Application of three-dimensional vision perception technology to industrial robots

Rui Yang1,2, Qinglong Mo1, Yuhong Li1*, Lin Gan1 and Ruihan Hu1,3,4,5*

1 Guangdong Institute of Intelligent Manufacturing, Guangdong Key Laboratory of Modern Control Technology, Guangzhou 510070, China
2Cloud Computing Center Chinese Academy of Sciences, Dongguan 523808, china
3Guangdong Key Laboratory of Modern Control Technology, Guangdong Institute of Intelligent Manufacturing, Guangzhou, China
4Engineering Research Center of Intelligence Perception and Autonomous Control, Ministry of Education, Beijing, China
5Xinjiang Production and Construction Corps Key Laboratory of Modern Agricultural Machinery, Shihezi University, Shihezi, China
*Corresponding author’s e-mail: yh.li@giim.ac.cn, rh.hu@giim.ac.cn

Abstract. The three-dimensional vision system can improve the active perception ability of the robot, and then guide its flexible operation. This system has been widely used in industrial production processes, such as disorderly sorting, assembly, flexible welding, and defect detection. In sorting, assembly and other applications, accurate perception in a complex and changeable industrial environment is essential. Moreover, the control and other operations should be completed under the guidance of feedback information based on the collected three-dimensional perception results. Nonetheless, improvements are still required, such as accurate three-dimensional detection and positioning of work-in-progress and autonomous guidance in a complicated industrial context with continuous changes.

1. Introduction

The three-dimensional visual perception of industrial robots mainly covers the detection, recognition and positioning of targets (operating objects). In industrial applications, most of the existing work-in-progress detection algorithms take the edge of the work-in-progress as the matching feature to specify the similarity measurement criteria of the contour of the target object. Then, the search algorithm is used to realize the matching recognition of the target object in the distance transformation space. Generally speaking, the commonly used target detection, recognition and location technologies can be roughly divided into two categories: the traditional feature methods based on "voting", "matching" or other mechanisms, and the feature extraction methods based on the deep neural network[1-2]. Examples of the "voting" mechanism-based methods include Hough transform and its variants, RANSAC, etc. Hough transform is subject to such limitations as the poor anti noise ability and high computational complexity, since noise will interfere with the detection of real vertices in Hough parameter space. Moreover, its computational and storage complexity increases exponentially with the growing shape free parameters[3-5]. Although RANSAC has good anti-noise ability when processing a single shape (the image contains only a single target) (the amount of calculation may be large), its effect in multi-
target detection is even worse than Hough transform. This situation is particularly obvious in an industrial environment involving the accumulation or occlusion of work-in-progress. For methods under the "matching" mechanism, the early template-based matching directly uses the image pixel information to obtain the similarity measurement results. This type of technology is simple and mature. However, its matching accuracy and operation speed are constrained by changes in the illumination and viewing angle, and by the significantly improved image resolution in recent years, respectively. These two aspects greatly hinder its industrial applications. It is worth noting that the methods based on "voting" and "matching" mechanisms only achieve target detection and recognition, while an additional positioning process is still needed. This leads to the isolation of algorithm steps and the difficulty to optimize the global optimal algorithm. The target detection algorithm based on the deep neural network, with the rapid improvement of computer computing ability, has been increasingly applied to computer vision. Through the end-to-end connection, the deep neural network facilitates the construction of a system that integrates detection, recognition and positioning and can be optimized by back-propagation algorithm. Firstly, a deep convolution neural network is needed for feature extraction. The possible heavy workload of calculation makes it difficult to meet the real-time requirements\[^{6-7}\]. Then, because the commonly used object detection algorithms only use rectangular boxes to mark the target position, the anchor setting is monotonous, thus failing to accurately estimate the target position and attitude in complex and changeable industrial environments. Given the above problems and the challenges brought by the disorderly occlusion of work-in-progress and defect detection tasks in the industrial environment, this project proposes an efficient deep learning framework for target detection, recognition and positioning\[^{8-10}\]. It can use the powerful expression ability of the deep neural network to adaptively learn the characteristics of work-in-progress and defects in the industrial environment, thus improving the forward reasoning process to meet the real-time performance. Moreover, the end-to-end deep neural network is adopted to optimize the core steps such as network topology construction, anchor design and feature enhancement. These advantages promote the detection and recognition of occluded targets and flexible targets, and increase the efficiency of the algorithm.

2. Three-dimensional visual perception

The image based on a two-dimensional plane lacks the vision information related to three-dimensional depth. In the industrial environment, the spatial structure of the work-in-progress is a key feature for the task of detection, recognition and positioning. Moreover, it faces the challenge to accurately distinguish pits, defects, scratches and other defects. These two issues can be well solved by three-dimensional vision. Compared with traditional RGB image data, RGB-D data covers the depth information of the scene, thus improving the performance of features. In recent years, spatial information has been introduced for target detection, recognition and positioning based on RGB-D data. Specifically, conventional methods such as sparse representations, hog features, and support vector machines are mainly adopted, which take both RGB information and depth information as inputs for the shallow fusion of data sets or prediction levels. The fusion between information is not sufficient, which requires further improvement. Nowadays, with the advancement of industrial manufacturing technology and processing technology, higher requirements for detection content, detection efficiency and accuracy are proposed. Therefore, to achieve efficient 3D visual information fusion in industrial environments and learn more sufficient features, the thorough combination of information between multiple modes in the algorithm has become an urgent task. Herein, the depth information generated by color images and structured light is taken as the input of the algorithm framework. In the processing of 3D visual input, the challenges are mainly embodied in the following two aspects(Fig.1).
2.1. Feature representation and transformation
Since the feature representation of work-in-progress varies with the modes in 3D vision, the complementarity and redundancy between modes should be fully exploited to obtain a unified feature representation. However, most of the existing fusion methods adopt simple addition, multiplication and maximization, failing to make full use of cross-modal information. In addition, in practical industrial applications, due to the limitations of environments and equipment, modal losses of input information may occur. To complete the simulation of missing modes, an effective mapping method should be generated from the complete training data of modes (Fig. 2).

2.2. Deep learning and joint training
The industrial environment usually involves a variety of work-in-progress and defects, as well as the appearance in the class, which increases the difficulty to define them with a regular contour or
structure. For this reason, a more general detection framework based on deep learning is proposed. Using the robust feature extraction ability of the neural network, the back-propagation algorithm adaptively learns the convolution filter to deal with the detection targets with irregular appearances. For high detection accuracy, the algorithm framework quickly obtains the semantic feature representation of the image to be detected through the core steps of network topology construction, anchor design and feature enhancement. It hence accurately and efficiently judges the target in the image to promote further recognition and positioning. To meet the real-time requirements, model compression, pruning and other algorithms are put forward to improve the detection efficiency (Fig. 3).

![Figure 3. Detection and identification of cable cover reinforcement](image)

3. **Cable cover reinforcement welding target positioning**

Occluded targets, incomplete features and other target interference in the industrial environment tend to cause positioning errors. An occlusion-sensitive positioning algorithm is introduced to model the occlusion relationship between objects and improve the accuracy of initial target positioning. Notably, the traditional square anchor only contains the position and scale information, but not the important information such as the target altitude and inclination in the industrial environment. Thus, a rotatable anchor algorithm is used to display and model the inclination angle of the target, thereby increasing the accuracy of initial target positioning. With the three-dimensional CAD model of cable cover reinforcement welding as the input, a hierarchical two-dimensional view model is automatically generated. Based on the initial positioning results, the pose is optimized to further obtain the accurate three-dimensional pose of the target (Fig. 4).

![Figure 4. Cable cover reinforcement positioning](image)
4. Conclusion
This paper presents a method to efficiently fuse the multi-modal information from three-dimensional sources, make full use of the complementarity between modes, and eliminate ambiguity and redundancy. For different input data, adaptive adjustment of the fusion strategy can be realized by reasonable modeling and use of the interactive information between modes. Furthermore, the characteristics of industrial conditions are organically combined with the deep learning algorithm to achieve accurate and fast target detection, recognition and positioning. Specifically, when various targets (occlusion targets, flexible targets, defects, etc.) in the industrial environment are considered, the network structure, loss function constraints and training paradigm are designed to improve the feature expression ability of targets. This will further improve the intra class differences of targets and reduce the inter ones, finally elevating the efficiency and generalization performance of the algorithm.

Acknowledgments
This project is supported by the National Science Foundation of China (Grant No. 51975219), as well as in part by the Guangdong Academy of Sciences’ (GDAS’) Project of Science and Technology Development (Grant 2017GDASCX-0115 and Grant 2018GDASCX-0115), the Guangdong Academy of Science for the Special Fund of Introducing Doctoral Talent (Grant 2021GDASYL-20210103087), the Opening Foundation of Xinjiang Production and Construction Corps Key Laboratory of Modern Agricultural Machinery (Grant BTNJ2021003), and the Foundation of Engineering Research Center of Intelligence Perception and Autonomous Control, Ministry of Education of China (Grant K100052021008).

References
[1] Ferraris Claudia, et al."Monitoring of Gait Parameters in Post-Stroke Individuals: A Feasibility Study Using RGB-D Sensors." Sensors 21.17(2021): doi:10.3390/S21175945..
[2] Zhao Yifan, et al."RGB-D Salient Object Detection with Ubiquitous Target Awareness." IEEE transactions on image processing : a publication of the IEEE Signal Processing Society PP.(2021): doi:10.1109/TIP.2021.3108412.
[3] Cao Jinming, et al."RGB[formula omitted]D: Learning depth-weighted RGB patches for RGB-D indoor semantic segmentation." Neurocomputing 462.(2021): doi:10.1016/J.NEUCOM.2021.08.009.
[4] Skoczeń Magda, et al."Obstacle Detection System for Agricultural Mobile Robot Application Using RGB-D Cameras." Sensors (Basel, Switzerland) 21.16(2021): doi:10.3390/S21165292.
[5] Skoczeń Magda, et al."Obstacle Detection System for Agricultural Mobile Robot Application Using RGB-D Cameras." Sensors 21.16(2021): doi:10.3390/S21165292.
[6] Tokas Theodoros, et al."Impact of three-dimensional vision in laparoscopic partial nephrectomy for renal tumors." Turkish journal of urology. (2020): doi:10.5152/TUD.2020.20439.
[7] Mongelli Francesco, et al."Three-dimensional laparoscopic Roux-en-Y gastric bypass with totally hand-sewn anastomoses for morbid obesity. A single center experience." Acta cirurgica brasileira 35.8(2020): doi:10.1590/s0102-8650202008000006.
[8] Kim SungHyun,and Lee* Won Hyung."Rotation Correction Method Using Depth-Value Symmetry of Human Skeletal Joints for Single RGB-D Camera System." Tehnički vjesnik 28.4(2021): doi:10.17559/TV-20210116090844.
[9] Teng Xiaowen, et al."Three-Dimensional Reconstruction Method of Rapeseed Plants in the Whole Growth Period Using RGB-D Camera." Sensors (Basel, Switzerland) 21.14(2021): doi:10.3390/S21144628.
[10] Preksha Pareek,and Ankit Thakkar."RGB-D based human action recognition using evolutionary self-adaptive extreme learning machine with knowledge-based control parameters." Journal of Ambient Intelligence and Humanized Computing .prepublish(2021): doi:10.1007/S12652-021-03348-W.