August 25, 2021

Analysis of RGB-D images through Computer Vision for robot grasping of industrial parts

Daniel Horcajo de la Cruz

Master Thesis. Universidad Pontificia de Comillas

Master in Smart Industry

Tutors: Ignacio de Rodrigo Tobías, Álvaro Jesús López López

Abstract

The development of new technologies in recent years has brought with it a high level of progress in areas such as robotics. In turn, the growing interest in increasingly flexible and robust algorithms has led to the emergence of new types of artificial intelligence techniques, including, among others, object detection models. Despite these advancements, the use of robotic arms currently found in logistics and manufacturing centers is still limited by the type of gripper they are equipped with, thus restricting their gripping capabilities.

This project seeks to explore different solutions to this problem by developing an algorithm to overcome this limitation. Thus, by using color and depth cameras, in combination with traditional and artificial vision algorithms, it is intended to provide an ambidextrous robot with the ability to grasp any of the different types of parts found in a specific set of automotive components. These grips can be carried out by means of a parallel or suction cup type gripper, depending on the morphological characteristics and texture of the part to be gripped. As a result, the gripping capacity of the robotic arm is enhanced, providing it with the flexibility to perform grasping operations without being limited by a particular type of gripper.

KEYWORDS

Grasp, YOLO, Dex-Net, Computer Vision, grip, suction cup, depth camera, RGB-D

I. INTRODUCTION

Using a robotic arm capable of providing a sufficiently robust grip for a wide variety of objects is currently a major challenge present in a multitude of domains, ranging from online retail logistics to manufacturing processes. One of the existing limitations is the difficulty in making a single type of gripper capable of grasping objects with disparate geometries and textures in a sufficiently robust manner, leading to the need for multiple types of grippers depending on their application.

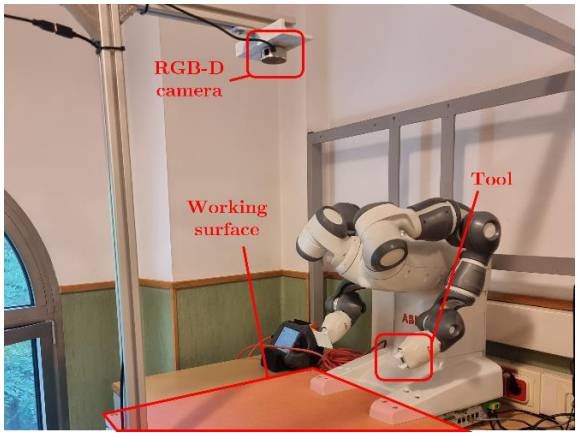

Figure 1. Setup of the YuMi robot in the laboratory at ICAI's Institute for Research in Technology.

Thus, this project seeks to explore different ways to solve this problem and will be carried out in collaboration with the Institute of Technological Research of the Universidad Pontificia Comillas (ICAI). To this end, the project seeks to combine disciplines such as robotics and artificial vision to find an algorithm that allows a robotic arm to pick up any part of a set of automotive components, using a single gripper (also known as universal picking). This process will be carried out through a set of techniques ranging from

computer vision to the use of depth imaging.

To achieve this goal, the configuration shown in **Figure 1** is arranged. The robot used will be an ABB YuMi - IRB 14000, in front of which there is a work surface on which the different parts to be grasped will be positioned. The detection of these parts will be carried out through an RGB-D camera capable of providing both a color image (RGB) and a depth map (Depth) that will allow the measurement of the distance from its sensor to each point in the scene.

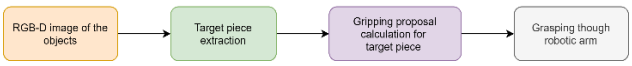

Figure 2. Information flow throughout the project.

The flow of information throughout the project is as shown in **Figure 2**. Using the RGB and depth images captured by the camera, the aim is to detect and extract the target part, on which a gripping proposal will be calculated – either by means of a parallel gripper or by means of a suction cup. This grasping proposal will finally be sent to the YuMi robot, which will be responsible for executing it and grasping the part. It is important to mention that the third stage of the information flow will explore the possibility of using the Dex-Net GQ-CNN [3] to compute a grasp on an already detected part.

The set of parts to be used in the project is known and already fixed, and is made up of different automotive components that are regularly handled in an industrial environment. Hence, the development context of the project will revolve around the feasibility of applying the solution found in such an environment, which will impose working conditions such as light changes or a limited number of part types.

II. SOLUTION

The flow of information throughout the project previously shown in **Figure 2** consists of three different phases: capturing the RGB-D image through a depth camera, detecting and capturing the target part, and calculating the proposed grip on it.

A. RGB-D image capture

The RGB-D cameras available are the D435 [1] and L515 [2] models from Intel's RealSense line. Both cameras will undergo a calibration method that will make it possible to approximate their respective intrinsic parameters in order to eliminate the different types of distortions that their images may suffer [6]. In addition, given that the former uses stereoscopic vision as a method of measuring the depth of the scene – compared to the LiDAR technology of the latter – a quality analysis of the depth image provided by each of them will be conducted.

Upon completion of this analysis, it is concluded that the depth image provided by the LiDAR L515 model is superior to that of the D435 model, offering an actual accuracy of fewer than 3 millimeters compared to the 10 millimeters of its competitor. It is also observed that, despite the post-processing filters used to improve the quality of the depth image in both cameras, the LiDAR model shows much fewer fluctuations in the depth images produced between shots than its stereoscopic analog. The L515 model is therefore chosen as the main RGB-D camera to be used throughout the project.

B. Extraction of the target piece

Throughout the project, various techniques oriented to image object detection and extraction are explored using the Python OpenCV library [4]: depth segmentation, background subtraction, and object detection using convolutional neural networks (YOLOv3) [5].

The depth segmentation method is the most straightforward of the three and consists of extracting from the image everything whose distance to the camera (positioned as indicated in **Figure 1**) is less than the distance to the workbench. Background subtraction, on the other hand, compares an image where the parts appear on the workbench with another image where the workbench is empty. This way, the difference between the two images provides a mask capable of locating the one thing that changes between the two shots – i.e., the workpieces. Although these two methods provide adequate results under controlled laboratory conditions, both have major limitations that make their use in an industrial environment unfeasible – such as low robustness to changes in lighting conditions or low accuracy.

Consequently, it is decided to explore the possibility of using object detection algorithms such as YOLOv3. For this purpose, a custom dataset consisting of the 46 available car components will be created and will then be used to train one specific model for each type of component of the set – thus only being able to detect said type of component. Since the types of components to be detected will be a known parameter beforehand, this approach allows to reduce the number of false positives that may appear, compared to using a single model capable of detecting any type of

component.

The masks obtained by each of these three methods are shown in **Figure 3**. While the quality of the background subtraction mask is highly detailed compared to depth segmentation, the slightest change of the background with respect to the original background image used by the algorithm is capable of completely invalidating any detection capability. On the other hand, it should be noted that YOLO, despite offering a bounding box instead of a detailed mask around the part, makes it possible to detect the part itself, and not a change in the image conditions, as is the case with the other two methods. Additionally, in order to obtain a more detailed mask out of the YOLO detections, depth segmentation is performed on the crop of its bounding box.

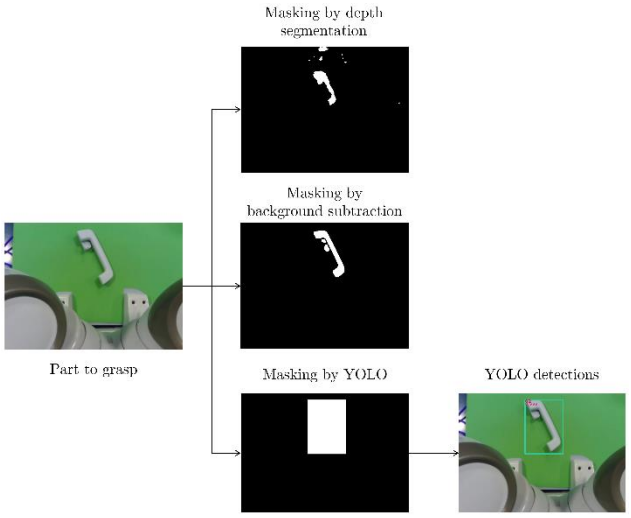

Figure 3. Comparison of the output masks after the implementation of depth segmentation, background subtraction, and YOLO methods.

C. Calculation of grip proposals

Since the YuMi robot has two different arms, the possibility of using two types of grippers to grasp parts according to their morphology and texture is explored: using a parallel gripper on parallelepiped-type parts, and using a suction cup for parts with flat surfaces.

Given a depth image of the scene and a binary mask of the target part, we explore the possibility of implementing the Dex-Net GQ-CNN [3] on the parts of the custom dataset. However, as shown in **Figure 4**, due to an obtained confidence score lower than 5% for most of the proposed grips, as well as the fact that these grips are always perpendicular to the work plane and not to the part itself, its use is discarded and it is decided to develop a custom algorithm for each type of tool available – parallel and suction cup.

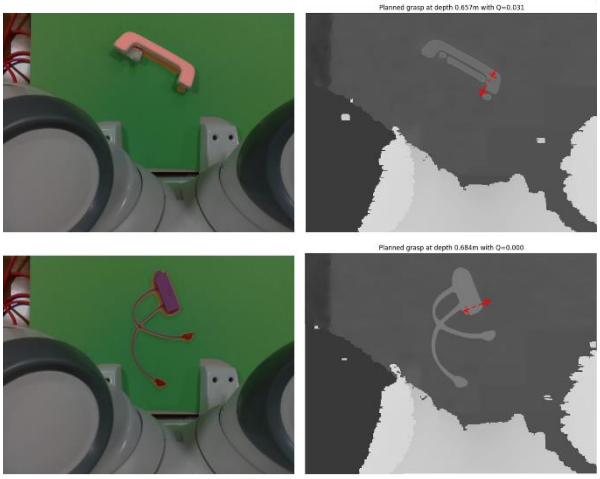

Figure 4. Parallel grip proposals offered by Dex-Net (right) given a mask of the part (in red, left) and a depth image.

Parallel gripping is performed perpendicular to the working surface. Using a cutout from the YOLO bounding box, edge detection is used to find the contours of the relevant workpiece. This way, by taking the center of the cutout and its depth measurement as the terminal grip point $(x, y, z)_{TP}$, and knowing that the part is a parallelepiped, it is possible to use the contours to calculate the orientation θ_{TP} with which the gripper should be closed. This process is schematized in **Figure 5**.

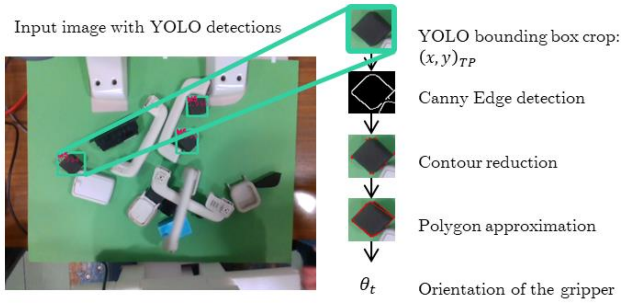

Figure 5. Calculation process of the orientation of the parallel grip.

Suction cup gripping, while following the same operating principle as parallel gripping, offers some additional challenges. As in parallel gripping, the starting point is a cutout according to YOLO's bounding box – however, the cropping is also performed on the depth image in this case.

Depth segmentation is then performed on this crop so as to obtain a mask of the target piece, whose centroid can be calculated, thus obtaining the point of contact $(x, y, z)_{TP}$ of the suction cup on the upper surface of the piece, as shown in **Figure 6**.

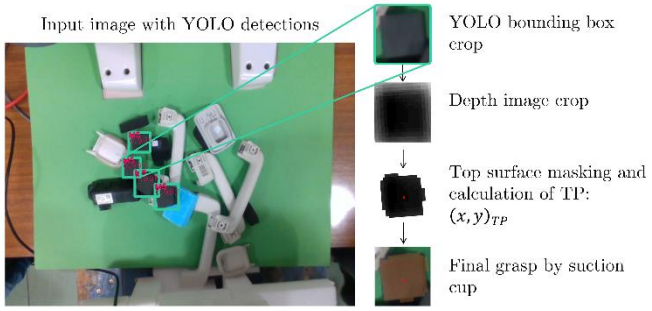

Figure 6. Calculation of the terminal point for the suction cup grip.

Lastly, in order to offer a more flexible solution that allows the use of the suction cup on tilted surfaces, the implementation of the RANSAC algorithm for the detection of oblique planes is carried out. In this way, the calculation of the orientation of the terminal point of the robotic arm is equivalent to that of the normal vector to the plane of the detected part, as shown in **Figure 7**.

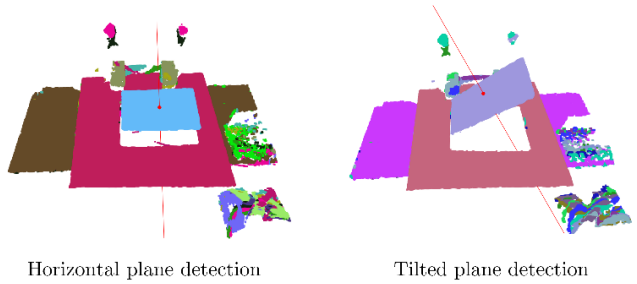

Figure 7. Horizontal and tilted plane detection through the RANSAC algorithm.

As a result, an algorithm capable of proposing both parallel grips and suction cup grips – even on angled pieces – is obtained.

III. RESULTS

Throughout this project, different types of algorithms that allow to detect and extract a target piece given an RGB image and a depth image have been analyzed.

While the depth segmentation method is robust to the environmental changes that may occur in an industrial environment, the quality of the mask provided was not sufficient to allow for the calculation of a quality grasp. Depth subtraction, while providing a clear and quality mask, is a technique that is severely affected by both light fluctuations and changes in the working table, causing it to suffer from a level of robustness that is far too low for industrial environments. Furthermore, both techniques are based on the fixed detection of parts, without being able to actually identify them. In addition, the possibility of implementing Dex-Net's GQ-CNN on the custom dataset is ruled out due to excessively low confidence scores on the calculated grips, as well as the restriction of the grips to an orientation normal to the working plane.

This problem is overcome by using the YOLO algorithm, which is able to detect, locate and identify the part to be grasped – making it much more robust to changes in its environment compared to the two previous masking methods. Hence, by combining the YOLO detections with edge and plane detection techniques (see **Figure 8**), an algorithm capable of gripping any part from the available set of parts – either by means of a parallel gripper or a suction cup – is achieved, even for planes not parallel to the working surface.

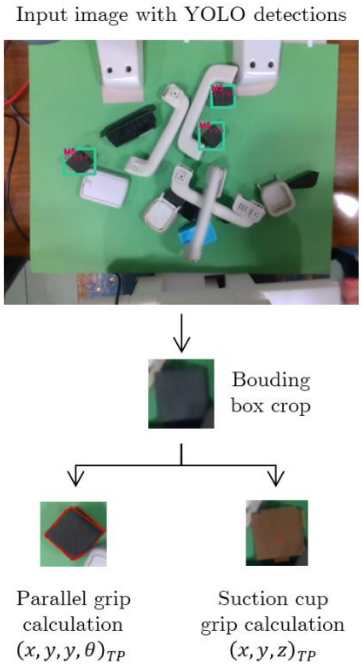

Figure 8. Comparison of the results obtained using the different gripping options (parallel and suction cup)

IV. CONCLUSIONS

As mentioned in the previous section, the results obtained in this project offer the possibility to calculate parallel and suction cup grips in a flexible way depending on the location and type of the part.

Thus, the following conclusions can be drawn:

- Traditional algorithms such as depth segmentation and background subtraction, although adequate in laboratory conditions, are not suitable for use in industrial environments due to their low robustness.

- The creation of a proprietary *ad-hoc* dataset for a specific project is a time-consuming and resource-intensive task, which can present a barrier to entry for the application of artificial intelligence algorithms.

- The use of object detection algorithms such as YOLOv3, once the dataset barrier is overcome, offers a level of flexibility and robustness that leads to very satisfactory results, easily transferable to an industrial environment.

- While plane detection using RANSAC provides good results in laboratory conditions, its application on containers of pieces in an industrial environment, where these are chaotically stacked and interlocked with each other, may not be sufficient.

REFERENCES

[1] Intel, “Intel RealSense D400 Series Product Family Datasheet.” [Online]. Available: https://dev.intelrealsense.com/docs/intel-realsense-d400-series-product-family-datasheet

[2] “Intel RealSense LiDAR Camera L515 Datasheet.” [Online]. Available: https://dev.intelrealsense.com/docs/lidar-camera-l515-datasheet

[3] J. Mahler, M. Matl, V. Satish, M. Danielczuk, B. DeRose, S. McKinley, and K. Goldberg, “Learning ambidextrous robot grasping policies,” Science Robotics, vol. 4, no. 26, 2019, Último acceso: 22/09/2020. [Online]. Available: https://robotics.sciencemag.org/content/4/26/eaau4984/

[4] OpenCV, “OpenCV.” [Online]. Available: https://opencv.org/

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” May 2016, arXiv: 1506.02640.Last access: 13/11/2020. [Online]. Available: http://arxiv.org/abs/1506.02640

[6] A. Romero-Manchado, "Calibración de cámaras no métricas por el método de las líneas rectas", Mapping, ISSN 1131-9100, Nº 51, 1999, pags. 74-80, Jan. 1999.