Research on pose and position estimation based on mask RCNN edge extraction for AR auxiliary assembly system

Teng Zhu[*ad], Xiaoyin Guo[c], Hansheng Yan[cd], Yongbin Chen[b], Bin Shu[b]
[a]School of Surveying and Remote Sensing Information, Guangdong Institute of Industry and Commerce, Guangzhou 510550, Guangdong, China; [b]School of Mechanical and Electrical Engineering, Guangdong University of Technology, Guangzhou, Guangdong, 510039, China; [c]School of Mechanical and Electrical Engineering, Guangdong Institute of Industry and Commerce, Guangzhou 510510, Guangdong, China; [d]National Virtual Simulation Training Base of Surveying and Mapping Geographic Information Technology, Guangdong Polytechnic of Industry and Commerce, Guangzhou 510510, Guangdong, China

ABSTRACT

To solve the registration problem of complex texture objects in monocular images in an AR auxiliary assembly system, a 3D object registration method based on Mask RCNN edge extraction is proposed in this paper. This method mainly adopts the edge contour feature of long plate workpiece and enhances the robustness of the algorithm by constructing the matching relationship between 3D model and 2D feature. The experimental result shows that compared with the contour template matching of the Holcon, the Mask CNN-based pose estimation method proposed in this paper can effectively improve the effectiveness of the auxiliary system.

Keywords: Augmented reality, pose estimation, Mask RCNN, assembly guidance

1. INTRODUCTION

Augmented reality (AR) is a technology that uses computers and related display devices to superimpose virtual information on users' vision to enhance real scene information[1]. Instead of text information superposition, the main point of those current AR researches is how to "place" a 3D virtual model in the real world accurately and stably[2]. It is necessary to calculate the pose and position of the target in the virtual scene so that the 2D projection can be displayed without deviation, Then the 3D virtual object can be seamlessly added to the real environment[3].

Currently, target pose detection is one of the research hotspots in the field of computer vision[4,5]. The methods are mainly divided into two categories: target reconstruction and matching based on binocular vision combined with depth information[6], feature matching based on monocular cameras and image edges[7], and feature points or texture[8]. Among them, binocular vision methods can effectively recover depth information, but they require a large amount of computation and are difficult to meet the needs of real-time tracking[9]. However, the real-time and economic requirements of the actual production environment make the monocular vision become a more concerning research direction.

Aiming at the problem that the classical experimental algorithms do not match the application requirements, a novel 3D position and pose monitoring method was proposed in this paper based on Mask Region Convolutional Neural Network (Mask RCNN) and line feature. Experiment results show that the algorithm can overcome the problems of unstable illumination, easily occluded natural characteristics and metal surface materials in the actual production environment.

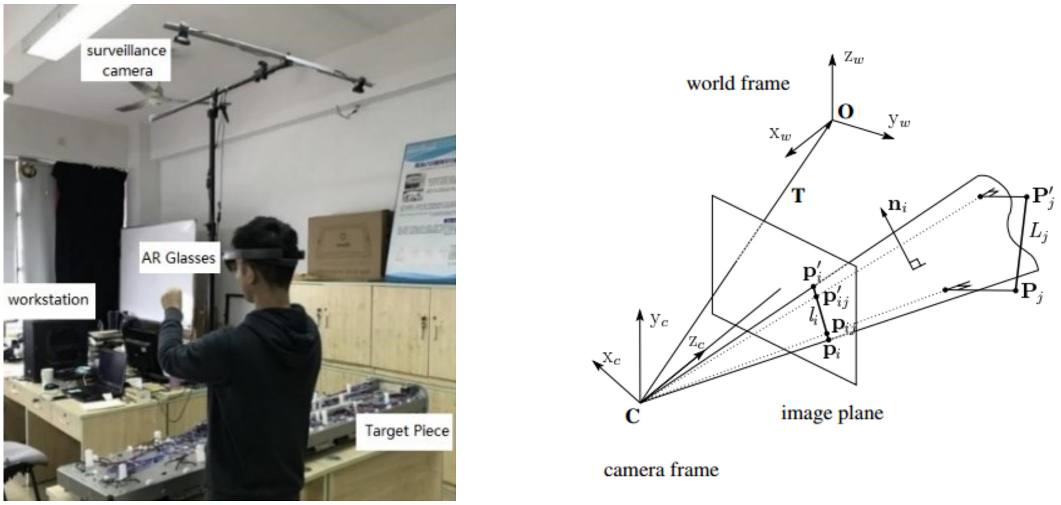
2. PRODUCTION ENVIRONMENT AND PRINCIPLE OF MONOCULAR POSE TRACKING

The research object of this paper is an AR auxiliary system in the process of antenna board wiring of the large base station. Its main structure includes AR glasses, graphics processing terminal, scene surveillance camera, workbench and target workpiece. For cost control and practical work needs, no manual marking is allowed in monitoring scenes such as antenna

[*] jhgf_1234@sina.com

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 125066F
© 2022 SPIE · 0277-786X · doi: 10.1117/12.2661820

board and workbench. At the same time, a common optical camera is used for scene monitoring, as shown in Figure 1a.



(a) work scenario

(b) monocular vision pose estimation model

Figure 1. Research environment and theoretical basis.

In this AR auxiliary system, the overhead camera is the main information acquisition device. Its images were used to position and pose transformation information of the target antenna panel and control the virtual model to complete the corresponding movement in the AR scene. Although monocular vision can hardly obtain depth information, it has the combined advantages of fast computation, low cost and moderate accuracy. Considering the long and narrow shape of the target plate, the actual 6-D solution can be reduced to the constrained 4-D case, i.e., both the depth displacement and the long-edge turning angle can be considered within a very small variation.

3. LINE EDGE EXTRACTION BASED ON MASK RCNN

The basic structure of Mask RCNN and fast RCNN adopts the same two-state steps. First, the method finds the Region Proposal Network (RPN), then classifies and locates each ROI found by the RPN, and finally finds the binary mask, which is different from other network frameworks that find the mask first. The process of Mask RCNN for the AR auxiliary system is shown in Figure 2.

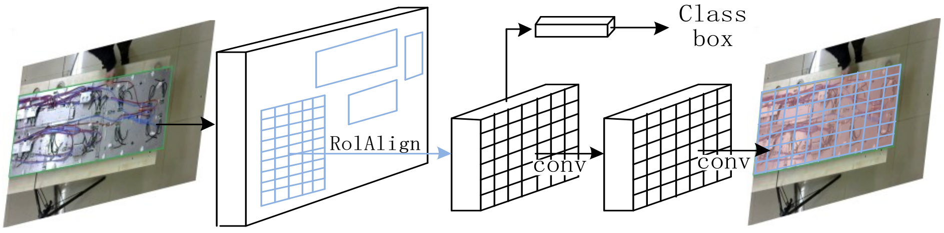
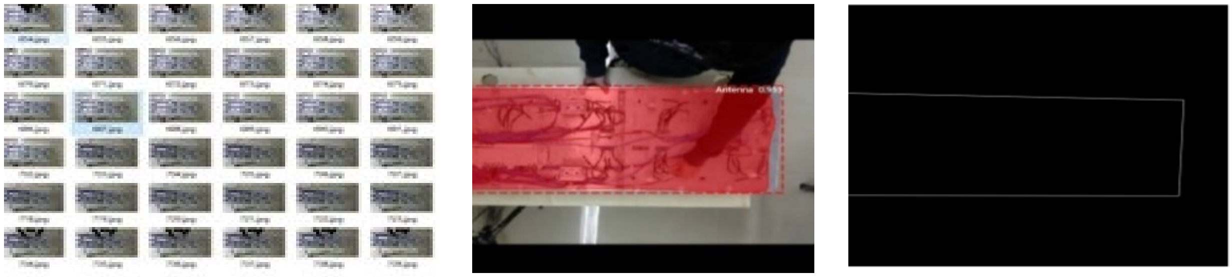


Figure 2. Frame of the Mask RCNN instance split.

Compared with fast RCNN, Mask RCNN adds a mask prediction branch[10], improves ROI pooling and proposes ROI alignment, which can realize one-to-one correspondence between output and input pixels. The main process of the training sector is as follows: first, an image was input to the model, which will have 2000 regions to be detected and proposed on. Then the features of every region are extracted one by one through a convolution neural network. Finally, these extracted features are classified by a support vector machine (SVM) to obtain the category of the object, and the size of the target bounding box is adjusted through bounding box regression.

The surface parts and wiring of the antenna board will change with the progress of the process. Therefore, this paper mainly uses the edge line of the antenna board as the matching feature, and the edge segment extraction method is completed by using the excellent target recognition and segmentation ability of the Mask RCNN neural network[11]. The process of edge extraction includes training set selection, feature extraction, SVM classification, parameter training, neural network segmentation and edge fitting. The intermediate results are shown in Figure 3.



(a) training set (b) image segmentation result (c) edge lines

Figure 3. Mask-RCNN line edge extraction.

After the Line edges were found. Rotation variation of the target sheet along the *Z* axis in the world coordinate system *θ*. It can be calculated by two long edges that are basically parallel in the target contour line extracted from the template image and the key frame image. The calculation formula is:

$$\theta = \tan^{-1}(|k_1 - k_2| / [1 + k_1 k_2]) \tag{1}$$

where k_1 and k_2 respectively refer to the average slope of the two parallel long edges of the template graph and the key frame, The direction of *θ* is judged by the size of k_1 and k_2. According to the right-hand rule, if $k_2 > k_1$, it is clockwise, otherwise, it is counterclockwise. The flip angle *ϒ* of the target sheet along the X axis in the world coordinate system can be obtained from the curve fitting the included angle of the contour edge line and the turning angle *φ*, which can be obtained by the following formulas.

$$\begin{aligned} v_1 &= \overrightarrow{pq}, v_2 = \overrightarrow{ps} \\ v_1 \cdot v_2 &= \|v_1\| \|v_2\| \cdot \cos \varphi \\ \varphi &= \cos^{-1}(v_1 \cdot v_2 / \|v_1\| \|v_2\|) \end{aligned} \tag{2}$$

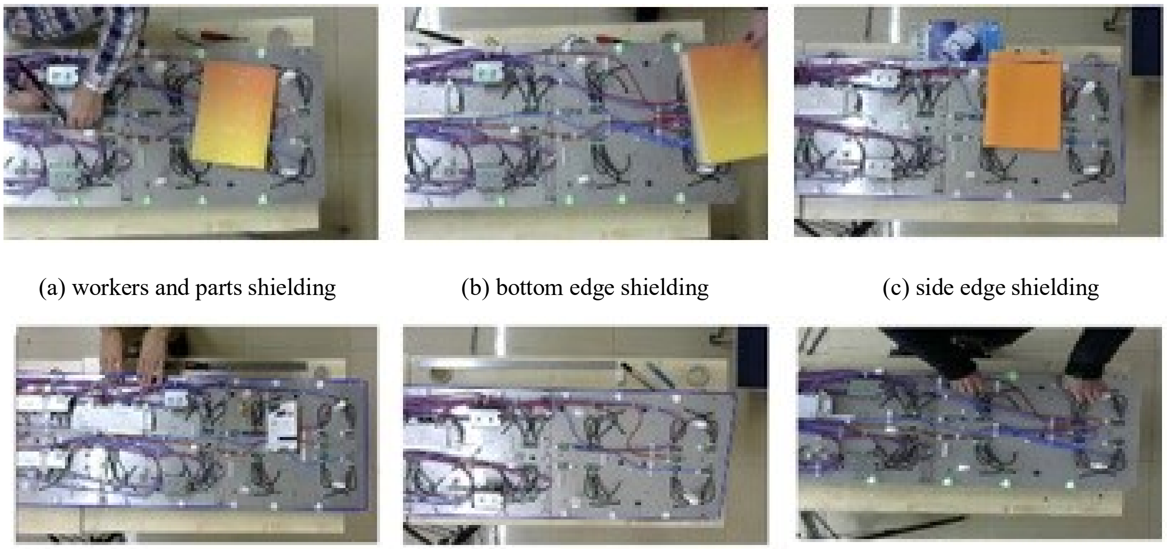
where v_1 and v_2 is the direction vector of line edge k_1 and k_2. And the flip angle *ϒ* And **Z** axis displacement *z* is:

$$\Upsilon = h_\theta(\varphi), z = l - l\cos(|\alpha|) \tag{3}$$

4. EXPERIMENT AND DISCUSS

The robustness of the ANN-based algorithm in the actual production environment was tested in the experimental part. And the fitting accuracy and solution speed were compared with the local template matching method of Holcon 18.11. In the intelligent production case studied in this paper, 7 groups of sampling data between [-15°, 15°] are selected Homogeneous as polynomial fitting reference. The test part of the pose solution method mainly tests the robustness under three interference factors: occlusion, illumination and overturning. The results are shown in Figure 4.

As can be seen from Figure 4, the Mask CNN-based pose estimation algorithm adopted in this paper has good adaptability to different illumination and occlusion interference. At the same time, to verify the accuracy of this algorithm in the pose tracking process, the fitting results are compared with the real values according to seven groups of sampled data between [-15°, 15°]. The comparison results show that the error in the Y direction is not more than 5 pixels, the error in the X direction is not more than 2 pixels, and the fitting accuracy rotation of the angle *φ* can reach 1°, the turning angle can reach 2°, and the error with the real value is less than that of the local contour template matching method of Holcon. For the speed of the pose solution, the average time of the Mask CNN-based method is 643ms while the Holcon matching is 864ms, which improves the efficiency by more than 25%.

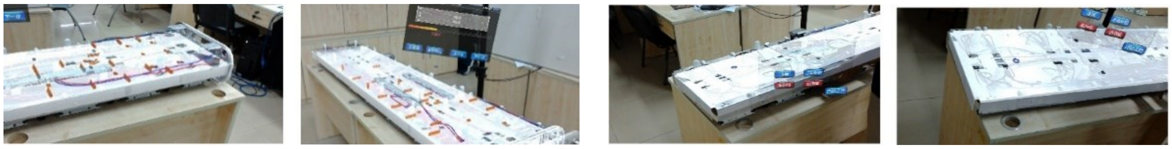


(a) workers and parts shielding (b) bottom edge shielding (c) side edge shielding

(d) strong reflection (e) side rotation with high light (f) edge shielding with weak light

Figure 4. Shape matching results in a variety of complex working environments.

Finally, Figure 5 shows the augmented reality fusion scenario from the perspective of the AR cabling auxiliary system. It can be seen from the figure that since the target plate is a large target, the visual impact of the error of the polynomial fitting algorithm on the human eye is difficult to detect, which does not affect the operator's wiring operation.



(a) $\Upsilon = 0$ (b) $\Upsilon = 10$ (c) $\Upsilon = 15$ (d) $\Upsilon = -15$

Figure 5. View scene in AR glass.

5. CONCLUSION

The edge line feature extracted by Mask RCNN was used to solve the three-dimensional pose of the target workpiece, combined with the analysis of the pose change in the actual production environment, improving the pose solution speed by reducing the dimension and reducing the search space, completes the fast and robust extraction of the edge of the antenna board by using the anti-interference of Mask RCNN neural network. The experimental results show that the accuracy of the pose solution method based on Mask RCNN can reach 2-5 pixels, and the rotation angle fitting accuracy is better than 1°, both of which exceed the local template matching algorithm of the software Holcon. And the key frame relocation efficiency is improved by more than 25%.

ACKNOWLEDGEMENTS

The basic and applied basic research project of Guangzhou basic research plan, "Research on three-dimensional tracking and registration technology of multi person collaborative virtual simulation experiment platform" (20210201298), the special project for high-level talents of Guangdong industry and Trade Vocational and technical College "Research on the key technology of rapid generation method of VR scene and model" (2021-gc-06), and special project of National Virtual Simulation Training Base "Research on Key technologies of Remote Sensing virtual simulation training system", the special project of Guangdong science and technology innovation strategy "Research on VR/AR key technologies in online commodity display in the post epidemic era" (pdjh2021a0741).

REFERENCES

[1] Chen, Y. C., He, H. W., Chen, H. E., et al., “Improving registration of augmented reality by incorporating DCNNS into Visual SLAM,” International Journal of Pattern Recognition & Artificial Intelligence 32(12), 1855022 (2018).

[2] Wu, Y. M., He, H. W., Sun, J., et al., “Vision-based hand tracking and gesture recognition for augmented assembly system,” Key Engineering Materials 392(13), 1030-1036 (2009).

[3] David, P., DeMenthon, D., Duraiswami, R., et al., “Simultaneous pose and correspondence determination using line features,” IEEE Computer Society Conference on Computer Vision & Pattern Recognition, 1211499 (2003).

[4] Chandra, A., Manojkumar, P. C., Tharun, M. B., et al., “Object detection, 3-D coordinate extraction and pose estimation for an autonomous medical robot,” 2021 5th Inter. Conf. on Computer, Communication and Signal Processing (ICCCSP), 9465501 (2021).

[5] Liu, D., Arai, S., Xu, Y., et al., “6D Pose estimation of occlusion-free objects for robotic bin-picking using PPF-MEAM With 2D images (Occlusion-Free PPF-MEAM) IEEE Access 9, 50857-50871 (2021).

[6] Zhu, J., Pan, Z., Chao, S., et al., “Handling occlusions in video-based augmented reality using depth information,” Computer Animation & Virtual Worlds 21(5), 509-521 (2021).

[7] Zhu, M., Derpanis, K. G., Yang, Y., et al., “Single image 3D object detection and pose estimation for grasping,” IEEE Inter. Conf. on Robotics & Automation, 3936-3943 (2014).

[8] Shu, C., Liang, L., Liang, W., et al., “3D pose tracking with multi-template warping and SIFT correspondences,” IEEE Transactions on Circuits & Systems for Video Technology 26(99), 1-10 (2015).

[9] Mkhoyan, T., Visser, C. and Breuker, R. D., “Parallel real-time tracking and 3D reconstruction with TBB for intelligent control and smart sensing framework,” AIAA Scitech 2020 Forum, 2020-2252 (2020).

[10] Girshick, R., “Fast R-CNN,” 2015 IEEE Inter. Conf. on Computer Vision (ICCV), 1440-1448 (2015).

[11] He, K., Gkioxari, G., Dollar, P., et al., “Mask R-CNN,” 2017 IEEE Inter. Conf. on Computer Vision (ICCV), (2017).