Grasping position recognition of spatial non-cooperative targets based on active light detection

Zhihong Chen¹,²*, Junqin Lin¹,², Yanbo Wang¹,², Binyan Liang¹,², and Yao Wang¹,²

¹ Beijing Institute of Precision Mechatronics and Controls, Beijing, 100076, China
² Laboratory of Aerospace Servo Actuation and Transmission, Beijing, 100076, China
*Corresponding author’s e-mail: zhihongch@laset.com

Abstract. The current mainstream operation methods are mainly oriented to cooperation goals and pre-designed signs, which are difficult to meet the needs of complex tasks. Taking the non-cooperative target as the research object, this paper proposes a method of active light source detection combined with different multi-source data processing networks. When the image brightness is relatively high, the residual convolutional neural network is used to identify the target capture area in the multi-source data; when the image brightness is dark, the PointNet++ network is used to identify the capture area from the point cloud data, thereby adaptively achieving a better segmentation effect. The final experiment proved the effectiveness of this method.

1. Introduction
With the development of computer technology these years, computer vision, which is an important research field of artificial intelligence, has been widely used in all walks of life. Vision-based robotic arm grasping has gradually become a current research hotspot. From the previous research work [1-4], a vision-based robotic arm grasping task was introduced. The common point is that they all use traditional feature extraction methods to process image information. These methods generally use artificially designed features to recognize and capture targets, but they are affected by factors such as the shape, size, angle change, and external illumination of the target object. The generalization ability of artificial features is not strong, the robustness is poor, and it is difficult to adapt to new objects. The concept of deep learning [5] was first proposed by Hinton et al. in 2006. After Krizhevsky and others used the deep learning method [6] to achieve breakthrough results in the 2012 ImageNet competition, there is an explosive development in deep learning technology. Compared with the traditional artificial feature-based methods, the advantage of deep learning is that there is no need to design specific artificial features, but a general learning process is used to enable the model to learn features from large-scale data independently [7].

Deep learning also has its applications in robotic grasping tasks. In [8], a deep neural network is used to directly detect the position of the palm and fingers of the grasping hand from the local target view containing stable grasping, and then complete the grasping task. Lenz et al. [9] used the grab function to predict the scores of all possible positions of the target. The smooth grab function made the model more robust to uncertain grab positions. The grab function was controlled by the convolutional neural network which was obtained by online learning. Johns et al.[10] proposed a two-step cascade
detection system, which detected the optimal grasping position of the target object by processing RGB-D multi-modal information, and obtained a better grasping effect.

This paper proposes a detecting method for the optimal grasping position of a robotic arm that combines active light source detection with different multi-source data processing networks. When the brightness is relatively high, the residual convolutional neural network is used to identify the target capture area in the multi-source data to achieve a better segmentation effect; but when the brightness is relatively low, the effect is significantly reduced. At this time, the point cloud segmentation network is used to achieve better segmentation. Compared with the features set up manually, the features learned by the deep neural network method can adapt to new objects that have never appeared in the training set, and have stronger generalization and stability. Compared with traditional recognition methods based on images or point clouds, this method can be applied to complex lighting environments in space, and has strong environmental adaptability and robustness.

2. Grab position segmentation system frame

2.1. System framework

The vision-based manipulator grasping system is shown in Figure 1. The system is composed of a robotic arm, a vision sensor, a visible light brightness sensor, an image processing module, a main control system, and a robotic arm servo drive. The system takes the main control part as the core, perceives the brightness information of the surrounding environment through the visible light sensor, obtains the image information through the visual sensor, and processes the image information by the image processing program to realize the three-dimensional positioning of the grasping position, and then the grasping operation.

![Figure 1. Manipulator grasping system based on vision](image)

The framework of the image processing program module is shown in Figure 2. This framework adopts the method of active light source detection combined with different multi-source data processing networks. Using RGBD camera to collect environmental data, image data and point cloud data of the environment can be obtained. The visible light brightness sensor uses the brightness information of the surrounding environment to determine whether the brightness reaches the threshold. When the brightness is relatively high, the residual convolutional neural network is used to identify the target capture area in the multi-source data; when the brightness is dark, the point cloud capture position segmentation network is used to achieve the segmentation of the capture area.
2.2. Coordinate system definition

For the task of positioning the grasping area, this article uses the following formulas to express the grasping posture and grasping quality in the world coordinate system and the image coordinate system:

1) Global coordinate system

As shown in Figure 3, this article uses $G_r = (P, \Phi, W, Q)$ to represent the grabbing action, where $P = (x, y, z)$ represents the center position of the grabbing action in the global coordinate system; $\Phi$ represents the rotation angle of the arm around the $Z$ axis when grasping down; $W$ is the width of the robot arm that needs to be opened; and $Q$ represents the quality evaluation score of the grasping inspection.

2) Image coordinate system

This article uses $G_i = (P, \Phi, W, Q)$ to represent a grabbing action, where $P = (u, v)$ represents the grabbing action at the center of the image; $\Phi$ represents the rotation angle of the robotic arm around the $Z$ axis; $W$ is the width of the robot arm that needs to be opened; $Q$ represents the quality evaluation score of the grasping inspection. Figure 4 shows the prediction result of a network model, which is the output grab box under the image coordinate system, and the grab center coordinates, width and other information are respectively marked.
3) Formula conversion:
Taking into account the need for actual grasping, we can convert the two-dimensional grasping pose to the robot coordinate system pose through coordinate transformation. The transformation formula is as follows:

$$G_r = T_{rc}(T_{ci}(G_i))$$

Among them, the $T_{rc}$ matrix realizes the conversion from the camera coordinate system to the machine coordinate system, and the $T_{ci}$ matrix realizes the conversion from the image space to the 3D space.

3. Multi-source information-based grasping location recognition network architecture
The main architecture of the multi-source information-based capture area segmentation network module in this paper is a convolutional neural network, which is also the most commonly used network architecture for processing problems based on multi-source information. On the basis of using the traditional convolutional network, we also added five residual layers to better learn the objective function of the network.

Figure 5 shows the convolutional neural network architecture for grabbing detection. The network includes 3 convolutional layers, 5 residual layers, and 3 convolutional replacement layers.

The network input image can be a multi-channel image, such as an RGB image or a depth image, or a combination of the two can be used as a network input. The required input image size is 224×224. For input of a specific size, we also designed a preprocessing module for the original image taken by the camera to process the original input into a size suitable for network input.

The network finally outputs 3 pictures for visualization of the results, which are the quality of the grasping, the angle of grasping, and the width of the robot arm.

From the mathematical expression $G_i = (P, \Phi, W, Q)$ in the image coordinate system, these three pictures respectively represent the $Q$, $\Phi$ and $W$ corresponding to each point of the network input picture. The capture $P$ in the formula is the coordinate of the point. We chose the point with the highest crawl quality in the picture as the final crawl prediction box.
4. Point cloud capture location recognition network architecture

The architecture of the point cloud capture and segmentation network module in this article is shown in Figure 6. The network input is the down-sampled point cloud data of the object to be segmented (the point cloud data of the training set is too dense, and the amount of calculation is reduced by down-sampling, as shown in the figure 6. The network first uses the SA module (set abstraction layer) to use the FPS algorithm for the input point cloud data of each SA layer to extract its local features, and output the corresponding central point set.

First, the FPS algorithm randomly selects a point, adds it to the center point set, selects the point farthest from the point in the center point set to the center point set, and iteratively selects the center point (the point selected later needs to be compared with all the points in the previous center point set. Do the distance calculation metric distance) until the number of concentrated points in the center point reaches the threshold.

Secondly, through multi-layer FPS feature extraction, the network obtains a set of center points (including 64 points) containing global features. This set of points is then segmented and classified into two categories (i.e., classified into handles and non-handles).

Finally, use the segmented and classified global feature center point set to interpolate with the output point set of each SA layer in the feature learning part to realize the transfer of global features, and finally segment the point cloud data of the original network input size. The position of the handle.

Figure 6. Network architecture

After processing the identified point cloud, the capture parameters of non-cooperative targets can be obtained: \( G_r = (P, \Phi, W, Q) \) means grabbing action; Among them:

\[
p = \frac{\sum_{i=1}^{n} (x_i, y_i, z_i)}{n}
\]

represents the center position of the grasping action in the world coordinate system, and \( n \) is the number of divided point clouds Number, \((x_i, y_i, z_i)\) is the space coordinate value of each point cloud;

\[
\Phi = \min_{c,n,||n||=1} \sum_{i=1}^{n} ((x_i - c)^T n)^2
\]

means that the robotic arm is grabbing downward The rotation angle of the time around the Z axis, that is, the main direction of the segmented point cloud; \( W \) is the width of the robot arm that needs to be opened, that is, the maximum width of the platform perpendicular to the main direction of the point cloud; And \( Q \) represents the quality evaluation score of the crawl detection.

5. Experiment and results

The data set involved in this system contains point cloud, RGB and its corresponding segmentation area labeling. The data set for constructing the real scene is relatively large, so this data set is generated by the V-rep simulation environment which contains the target object and the target. The
two-dimensional image of the handle, as shown in Figure 7, is an RGB image and depth image obtained in a simulated environment.

![Figure 7. Example of RGB image and depth map obtained in the simulation environment](image)

In the simulation environment, target objects with different angles and colors and their label data are randomly generated, 90% of which are used for training the network, and the remaining 10% is used for network testing. Training and testing use a data set with a data size of 1000. We use the intersection ratio between the predicted segmentation result and Ground Truth (IOU, the intersection ratio between 0 and 1, the intersection of the two/the union of the two Set) as a technical evaluation index for grabbing and segmenting. When the IOU value is larger, its accuracy is higher.

![a. ground truth b. Multi-source information segmentation effect in bright scene c. Point cloud segmentation effect in bright scene](image)

![d. Multi-source information segmentation effect in dark scene e. Point cloud segmentation effect in dark scene](image)

Figure 8. The segmentation effect of the capture area of different networks in different environments

The experimental results are shown in Figure 8, a is the Ground truth, b is the segmentation effect of multi-source information in a bright scene, and c is the segmentation effect of point cloud in a bright scene. It can be seen that the segmentation effect of multi-source information in this environment is slightly Better; d is the segmentation effect of multi-source information in the dark scene, and e is the point cloud segmentation effect in the dark scene. It can be seen that the segmentation effect of the point cloud network in this environment is better. The final test result found that the average IoU index of the network prediction results of this system is 93.2%, which is slightly higher than the 91.4% of the network segmented by multi-source information and the 89.7% of the network segmented by only the point cloud.
6. Conclusions
This project proposes a method of active light source detection combined with different multi-source data processing networks. By capturing images and point cloud data of the recognition area, the visible light brightness sensor is used to use the brightness information of the surrounding environment, and the system is guided to adopt different segmentation networks. The method we propose can achieve intelligent segmentation and recognition of the target, and provide environmental information for the subsequent intelligent capture and planning of the robotic arm. In the simulation environment, the non-cooperative target capture location recognition accuracy rate is 93.2%, which verifies the effectiveness of the algorithm.

References
[1] Kemp, C. C. , Edsinger, A. , & Torres-Jara, E. . (2007). Challenges for robot manipulation in human environments [grand challenges of robotics]. Robotics & Automation Magazine IEEE, 14(1), 20-29.
[2] Arnau R, Guillem A, Francesc M, Carme T.(2014). Learning RGB-D descriptors of garment parts for informed robot grasping. Engineering Applications of Artificial Intelligence. 35:246-258.
[3] Yan, X. , Hsu, J. , Khansari, M. , Bai, Y. , Pathak, A. , & Gupta, A. , et al. (2017). Learning 6-DOF Grasping Interaction via Deep Geometry-aware 3D Representations. 2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE.
[4] Lin, Y. , & Sun, Y. . (2015). Robot grasp planning based on demonstrated grasp strategies. The International Journal of Robotics Research, 34(1), 26–42.
[5] Hinton, G. E. , & Salakhutdinov, R. R. . Reducing the dimensionality of data with neural networks. Science, 313.
[6] Krizhevsky A, Sutskever I, Hinton G E. (2012) ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems. Cambridge, USA: MIT Press, 2012: 1097-1105.
[7] Lecun, Y. , Bengio, Y. , & Hinton, G. . (2015). Deep learning. Nature, 521(7553), 436.
[8] Varley, J. , Weisz, J. , & Allen, P. . (2015). Generating multi-fingered robotic grasps via deep learning. 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE.
[9] Lenz, I. , Lee, H. , & Saxena, A. . (2013). Deep learning for detecting robotic grasps. The International Journal of Robotics Research, 34(4-5).
[10] Johns, E. , Leutenegger, S. , & Davison, A. J. . (2016). Deep Learning a Grasp Function for Grasping Under Gripper Pose Uncertainty. IEEE/RSJ International Conference on Intelligent Robots & Systems. IEEE.