DOI: https://doi.org/10.52651/sam.a.2024.1.39-44

OBJECT RECOGNITION SYSTEM FOR THE SPINBOTICS ROBOTIC ARM

Patrik ŠTEFKA, Peter PÁSZTÓ, Marian KĽÚČIK, Martin SMOĽÁK, Matej VARGOVČÍK, Jakub LENNER

Abstract: This study focuses on the development of a visual system designed to facilitate object detection for the Spinbotics robotic arm in spatial environments. The primary objective is to enable accurate detection and classification of diverse objects, enhancing the arm's capability to grasp and manipulate items effectively. The system employs the YOLOv7 deep neural network, fine-tuned using transfer learning on a local computing infrastructure. Compared to traditional methods like R-CNN and SSD, YOLOv7 offers superior real-time processing capabilities and efficiency, making it well-suited for dynamic environments. Through extensive training and testing, the system demonstrates robust performance in detecting objects across varied scenes and identifying optimal grasp points. This research underscores the effectiveness of integrating advanced computer vision techniques to enhance the operational efficiency and versatility of robotic manipulators in real-world applications.

Keywords: Robot arm; Visual system; YOLO; Object detection.

1 INTRODUCTION

The aim of this task is to design and test a visual system capable of detecting various objects in space that the robotic manipulator can grasp and transfer. The system needs to learn, reliably detect, and classify these objects in different scenes and determine the position of the point where the object can be grasped by the arm. The process of learning and transferring objects is designed with user assistance.

Fig. 1 Spinbotics 6-axis Serial Modular Robot
Source: www.spinbotics.com

The chosen sensor for this task is the Intel RealSense D455 camera. This is a stereoscopic depth camera with a global shutter and compact dimensions (124 mm x 26 mm x 29 mm) with USB-C connectivity. The camera is supported in ROS2 (robot operating system) thanks to the realsense-ros driver. The sensor outputs both RGB and depth images, combining which we obtain a colored 3D point cloud. To train and use the model, a computer with an Nvidia graphics card is required, the GTX1070m card is used. The camera is mounted on the Spinbotics robotic arm (Fig. 1) using a 3D printed holder, positioned between the fifth and sixth axes of the arm near the end effector (Fig. 2). The camera is temporarily connected via a USB-C cable routed externally along the robot, with future plans for connection through the robot's tool connector.

Fig. 2 Placement of the Intel RealSense D455 camera on the Spinbotics Robotic Arm
Source: author.

2 DATA COLLECTION AND PREPARATION

For development and initial testing, six objects were selected: a wooden cube, a wooden prism, and a wooden cylinder. A rod, a coin, and a ring are made of metal. The challenge is to correctly distinguish individual objects despite their very similar shapes from certain viewing angles (cube/prism and coin/ring).

The training dataset was created by capturing the objects on a contrasting surface from multiple viewing angles and distances (Fig. 3), as well as various groups and arrangements of objects. Hundreds of such captured samples underwent manual annotation in the YOLO (You Only Look Once) [1] standard format. Each image was accompanied by a corresponding .txt file with the position of the rectangle and the object identifier on a separate line.

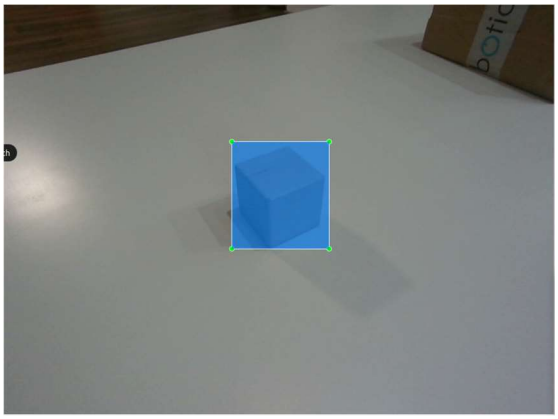

Fig. 3 Cube annotation
Source: author.

To expand the dataset, several augmentation methods were used, where the resulting training image is composed of multiple images from the dataset with slightly altered properties such as HSV (hue, saturation, value) [2], rotation of the original image, translation, scale, etc. We used the following methods: changing HSV parameters within a range of ±30 %, which allows us to simulate different lighting conditions, color distributions and brightness levels. We also used rotation within a range of -15 to 15 degrees, translation within ±30 %, and scaling within ±20 %. Finally, we used vertical mirroring method to the images. We did not use perspective changes or shear deformations.

These parameters contribute to robust detection in other environments.

The collection and extensive annotation will be automated using depth camera data and the position of the robotic arm in the future. For automated dataset collection of new objects, a single object on a contrasting surface at a known position is scanned, around which the robot plans a trajectory to capture the object from as many viewing angles as possible.

Each training batch includes a diverse set of images, generated by applying different augmentations (Fig. 5).

Fig. 4 Multiple augmentation methods were applied to the original image of wooden prism object. First row (left-to-right): original image, altered HSV values; second row: rotation + scale, altered HSV values, vertical mirroring + rotation.
Source: author.

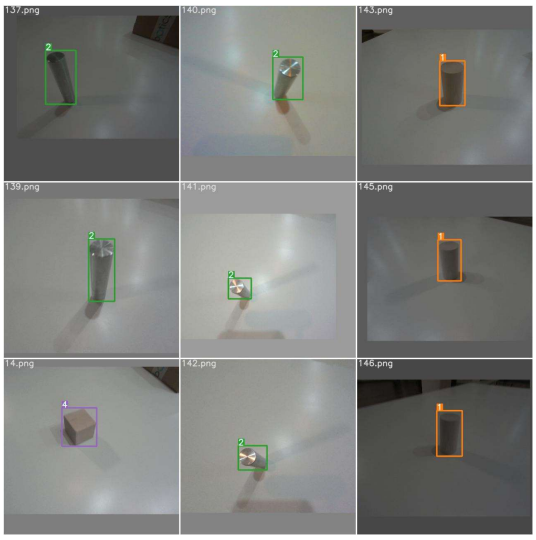

Fig. 5 Sample training batch
Source: author.

3 MODEL TRAINING

The detector is based on the YOLOv7 [1] deep neural network. This single-shot detector was chosen based on previous experience with this model and its good real-time detection capabilities. YOLOv7 is fine-tuned on custom data using transfer learning. In this method, parameters, weights resulting from long training on different objects from the COCO (Common objects in context) dataset [3]. Significantly more powerful hardware was used. In our case, we train locally on a computer with an Nvidia graphics card, and such training takes only

a few hours since only the weights of the last layers of the network are changed.

For training, we always create a dataset split, typically in an 80:20 ratio, where 80 % of the images are used for training and 20 % for testing detection accuracy [4].

The success of the training is indicated by a confusion matrix (Fig. 6). Training continues until detection accuracy stabilizes; in this case, we trained for 700 epochs.

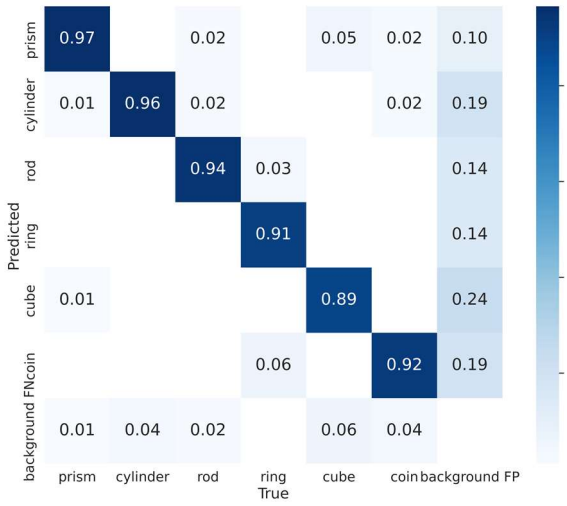

Fig. 6 Confusion Matrix after four hours of training
Source: author.

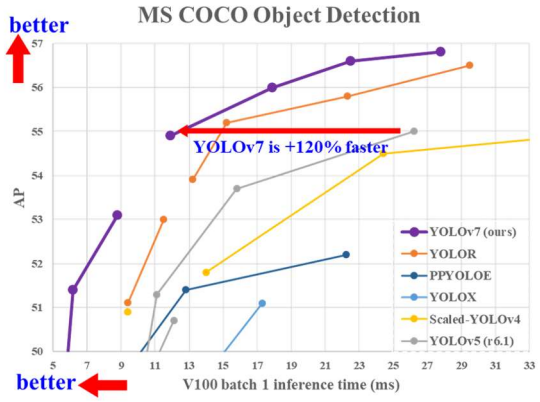

Fig. 7 Comparison of the YOLOv7 model's success and detection time with its previous versions
Source: [1].

4 TESTING THE TRAINED NEURAL NETWORK

The robotic arm should be able to select various stored objects from a view of the scene and move them from their original location to a new place. A database of known objects will be available, from which the user can choose which objects the robot should manipulate.

The trained detector is then integrated into the ROS2 environment. Object detection occurs only from the RGB image, with the detector outputting the coordinates of the rectangle in the image for the detected object.

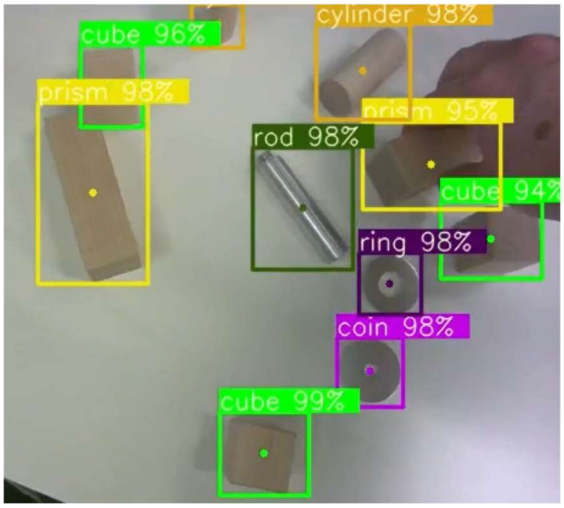

Fig. 8 Detected object in the RGB image
Source: author.

After determining the centroids of the rectangles, their position in the 3D space is calculated by projecting the 2D point from the calibrated depth image. The depth image is pre-processed using algorithms to fill in holes and filter out noise. With our hardware, the detector reliably distinguishes objects at a sampling rate of 15fps, which is sufficient for this type of task. Thus, the robot has information about the position of all recognized objects in the given scene. Measuring the distance of objects with the chosen camera is possible from a distance of 400 mm.

The method for selecting a candidate for grasping and the method of grasping is determined by the user. Either the object closest to the robot's end effector or the object with the most space around it to avoid collision with surrounding objects can be selected.

The images show the detection of objects in the scene (Fig. 9) and the estimation of the centroid distance of all “cube” type objects in the point cloud (Fig. 10), with the cube position marked in green in the rviz2 environment (Fig. 11).

Fig. 9 Output of the trained neural network on a series of test images from the dataset
Source: author.

Fig. 10 Position of points detected on objects in 3D space
Source: author.

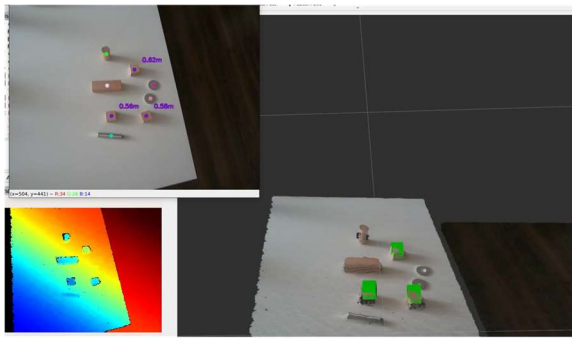

Fig. 11 Candidates for the robot to grasp objects labeled as "cube"
Source: author.

5 DISCUSSION

Our YOLO-based detector offers several key advantages that make it an excellent choice for a visual system. Firstly, it excels in speed and efficiency. Unlike traditional methods that process images multiple times at varying scales, YOLO treats object detection as a single-step process, predicting bounding boxes and class probabilities directly from full images. This streamlined approach reduces computation time, making YOLO well-suited for real-time applications.

Moreover, YOLO's unified architecture processes images in a single pass, enhancing accuracy and performance. It learns generalized object representations effectively, which improves its ability to detect a wide range of objects, even in complex scenes. This contrasts with other detectors that may compromise between speed and accuracy.

Compared to alternatives like R-CNN (Region-based Convolutional Neural Network) [6],[7] and SSD (Single Shot Multibox Detector) [8], YOLO stands out for its real-time processing capability. Traditional methods often rely on multi-stage pipelines, resulting in slower performance. While SSD offers real-time processing, YOLO frequently matches or exceeds its detection accuracy with simpler implementation.

Additionally, YOLO's flexibility in model size allows deployment across various hardware platforms, from powerful servers to edge devices with limited computing resources. This adaptability is crucial in robotics applications, where deployment environments vary widely.

In contrast to expensive industrial vision systems, our YOLO-based solution offers cost advantages. Industrial systems typically require specialized hardware and software, driving up costs. In contrast, our system leverages affordable, off-the-shelf components without sacrificing performance.

Furthermore, our visual system operates locally, eliminating the need for reliance on costly cloud services. Many industrial systems depend on cloud-based processing, which incurs ongoing expenses and requires consistent internet connectivity. Our local processing approach ensures data security and reduces operational costs, making it suitable for environments with unreliable connectivity or strict privacy requirements.

Overall, the YOLO-based visual system strikes a balance between speed, accuracy, and versatility. Its real-time performance and cost-effectiveness are particularly beneficial for applications such as autonomous driving, security surveillance, and industrial automation, where precise object detection is critical.

By adopting the YOLO detector, our visual system not only meets the demands of real-time object detection but also provides a practical, cost-efficient solution compared to many industrial vision systems.

6 FUTURE DEVELOPMENTS

The current solution has not yet been tested with the actual Spinbotics 6-axis Serial Modular Robot. Further development of the task will proceed after verifying the correctness of the proposed system. Gradually, efforts will focus on automating data collection and annotation, optimizing the learning

process, designing the user interface, and specifying the use of the proposed system for a real task with a different set of objects.

7 CONCLUSION

This study presents the design and implementation of an object recognition system for the Spinbotics robotic arm, leveraging the Intel RealSense D455 camera and YOLOv7 neural network. Challenges such as distinguishing similar objects and ensuring robust detection in varied environments were addressed. The system reliably detects and classifies objects, crucial for pick-and-place tasks.

In comparison to traditional methods like R-CNN and SSD, the YOLOv7-based system stands out for its speed and efficiency, making it suitable for real-time applications. Unlike many industrial vision systems that rely on expensive, specialized hardware and cloud-based processing, our solution operates entirely on affordable, off-the-shelf hardware and runs locally. This not only reduces costs but also ensures data security and independence from prepaid cloud services.

Future efforts will focus on real-world testing, automating data handling, optimizing training processes, enhancing user interfaces, and adapting the system to different object sets. These steps aim to refine the system's performance and usability further. By continuing to improve the system's capabilities and versatility, we aim to create an effective and accessible solution for a wide range of industrial applications.

Additionally, the local operation of the system eliminates dependency on internet connectivity and cloud services, which can be a limitation in environments with unreliable connectivity or stringent data privacy requirements. This ensures that our system remains robust and functional in diverse settings.

In summary, the developed object recognition system for the Spinbotics robotic arm not only meets the demands of real-time object detection but also provides a cost-effective, secure, and versatile solution. This positions it as a competitive alternative to more expensive industrial vision systems, paving the way for broader adoption in various industrial applications.

References

[1] WANG, C. Y., BOCHKOVSKIY, A. and LIAO, H. Y. M. (2023). YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. [Online]. Available at: https://doi.org/10.1109/cvpr52729.2023.00721

[2] BRADSKI, G. (2000). "The OpenCV Library." Dr. Dobb's Journal of Software Tools.

[3] LIN, T., MAIRE, M., BELONGIE, S., HAYS, J., PERONA, P., RAMANAN, D., DOLLÁR, P. and ZITNICK, C. L. *Microsoft COCO: Common objects in context.* CoRR abs/1405.0312 (2014). Available at: https://doi.org/10.1007/978-3-319-10602-1_48

[4] McCLUER, N. *TensorFlow Machine Learning Cookbook. Packt*, 2017. ISBN 9781786462169.

[5] HUANG, J. et al. Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*". Honolulu, HI, 2017, pp. 3296-3297. Available at: https://doi.org/10.1109/CVPR.2017.351

[6] GIRSHICK, R., DONAHUE, J., DARRELL, T. and MALIK, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2014. Available at: https://doi.org/10.1109/CVPR.2014.81

[7] GIRSHICK, R. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV),* 2015. Available at: https://doi.org/10.1109/ICCV.2015.169

[8] LIU, W., ANGUELOV, D., ERHAN, D., SZEGEDY, C., REED, S., FU, C. Y. and BERG, A. C. SSD: Single Shot Multibox Detector. In *Proceedings of the European Conference on Computer Vision* (ECCV), 2016. Available at: https://doi.org/10.1007/978-3-319-46448-0_2

[9] SZELISKI, R. *Computer Vision: Algorithms and Applications.* London: Springer-Verlag, 2022. Available at: https://doi.org/10.1007/978-3-030-34372-9

Dipl. Eng. Patrik **ŠTEFKA**
RoboTech Vision s.r.o.
Červený kameň 61
900 89 Častá
Slovak Republic
E-mail: stefka@robotechvision.com

Dipl. Eng. Peter **PÁSZTÓ**, PhD.
RoboTech Vision s.r.o.
Červený kameň 61
900 89 Častá
Slovak Republic
E-mail: paszto@robotechvision.com

Dipl. Eng. Marian **KĽÚČIK**
RoboTech Vision s.r.o.
Červený kameň 61

900 89 Častá
Slovak Republic
E-mail: klucik@robotechvision.com

Dipl. Eng. Martin **SMOĽÁK**
RoboTech Vision s.r.o.
Červený kameň 61
900 89 Častá
Slovak Republic
E-mail: smolak@robotechvision.com

Dipl. Eng. Matej **VARGOVČÍK**
RoboTech Vision s.r.o.
Červený kameň 61
900 89 Častá
Slovak Republic
E-mail: vargovcik@robotechvision.com

Dipl. Eng. Jakub **LENNER**
RoboTech Vision s.r.o.
Červený kameň 61
900 89 Častá
Slovak Republic
E-mail: lenner@robotechvision.com

Patrik ŠTEFKA Studied robotics on Slovak University of Technology in Bratislava. Presently he is working in RoboTech Vision Ltd. company where he is responsible for development and integration of AI methods and neural network data processing into human interaction, navigation and localization algorithms.

Peter PÁSZTÓ studied robotics on Faculty of Electrical Engineering and Information Technology of STU in Bratislava. He is one of the founders and CEOs of RoboTech Vision Ltd. Throughout his university studies he has focused on image processing algorithms and their application in the field of mobile robotics (navigation and localization). During PhD studies he was awarded together with his team (now already CEOs of company) at a scientific conference in Rijeka, Croatia for an image processing algorithm detecting obstacles in front of a mobile robot navigated only by a smartphone running on Android OS.

Marian KĽÚČIK studied robotics on Faculty of Electrical Engineering and Information Technology of STU in Bratislava. He is one of the founders and CEOs of RoboTech Vision Ltd. During his studies he focused on the development of a robot with a combined chassis. He was also working on development of genetic algorithms for navigation and localization of his robotic platform. His skills include programming in multiple programming languages, controlling various operating systems, control system development, mechanics, control software and electronics. He focuses on the development of communication layer software and the configuration of OS in his company.

Martin SMOĽÁK has been involved in mobile robotics since secondary school. He was also involved in telemedicine projects and developed smart cars for the police in the past. He studied robotics on Faculty of Electrical Engineering and Information Technology of STU in Bratislava. Within RoboTech Vision Ltd. (of which he is also on of founders and CEOs), he focuses mainly on the development of lower-level control software, hardware, mechanics, and navigation algorithms and mobile robotic platforms design.

Matej VARGOVČÍK studied robotics on Faculty of Electrical Engineering and Information Technology of STU in Bratislava. During his studies he was developing number of significant scientific works. He implemented his knowledge in RoboTech Vision Ltd. company in which he is currently working. He is responsible for development of autonomous navigation an localization algorithms. He also actively collaborates with the scientific comunity in his field and is contributing his solutions on GitHub platform.

Jakub LENNER studied robotics on Slovak University of Technology in Bratislava. Presently he is working in RoboTech Vision Ltd. company where he is cooperating on development of visual navigation and path-planning methods. He is also author of several scientific publications in the field of precise robot path-planning for docking algorithms using visual systems and image processing.