large amount of training data and annotation, and they also characterized by higher inference speed. However, if the environmental conditions change dynamically, these methods become practically inapplicable. In this case, the optimal solution would be to train a previously pre-trained neural network model. However, the training dataset must be close to the real conditions in which recognition is planned, in particular, it is necessary that the background on which the target objects are located is the same as in real

conditions. Otherwise, the quality of detection and segmentation is significantly reduced.

Acknowledgements

This paper has been supported by the Kazan Federal University Strategic Academic Leadership Program ("PRIORITY-2030").

References

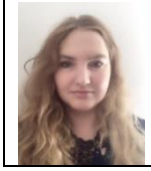
1. R. Hodson, How Robots are Grasping the Art of Gripping. Nature, vol 557, 2018, pp. 23–25.
2. M. Mustafin, E. Chebotareva, E. Magid, H. Li and E. Martinez-Garcia, Features of Interaction Between a Human and a Gestures-Controlled Collaborative Robot in an Assembly Task: Pilot Experiments, The 2023 International Conference on Artificial Life and Robotics (ICAROB), Oita, Japan, 2023, pp. 158–162.
3. K. Khusnutdinov, A. Sagitov, A. Yakupov, R. Lavrenov, E. A. Martinez-Garcia, K.-H. Hsia and E. Magid, Household Objects Pick and Place Task for AR-601M Humanoid Robot. Lecture Notes in Computer Science, 11659, 2019, p. 139-149.
4. K. Khusnutdinov, A. Sagitov, A. Yakupov, R. Meshcheryakov, K.-H. Hsia, E. A. Martinez-Garcia and E. Magid, Development and Implementation of Grasp Algorithm for Humanoid Robot AR-601M. Proceedings of the 16th International Conference on Informatics in Control, Automation and Robotics, 2019, pp. 379-386.
5. M. Mustafin, E. Chebotareva, H. Li and E. Magid, Experimental Validation of an Interface for a Human-Robot Interaction Within a Collaborative Task. In International Conference on Interactive Collaborative Robotics (ICR 2023), Baku, Azerbaijan, 2023, pp. 23–25.
6. R. Abdulganeev, R. Lavrenov, R. Safin, Y. Bai and E. Magid, Door handle detection modelling for Servosila Engineer robot in Gazebo simulator. Siberian Conference on Control and Communications, (SIBCON 2022), 2022, p. 1-4.
7. V. Myrzin, T. Tsoy, Y. Bai, M. Svinin and E. Magid, Visual data processing framework for a skin-based human detection. In Interactive Collaborative Robotics: 6th International Conference, ICR 2021, Proceedings 6, September 2021, pp. 138-149.
8. Y. Yu, C. Wang, Q. Fu, R. Kou, F. Huang, B. Yang, T. Yang and M. Gao, Techniques and Challenges of Image Segmentation: A Review. Electronics 12, no. 5, 2023, pp. 1199.
9. L. Wang, Z. Zhang, J. Su and Q. Gu, Robotic Autonomous Grasping Technique: A Survey. 2021 5th Asian Conference on Artificial Intelligence Technology (ACAIT), Haikou, China, 2021, pp. 287–295.
10. K. Kleeberger, R. Bormann, W. Kraus, and M. F. Huber Kleeberger, A Survey on Learning-Based Robotic Grasping. Current Robotics Reports 1, 2020, pp. 239–249.
11. Z. Xie, X. Liang and R. Canale, Learning-Based Robotic Grasping: A Review. Frontiers in Robotics and AI 10, 2023, p. 1038658.
12. Q. Zhang, J. Zhu, X. Sun and M. Liu, HTC-Grasp: A Hybrid Transformer-CNN Architecture for Robotic Grasp Detection. Electronics 12, no. 6, 2023, p. 1505.
13. C. Veiga Almagro, R.A. Muñoz Orrego, Á. García González, E. Matheson, R.M. Prades, M. Di Castro and M.F. Pérez, (MARGOT) Monocular Camera-Based Robot Grasping Strategy for Metallic Objects. Sensors 23, no. 11, 2023, p. 5344.
14. W. Prew, T. Breckon, M. Bordewich and U. Beierholm, Improving Robotic Grasping on Monocular Images Via Multi-Task Learning and Positional Loss, 2020, https://arxiv.org/abs/2011.02888
15. I. Lenz, H. Lee and A. Saxena, Deep learning for detecting robotic grasps. The International Journal of Robotics Research 34, 2013, pp. 705–724.
16. A. Depierre, E. Dellandréa and L. Chen, Jacquard: A Large Scale Dataset for Robotic Grasp Detection. 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 2018, pp. 3511–3516.
17. Paper with Code, https://paperswithcode.com/
18. S. Kumra and C. Kanan, Robotic grasp detection using deep convolutional neural networks. 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 2017, pp. 769–776.
19. S. Ainetter and F. Fraundorfer, End-to-end Trainable Deep Neural Network for Robotic Grasp Detection and Semantic Segmentation from RGB, 2022, https://arxiv.org/abs/2107.05287
20. H. Cao, G. Chen, Z. Li, J. Lin and A. Knoll, Lightweight Convolutional Neural Network with Gaussian-based Grasping Representation for Robotic Grasping Detection, 2021, https://arxiv.org/abs/2101.10226
21. Ultralytics YOLOv8 Tasks, Ultralytics Inc, https://docs.ultralytics.com/tasks/
22. J. Redmon, S. Divvala, R. Girshick and A. Farhadi, You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 779–788.
23. J. Terven and D. Cordova-Esparza, A Comprehensive Review of YOLO: From YOLOv1 and Beyond, 2023, https://arxiv.org/abs/2304.00501.

Authors Introduction

Mr. Nikita Kolin

He received his Bachelor's degree in Kazan Federal University, Russia, in 2022. Currently he is a Master's student at the Institute of Information Technologies and Intelligent Systems, Kazan Federal University, Russia.

Assistant Professor Elvira Chebotareva



She received her PhD in physics and mathematics from Kazan Federal University. She is currently an assistant professor in Laboratory of Intelligent Robotic Systems (LIRS) at Kazan Federal University, Russia.