Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Research on the Visual Servoing method based on semantic segmentation results of metal objects in a complex lighting environment

To cite this article: Shunkai Shi and Tingting Wang 2022 *J. Phys.: Conf. Ser.* **2402** 012022

View the article online for updates and enhancements.

Research on the Visual Servoing method based on semantic segmentation results of metal objects in a complex lighting environment

Shunkai Shi[a], Tingting Wang*

Department of Mechanical and Electrical Engineering, Hohai University, HHU Changzhou, China

[a]2801074247@qq.com

*wtt_624@163.com

Abstract. The problem of feature recognition and servo tracking for semantic segmentation results of metal parts under complex and variable lighting conditions is addressed. In this paper, we propose an algorithm that incorporates template matching and image-based visual servoing based on semantic segmentation. The method adopts a geometry-based template matching method to stably identify the image features of the pixel region obtained from semantic segmentation, introduces a feature pyramid dimensionality reduction method to improve the processing speed, and then introduces the extracted image moment features into the image-based visual servoing to operate the robotic arm to perform the servo process on the target.

1. INTRODUCTION

With the increasing level of industrial automation, robots [1] are being used more and more widely in complex work environments. A robot is a machine that can perceive changes in the external environment and complete the corresponding control process according to preset instructions to better complete a specific job instead of using humans [2]. Facing complex operations in dynamically changing scenes, combining a robotic arm with a vision system allows the robotic arm system to perceive the environment and perform movements and operations based on image-processed information.

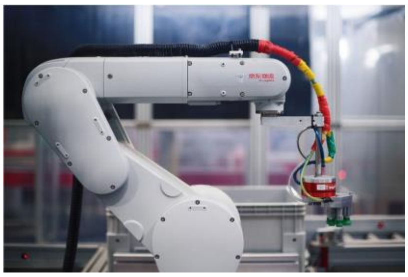


Figure 1: Application of Visual Servoing combined with an industrial robotic arm

The computer is able to process the external environmental variables in real-time and control the motion of the robot arm synchronously through network communication with the robot arm. This control method is called vision-servo [3]. Visual servoing can be classified into three main categories according to the servo characteristics. These are Image-based Visual Servoing (IBVS), Position-based Visual Servoing (PBVS), and Hybrid Visual Servo, which combines the first two [4]. Geometric features such as point features [5], edge features [6], line features [7], or elliptical features [8] are often used in visual servo control. Based on robotic vision systems, visual servoing has become one of the widely used methods that serve the needs of various industrial fields such as micro-assembly [9], rivet hole insertion [10], in-orbit service [11], autonomous capture [12], bio-injection [13], and automatic docking [14].

The scenario in this paper is a complex and variable lighting environment, and the effect of metal targets in such scenarios whose imaging varies with lighting. Image segmentation techniques such as binary segmentation as well as threshold segmentation are difficult to identify the target stably [15] [16]. In recent years deep learning methods have been more and more widely applied to image-related fields with great success in many computer vision applications such as image recognition and target detection. For metallic targets under complex illumination, the semantic segmentation method of deep learning can effectively identify the wanted region of the target, but due to the instability of its pixel region shape, it is difficult to directly extract relevant features for visual servoing. Therefore, this paper proposes a combination of template matching and image-based visual servoing methods based on semantic segmentation results to control the robotic arm for target localization and tracking.

2. SEMANTIC SEGMENTATION

The semantic segmentation method used in this paper is based on the U-Net semantic segmentation model. To facilitate network construction and better generalization, the U-Net structure used in this paper is improved in the enhanced feature extraction part by directly up-sampling twice before feature fusion, and finally obtaining the same height and width of the feature layer as the input image.

The experimental object of this paper is a circumferentially symmetric metal refueling port. The target is photographed under a variety of different lighting conditions. Different camera angles, as well as depths, are also selected, and special cases such as camera defocus and motion blur are also included. A total of 2000 photographs were taken to create the dataset, which was calibrated by using LabelMe to label the top surface of the fuel nozzle.

The results of the model prediction using the images within the dataset show that the recognition targets can be segmented effectively and more accurately in bright, dark, and occluded light environments, but the edges of the segmentation results are not smooth in overly bright and dark environments.

$$
\begin{aligned}
MIou &= \frac{Iou_p + Iou_n}{2} \\
&= \left(\frac{TP}{TP+FP+FN} + \frac{TN}{TN+FN+FP} \right) / 2
\end{aligned} \tag{1}
$$

For the accuracy evaluation criteria of the model, the Miou evaluation index is used in this paper. The average intersection ratio is obtained by summing and averaging the ratio of the intersection of the predicted results and the true values of each category in the model. The final Miou value is 0.9719, which proves that the accuracy of the semantic segmentation model meets the test criteria.

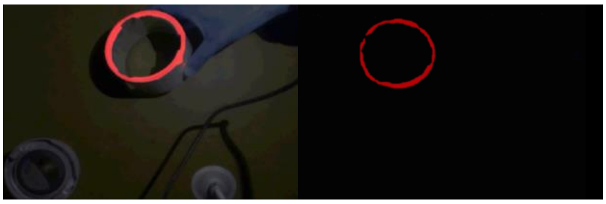


Figure 2: Predicted results in extremely bright and extremely dark environments

The visual information of the target in this case is relatively difficult to obtain. Template matching, etc., is required for visual servo control for subsequent target localization

3.TEMPLATE MATCHING METHOD

In a particularly intense lighting environment, the semantic segmentation pixel area will overflow the target area in a small area, as shown in the figure. The phenomenon of missing color blocks will appear in the pixel region of semantic segmentation under almost no illumination environment. If contour and other information are extracted directly from the prediction result map, the recognition error increases when the light conditions are more limited. Therefore, the method of template matching is adopted in this paper to extract the features of the target accurately.

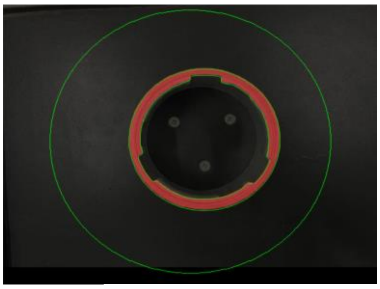


Figure 3: A template for matching

The template matching method adopted in this paper is template matching based on a geometric shape. The prediction results shown in Figure 3 were selected to make the template for matching.

The matching score of similarity is calculated as follows: There are a total of n edge points on the template. In the ROI area to be matched on the collected image, the cosine value of the Angle between the boundary points of the template and the direction vectors of the boundary points of the image area to be matched is calculated as $a_1, a_2, a_3 \cdots a_n$, and the matching value is expressed as:

$$P = \frac{a_1 + a_2 + a_3 + \cdots + a_n}{n} \tag{2}$$

The closer the P value is to 1, the higher the matching degree will be.

In the feature extraction part, three variables can be moved by the rotation and deformation of the template to extract the geometric features of the target. Image features can also be extracted by the contour of template matching results displayed in the image.

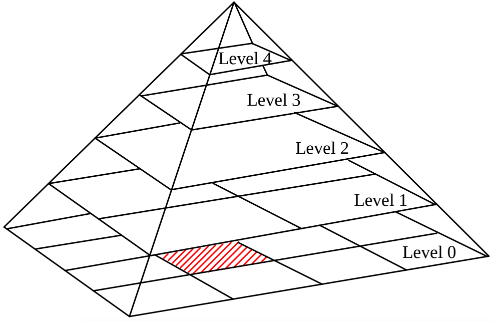


Figure 4: Pyramid dimension reduction method

Because the template needs to be rotated and scaled for matching, and for matching accuracy, the interval between rotation and scaling is small, which directly leads to the high time complexity of the algorithm and poor real-time performance. Given this situation, the pyramid down-sampling strategy is adopted in this paper to speed up the template matching speed. The same sampling dimension is made for the template and the image to be matched. Firstly, rough matching is carried out under the low-resolution image at the top of the pyramid. Secondly, the matching results are mapped to the next layer, and so on, until the original image of the last layer, and then fine matching is carried out. Through pyramid dimension reduction processing, the processing speed of template matching is greatly improved, which can reach about ten frames per second.

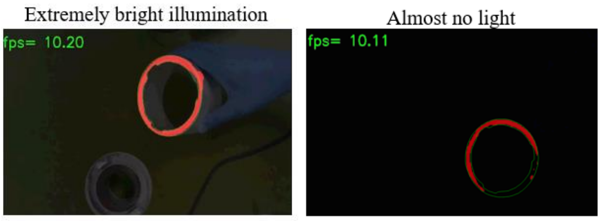


Figure 5: Template matching results

The experimental results of real-time template matching are shown in Figure 5. The geometry-based template matching can locate the recognition target in the image more accurately in both bright light and dark environments and can read the pixel coordinates where the target center point is located and the horizontal and vertical scaling ratio relative to the original template, which can be used as the basis for positional estimation.

4. DESIGN OF VISUAL SERVOING CONTROL METHOD

4.1.Experimental platform construction

Based on the above-mentioned target feature pixel area, a servo controller needs to be designed to perform servo positioning tracking or grasping. The two servo methods have their advantages and disadvantages, and for the experimental scenario of this paper, a more robust servo method is needed, so the solution of this paper is to use image moment features for IBVS fine positioning.

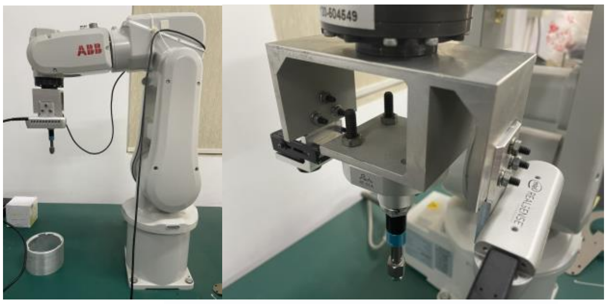


Figure 6: Robotic arm experiment platform

To achieve this first build the basic experimental platform as shown in Figure 6. Using Eye-in-Hand mode, the Real Sense depth camera is fixed to the end of the robotic arm ABBIRB120 by a metal connector. The robotic arm used in the experimental platform consists of a combination of six rotating joints and connecting rods with six degrees of freedom, which can realize the actions that need to be performed in industrial scenarios. The internal camera parameters M and aberration parameters $(k_1,k_2,k_3,p_1,p_2,)$ were determined by camera calibration, and the external camera parameters (chi-square transformation matrix between the robotic arm end-effector and camera coordinate system) were determined by hand-eye calibration.

Finally, RobotStudio is used to control the motion of the robot arm and to provide feedback on the motion parameters of the robot arm. The connection between the robot arm and the computer is established using socket communication so that the data, such as the position, can be transferred.

4.2.IBVS precise positioning

After the initial positioning in the previous section is completed, then according to the image characteristics of the recognition target, the image-based visual servo method is used to control the end of the robotic arm to precisely move to the desired position near the recognition target to complete the servoing process. In order to adapt the whole servo process to different shapes and sizes of servo objects, the control algorithm of the visual servo needs to choose suitable features, control rates, and desired poses for different visual servo objects. For this experimental object, the image moment feature is proposed to be used for fine positioning. After the previous image semantic segmentation and template matching, the closed pixel contour of the upper face of the identified object in the image has been obtained, and the set of all pixel points in the middle of the two edge contours is the target region *O* of the image moment feature as shown in Figure 7.

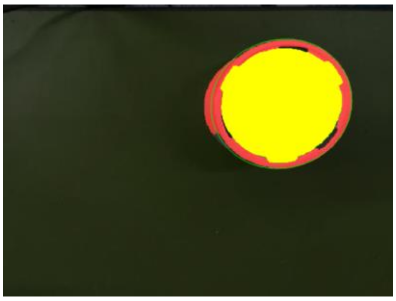


Figure 7: Image moment area *O*

For the yellow continuous region *O* in the image plane, the mathematical expression of its origin moment characteristic is:

$$m_{ij} = \iint x^i y^j I(x,y) dxdy \tag{3}$$

where (*i*+*j*) is the order of the image origin moments, *I*(*x*,*y*) denotes the luminosity at the point (*x*,*y*) on the image plane, and the luminosity function *I*(*x*,*y*) can be expressed as:

$$I(x,y) = \begin{cases} 1 & (x,y) \in 0 \\ 0 & (x,y) \notin 0 \end{cases} \tag{4}$$

The 0th-order origin moments of the binarized image are:

$$m_{00} = \iint x^0 y^0 I(x,y) dxdy = \iint dxdy = 0 \tag{5}$$

Its origin moment then represents the area of the target region.

In addition to the origin moment, another common image moment is the center distance, whose mathematical expression is:

$$\mu_{ij} = \iint \left(x - x_g\right)^i \left(y - y_g\right)^j dxdy \tag{6}$$

In Eq. (17), $x_g = m_{10} / m_{00}$, $y_g = m_{01} / m_{00}$; from Eq. (14) and Eq. (16), it is known that (x_g, y_g) is the regional center coordinate (form center) of objective region *O*. The image origin moments and center distance can be translated into the following mathematical expressions.

$$m_{ij} = \sum_{k=1}^{n} x_k^i y_k^j \tag{7}$$

$$\mu_{ij} = \sum_{k=1}^{n} \left(x_k - x_g\right)^i \left(y_k - y_g\right)^j \tag{8}$$

In the equation, (*x*,*y*) is the image coordinate point with luminosity 1 in the binarized image; *n* is the number of image coordinate points that satisfy the condition, i.e., the area of the target area on the image plane; at this time, $x_g = m_{10} / n$, $y_g = m_{01} / n$.

In the vision servo system, the higher-order moments combined from the image origin moments and center distance, which have translation and rotation invariance and are used to control the spatial degrees of freedom of the robot arm, are called image moment features.

In the image moment-based vision servo system, image moments that represent the center-of-mass coordinates (x_g, y_g), object size m_{00} , and azimuth angle α are generally selected as the four image features, which are used to control the four degrees of freedom of the camera $(v_x, v_y, v_z, \omega_z)$, respectively.

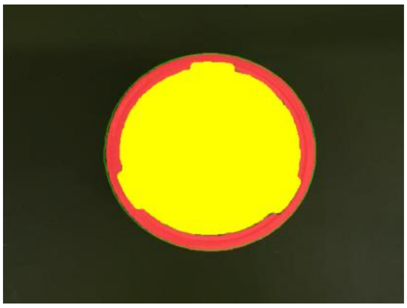
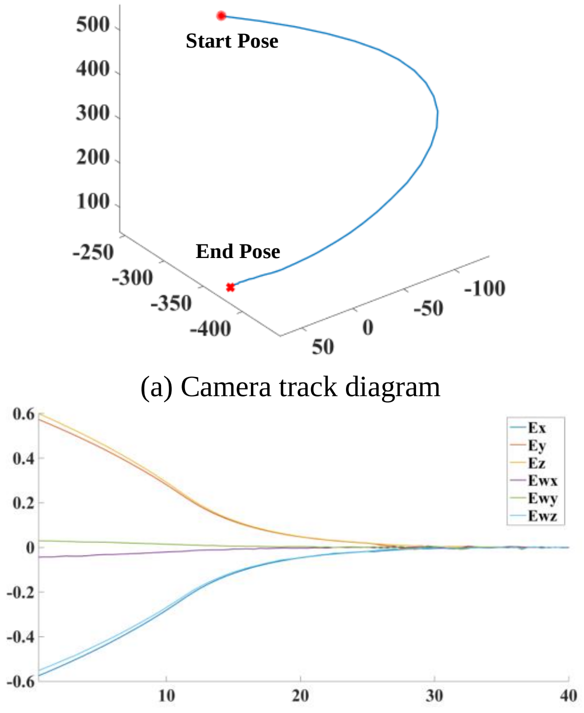


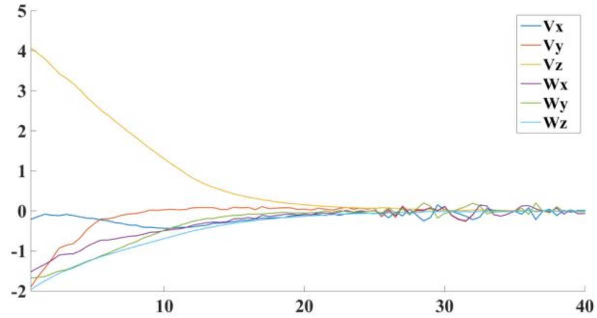
Figure 8: Image moment region of the desired pose

The image moment region at the desired position is shown in Figure 8. The difference between the image moment of the initial posture and the image moment of the desired posture is used as feedback to calculate the instantaneous speed of the camera and control the robot arm so that the camera mounted at its end gradually approaches the final desired posture and finally converges. This is the whole IBVS vision servo fine positioning solution.



(a) Camera track diagram

(b) Feature convergence graph



(c) Camera speed convergence graph
Figure 9: Image moment-based Visual Servoing

Finally, all the above programs, as well as algorithms, are integrated. In this paper, the data transfer between the computer and the robotic arm is implemented through socket communication, and the programs of each part are jointly programmed to conduct the experiments. The resulting graph of a set of experiments is shown in Figure 9, and the graph shows that the final curve reaches convergence, which indicates the success of the servo.

5. CONCLUSION

The proposed approach of incorporating deep learning into the vision servo to effectively cope with the complex and variable lighting environment significantly reduces the dependence on a stable light source during the gripping and docking process of vision servo-controlled robotic arms. The experimental results also demonstrate the feasibility of using a semantic segmentation model to extract target pixel regions that can be incorporated into vision servoing.

REFERENCES

[1] V. Jatla, M. S. Pattichis, C. N. Arge. Image processing methods for coronal hole segmentation, matching, and map classification[J]. *IEEE Transactions on Image Processing*, 2020, 29(01): 1641-1653.

[2] F. Wang, Y. Wu, M. Li, et, al. Adaptive hybrid conditional random field model for SAR image segmentation[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2017, 55(01): 537-550.

[3] B. Xue. Mask R-CNN-based power equipment rust detection[J]. *Computer Systems & Applications | Compute System*, 2019, 28(05): 248-251.

[4] Y. Gao, J. Guo, X. Li. Piglet image instance segmentation method based on deep learning[J]. *Transactions of The Chinese Society of Agricultural Machinery*, 2019, 50(04): 179-187.

[5] Y. Liu, Z. Meng, Y. Zou, et al. Visual object tracking and servoing control of a nano-scale quadrotor: system, algorithms, and experiments[J]. *IEEE/CAA Journal of Automatica Sinica*, 2021, 8(02): 344-360.

[6] R. Szeliski. Computer vision: Algorithms and applications[M]. *Springer-Verlag*, 2011.

[7] D. A. Forsyth, J. Ponce. Computer vision: A modern approach[M]. Prentice Hall, 2002.

[8] M. S. Nixon, A. S. Aguado, Feature extraction and image processing[M]. Academic, 2008.

[9] F. Ireta, A. I. Comport. Point-to-hyperplane RGB-D poses estimation: fusing photometric and geometric measurements[J]. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, Daejeon, South Korea, 2016: 24-29.

[10] E. R. Davies. Radial histograms as an aid in the inspection of circular objects[J]. Control Theory & Applications IEE Proceedings D, 1985, 132(04): 158-163.

[11] R. Gonzalez, R. Woods, Digital image processing, 3rded[M]. *Prentice Hall*, 2008.

[12] C. Lazar C, A. Burlacu. Predictive control of non-linear visual servoing systems using image moments [J]. *IET Control Theory and Applications*, 2012, 6(10): 1486-1496.

[13] F. Chaumette, S. Hutchinson. Visual servo control. I. Basic approaches[J]. *IEEE Robotics & Automation Magazine*, 2006, 13(4): 82-90