Robotic Bin-Picking System Based on Voice Recognition, Deep Learning, and Point Cloud Processing

Van-Dung Tran, Thanh-Hung Nguyen, Dinh-Ba Bui, and Minh-Ha Le

Abstract This paper presents an automatic object localization system, which is used to pick the random and on-demand objects in the workspace. The system includes a robotic arm system integrated with a RealSense camera. Firstly, the target object is estimated from the speech recognition algorithm. Secondly, the Yolo-V3 algorithm is applied to detect and classify the target from the color image. Then, individual feature point clusters were extracted using segmented 2-D features and depth maps. To determine the position and orientation of the target, each cluster is matched to the CAD model using the ICP algorithm. Finally, collision avoidance techniques are applied to select objects for the picking task. The feasibility and effectiveness of the developed system have been verified experimentally. The test ended again showing that the system was able to successfully locate and pick up 3-D target objects via voice commands.

Keywords Deep learning · 3-D object detection · 3-D object localization · Voice recognition · Robot bin picking · Yolo V3 detection

1 Introduction

Along with the strong development of the industrial revolution 4.0, the field of robotics develops rapidly and continuously, being applied in all fields and different uses such as: underwater robots, robots on the ground, robots in the sky, robots in space, … Some specific types of robots can reduce human work such as: Mobile robots, Robot Manipulators, Bio Inspired Robots, Personal Robot. In automated factories, smart factories, robotic arms are widely used to assist people in holding, lifting, moving objects, increasing work productivity or supporting people in a toxic and dangerous environments.

V.-D. Tran · T.-H. Nguyen (✉) · D.-B. Bui · M.-H. Le
Hanoi University of Science and Technology, Hanoi 10999, Vietnam
e-mail: hung.nguyenthanh@hust.edu.vn

B. T. Long et al. (eds.), *Proceedings of the 3rd Annual International Conference on Material, Machines and Methods for Sustainable Development (MMMS2022)*, Lecture Notes in Mechanical Engineering, https://doi.org/10.1007/978-3-031-57460-3_38

Robotic systems can be controlled manually by buttons, handles, …, or controlled by software using microcontrollers. A research study in [1] focuses on the design of a mobile robot equipped with a robotic arm utilizing a microcontroller and wireless communication. Another study in [2] outlines the design and control of a two-armed robot with seven degrees of freedom (DOF). In [3], a study explores the design of a robotic arm with 4 DOF, capable of performing individual tasks such as grasping, lifting, placing, and releasing objects. Another study in [4] presented on a 3D object recognition and pose estimation for random garbage selection using partition viewpoint feature histogram. More recently, a study in [5] presented a 3-D objects pose an estimation method for bin-picking using a combination of the semantic point pair feature method and the Mask-RCNN algorithm.

Building upon previous studies, the objective of this study is to design and control a robotic arm with six degrees of freedom, capable of picking up 3D objects through the integration of 3D image processing, voice recognition and deep learning technology.

2 System Overview

Figure 1 shows the object picking robot system. The system includes a robotic arm with gripper, RealSense camera, windows forms control interface. The RealSense camera is used to acquire the RGB-D images. The resulting images are used for object detection through deep learning technology. The RealSense camera also captures 3-D point clouds representing various objects. The evolved 3-D object recognition and localization algorithm is utilized to accurately determine the position and orientation of the target object. Voice control commands are integrated through the laptop's audio acquisition system. The robot uses these parameters to recognize and pick up randomly requested objects.

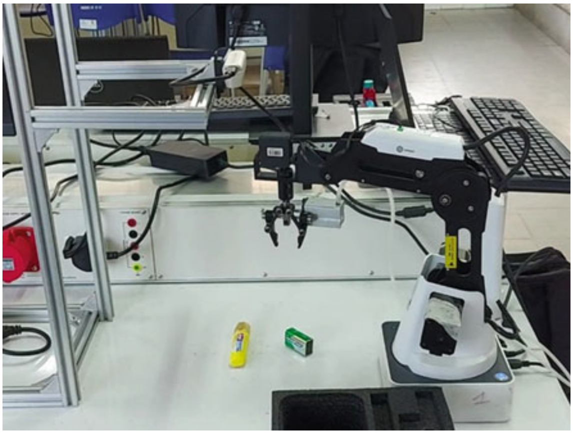


Fig. 1 The developed robotic bin-picking system

3 3-D Object Recognition and Segmentation

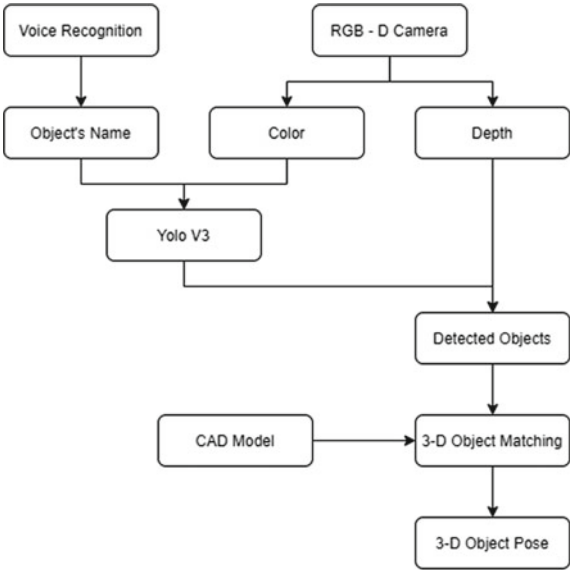
We have random objects with different properties on the table. The goal of this study is to identify individual objects through the prediction of object features and input point cloud segmentation. Each object will be represented by a name label and point cloud. The position of the name label and the point cloud are compared, thereby representing the final object in the orientation bounding box. Finally, the object's name label, location coordinates and orientation are determined for the robot to pick up the object as required. The flowchart of the proposed method is shown in the Fig. 2.

3.1 Voice Recognition

The voice recognition system supports humans in interacting with robots more flexibly. In this study, Microsoft's speech API [6] was used to simplify speech recognition and give commands to the robot to perform.

A simple system is consisting of 2 components: speech synthesis and speech recognition system. Speech synthesis is the process of creating sound or speech through a computer. The received sound will be heard by the computer and recognize the words and phrases, called speech recognition. Voices in predefined cases are recognized to give tasks to the robot to perform.

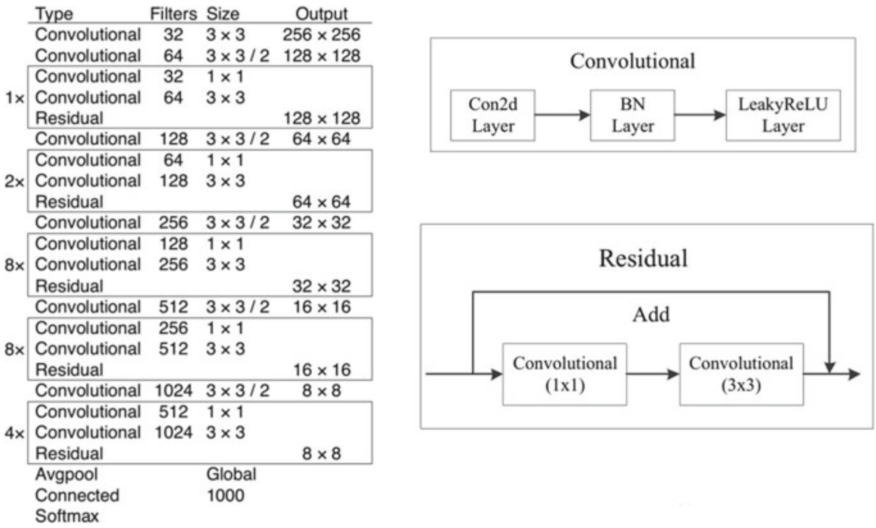
Fig. 2 Flowchart of the proposed method



3.2 2-D Object Detection

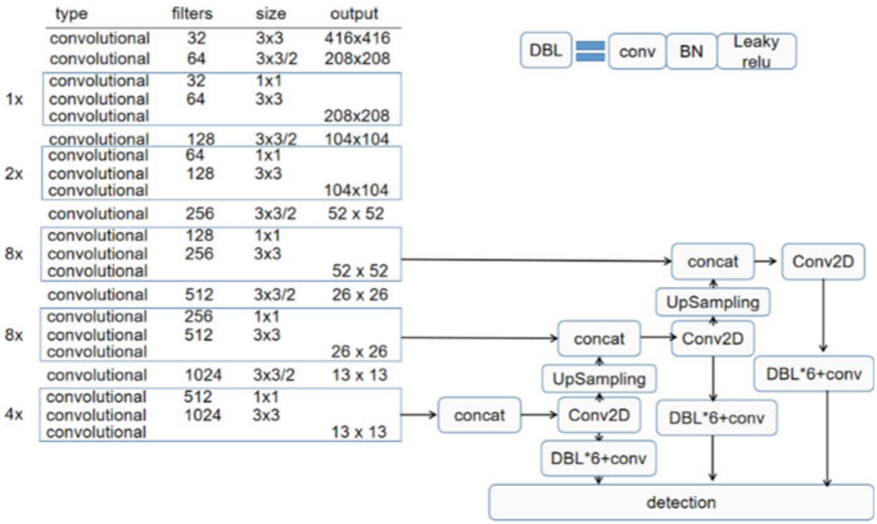
Object detection is one of the fundamental and important tasks of machine learning. Currently, there are many different algorithms that effectively support the detection and classification of interested objects. Such as deep learning algorithms based on convolutional neural networks (Fast R-CNN, Faster R-CNN, Mask R-CNN, etc.). Regression-based algorithms for fast detection of layers and bounding boxes objects such as Yolo, which can be used to recognize objects in real time. Deep learning algorithms based on convolutional neural networks are prioritized for using in many machine learning models. For example, the multi-layered fruit classification model using robot vision and Faster R-CNN network by Wan and Goudos [7], the segmentation model and damage detection in cars using Mask R-CNN by Zhang et al. [8]. The advantages of these algorithms are very good recognition performance and high accuracy. However, in the model we built, the chosen algorithm is Yolo-V3 due to its advantages of fast detection speed, which can be run in real time. The effectiveness of this algorithm is demonstrated through recent publications such as the tomatoes recognition model of Liu et al. [9], the real-time face recognition machine learning model of Chen et al. [10].

The Yolo-V3 model leverages the network architecture of Darknet-53, comprising of 53 convolutional layers and 5 maximum pooling layers. To mitigate overfitting, batch normalization and dropout operations are incorporated after each convolutional layer. The Darknet-53 architecture features five residual blocks, incorporating the concept of residual neural networks. The network diagram of Darknet-53 is depicted in Fig. 3, and the overall structure of the Yolo-V3 network can be seen in Fig. 4.



	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
	Convolutional	32	1 × 1	
1×	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
	Convolutional	64	1 × 1	
2×	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
	Convolutional	128	1 × 1	
8×	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
	Convolutional	256	1 × 1	
8×	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
	Convolutional	512	1 × 1	
4×	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

Fig. 3 Darknet-53 structure



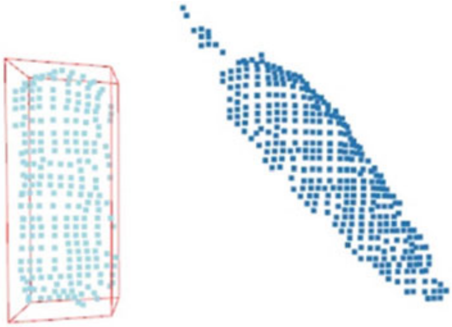
	type	filters	size	output
	convolutional	32	3x3	416x416
	convolutional	64	3x3/2	208x208
1x	convolutional	32	1x1	
	convolutional	64	3x3	
	convolutional			208x208
	convolutional	128	3x3/2	104x104
2x	convolutional	64	1x1	
	convolutional	128	3x3	
	convolutional			104x104
	convolutional	256	3x3/2	52 x 52
8x	convolutional	128	1x1	
	convolutional	256	3x3	
	convolutional			52 x 52
	convolutional	512	3x3/2	26 x 26
8x	convolutional	256	1x1	
	convolutional	512	3x3	
	convolutional			26 x 26
	convolutional	1024	3x3/2	13 x 13
4x	convolutional	512	1x1	
	convolutional	1024	3x3	
	convolutional			13 x 13

Fig. 4 Yolo-V3 network structure

3.3 *Point Cloud Segmentation*

The 3-D point cloud obtained from the camera contains information about the various objects in the scene. Splitting the input point cloud into smaller point clouds containing information that distinguishes each individual object, for later use. Several other studies have been published using Voxel Net, a LiDAR-based on the 3-D object detection network [11] or the 3-stage point cloud segmentation as introduced in [12]. The idea proposed in this study is to combine the detected objects in 2-D color image and the depth map to extract the target objects. After that, the extracted point clouds are filtered to remove noise and unnecessary data. The result of the point cloud segmentation is shown in Fig. 5.

Fig. 5 Point cloud segmentation



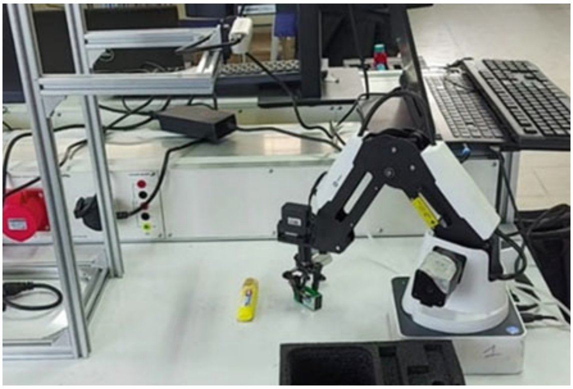


Fig. 6 The target is picked by the robot arm

3.4 Object Localization and Pose Estimation

To estimate 3-D position and orientation of the target object in the scene point cloud, the extracted target point cloud is aligned with the CAD model by employing the Iterative Closest Point (ICP) algorithm [13]. The ICP algorithm is used to obtain the transformation to refine the original estimated 3-D pose. The ICP algorithm is a matching process being employed to minimize the fitting deviation between two matching point clouds. The ICP algorithm iteratively revises the transformation needed to minimize the distance between the points of two raw scans. After the object recognition and localization, the target object will be picked by a parallel jaw clamp as depicted in Fig. 6.

4 Experimental Results

The experiment was tested with many kinds of objects with randomly positions and orientations. Minimum size of the objects is 0.01 × 0.01 × 0.01 m. Calibration results are highly accurate, so robot control can be applied to pick up the right objects as required. The data was processed on a computer with a core I7 processor (2.8 GHz and 8 GB RAM). The average processing time of 20 experiments is about 1.5 s for object localization. If using GPU, the result will be processed much faster (about 10 times faster than using CPU). Through voice control commands, the robot successfully determines the task of finding the location of the 3D objects that coincides with the required mask to carry out the object picking. In some cases, due to the influence of the environment and pronunciation, the robot may not be able to recognize the voice commands.

5 Conclusion

This research has achieved the following purposes:

1. Designing a robot arm model, which can detect, grab, and move objects to the required area.
2. Robot successfully identifies and locates objects through camera and 3-D image processing algorithm. Successfully controlling objects through voice commands and recognizing objects through using deep learning technology.

Acknowledgements This research is funded by Hanoi University of Science and Technology (HUST) [grant number T2021-SAHEP-008].

References

1. I.B. Alit Swamardika, I.N. Budiastra, I.N. Setiawan, N. Indra Er, Design of mobile robot with robotic arm utilising microcontroller and wireless communication. Int. J. Eng. Technol. **9**, 838–846 (2017)
2. J. Tarek, C. Zaoui, M. Aref, Design and control of a dual-arm robot. Int. J. Latest Res. Sci. Technol. **4**, 110–116 (2015)
3. R. Mourya, S. Amit, S. Sourabh, K. Sushant, B. Manoj, Design and implementation of pick and place robotic arm. Int. J. Recent Res. Civ. Mech. Eng. (IJRRCME) **2**, 232–240 (2015)
4. D. Li, N. Liu, Y. Guo, X. Wang, J. Xu, 3D object recognition and pose estimation for random bin-picking using partition view-point feature histograms. Pattern Recogn. Lett. **128**, 148–154 (2019)
5. C. Zhuang, Z. Wang, H. Zhao, H. Ding, Semantic part segmentation method based 3D object pose estimation with RGB-D images for bin-picking. Robot. Comput.-Integr. Manuf. **68**, 102086 (2021)
6. Microsoft's Speech API. https://docs.microsoft.com/en-us/dotnet/api/microsoft.cognitiveservices.speech. Accessed 20 May 2021
7. S. Wan, S. Goudos, Faster R-CNN for multi-class fruit detection using a robotic vision system. Comput. Netw. **168**, 107036 (2020)
8. Q. Zhang, X. Chang, S.B. Bian, Vehicle-damage-detection segmentation algorithm based on improved mask RCNN. IEEE Access **8**, 6997–7004 (2020)
9. G. Liu, J.C. Nouaze, P.L. Touko Mbouembe, J.H. Kim, YOLO-tomato: a robust algorithm for tomato detection based on YOLOv3. Sensors **20**, 2145 (2020)
10. W. Chen, H. Huang, S. Peng et al., YOLO-face: a real-time face detector. Vis. Comput. **37**, 805–813 (2021)
11. S.S. Shi, X.G. Wang, H.S. Li, PointRCNN: 3D object proposal generation and detection from point cloud. arXiv:1812.04244
12. G. Pang, R. Qiu, J. Huang, S. You, U. Neumann, Automatic 3D industrial point cloud modeling and recognition, in *14th IAPR International Conference on Machine Vision Applications (MVA)* (2015), pp. 22–25
13. P.J. Besl, N.D. McKay, A method for registration of 3-D shapes. IEEE Trans. Pattern Anal. Mach. Intell. **14**, 239–256 (1992)