Object Recognition and Location Based on Mask R-CNN and Structured Light Camera

Weipeng Mao, Hongyang Yu
Research Institute Electronic Science and Technology University of Electronic Science and Technology of China, Chengdu, China
hyyu@uestc.edu.cn

Abstract. In the indoor construction site scene, binocular cameras rely on natural light in the environment for collecting images. However, with the influence of environmental factors such as changes in illumination angle and changes in illumination intensity, the effect of the binocular vision algorithm is drastically reduced in the case of darker illumination. In order to make the robot recognize the object better and judge the position of the object, this paper proposes a method based on Mask R-CNN model and structured light camera for object recognition and localization. The Mask R-CNN model is used to segment the RGB image, and then extracting the depth information of the target object, and calculating the distance between the target object and the robot through the principle of triangulation. The experiment shows that the proposed object recognition and positioning system can still complete the recognition and location of the trained object in the case of dark indoor light.

Keywords: Mask RCNN, Instance segmentation, Structured light camera, Computer vision.

1. Preface

With the advent of the artificial intelligence era, robots are increasingly used in a wide range of industries. In indoor construction site scenarios, robots can reduce labor and save a lot of time. Vision can provide a large amount of information for robots, so it is important to study vision-based robot object recognition and positioning systems. Traditional depth cameras based on binocular vision are very sensitive to ambient lighting. In the case of strong and dark lighting, the effect of the binocular vision algorithm drops dramatically. At the same time, it is not suitable for monotonously lacking texture scenes, such as in the absence of visual features, there will be difficulties in matching.

In recent years, deep learning has developed rapidly in the field of computer vision. The single-task network structure has gradually become less noticeable. Instead, it is an integrated and complex multi-tasking network model. The representative is the instance segmentation model. Instance segmentation is a more comprehensive problem that combines target detection, image segmentation, and image classification. Yi L proposed an end-to-end instance partitioning full convolution solution FCIS [1] (Fully Convolutional Instance-aware Semantic Segmentation). FCIS improved FCN[2] (Fully Convolutional Networks for Semantic Segmentation), specifically canceled ROI-Pooling (Region of Interest Pooling), replaced by polymerization of ROI region, and also eliminated the fully connected layer (FC layer), replaced by the softmax classifier, and finally the same feature map is used for image segmentation and image classification. However, FCIS sometimes has overlap phenomenon in some instance segmentation tasks.

In order to solve the above problems, this paper uses Mask R-CNN [3] model combined with structured light camera to identify and locate objects. The Mask R-CNN model completes the instance segmentation of the image. It is based on Faster R-CNN [4-6], but it replaces the ROI-Pooling layer with ROI Align, which solves the pixel deviation problem in ROI-Pooling. At the same time, it adds a parallel FCN layer (MASK layer) to produce a precise pixel mask. The structured light camera does not depend on the color and texture of the object and adopts the method of actively projecting the known pattern to complete the fast and robust matching feature points, which can achieve higher precision, and the effect is still very strong in the case of indoor dark light. Thus the entire system can perform precise instance segmentation and positioning functions.

2. Robotic Visual Cognition System

2.1 System Principle

The system flow of this paper is shown in Fig. 1.

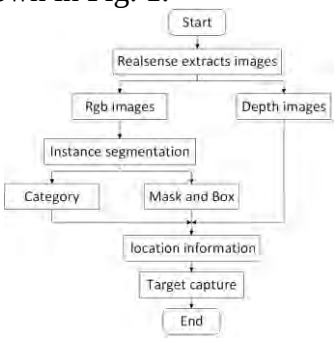

Fig. 1 System flowchart

System design ideas: This paper uses Intel's Realsense SR 300 [7] (hereafter referred to as SR300) camera for image acquisition, SR 300 can obtain RGB and DEPTH images as system input. The RGB image is segmented using Mask R-CNN to obtain object classification and accurate pixel mask. Then according to the Mask of the target object, we can get the depth information of the target object and finally calculate the distance of the target position.

2.2 Mask R-CNN

Mask R-CNN is an Instance Segmentation algorithm proposed by He Kaiming. It is an integrated and complex network model. It mainly accomplishes three tasks: target detection, target classification, and target segmentation. The target detection draws a detection frame for each object in the image. The target classification distinguishes the specific category of each object. The target segmentation distinguishes the foreground and the background at the pixel level for each target. The basis of Mask R-CNN is Faster R-CNN, but Mask R-CNN adds a masking task based on Faster R-CNN. At the same time, because ROI-Pooling uses the rounding quantization operation when the feature map ROI is mapped back to the original image ROI, there will be spatial misalignment, which causes errors, so Mask R-CNN replaces ROI-Pooling with ROI Align. ROI Align does not use rounding quantization but uses bilinear interpolation to achieve pixel-level alignment.

Mask R-CNN is mainly composed of five parts, namely feature extraction network, feature fusion network, region proposal network (RPN), ROI Align and final functional network. The structure of Mask R-CNN is shown in Fig. 2.

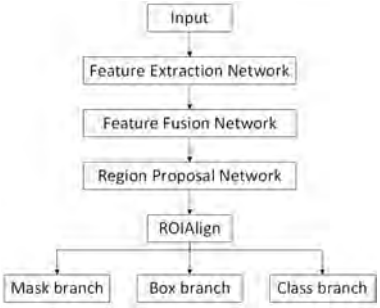

Fig. 2 Mask R-CNN structure chart

The feature extraction network is the backbone network of Mask R-CNN. It is used to extract the feature maps of images. It can use VGG16[8], VGG19, GoogleNet[9], ResNet50[10-11], ResNet101, etc. In this paper, the backbone network uses ResNet101 network. It is divided into 5 stages, which are respectively recorded as [C1, C2, C3, C4, C5], corresponding to 5 different scales of Feature map, used to establish the feature pyramid of FPN[12] network, and get new features, respectively recorded as [P1, P2, P3, P4, P5]. Actually, it does not use P1, because it takes a lot of time to calculate the feature map corresponding to C1. P6 is obtained by downsampling P5, and P6 is used to replace P1. For i =2, 3, 4, 5, 6, the specific correspondence is as shown in equation (1). Conv () is the convolution operation, sum() is the summation operation, upsample() is the upsampling operation, and downsample() is the downsampling operation.

$$\begin{cases} P_i = conv(sum(upsample(P_{i+1}), conv(C_i))) \\ P_5 = conv(conv(C_5)) \\ P_6 = downsample(P_5) \end{cases} \quad (1)$$

The feature maps are sent to the RPN, which uses a sliding window to scan the image and find the area where the target exists. The area scanned by the RPN network is called Anchor, which is a rectangle on the image. The RPN network generates two outputs for each Anchor. The first output is the category of Anchor. It is judged whether the Anchor is the foreground or the background. If it is the foreground, that means there may be an object in the Anchor box; the second output is the correction data for the border coordinates. As a foreground Anchor, it means that there is an object in the Anchor box, but it may not be perfectly at the center of the box. Therefore, the RPN network outputs the modified percentages△ x,△ y,△ w and△ h of x, y, w, and h for the Anchor to correct the Anchor box to better fit the object. The correction formula is as shown in equation (2).

$$\begin{cases} x = (1+\square x)\square x \\ y = (1+\square y)\square y \\ w = (1+\square w)\square w \\ h = (1+\square h)\square h \end{cases} \quad (2)$$

After prediction through the RPN network, a series of Anchors can be obtained and their position and size are corrected. If there are multiple Anchors overlapping each other, the Anchor will be filtered by non-maximum suppression (NMS) to get the Anchor with a higher foreground score and pass it to the next stage.

Before the next stage, the size of the Anchor box needs to be adjusted to a fixed size. The specific operation in the Faster R-CNN is to crop a part of the feature map by ROI-Pooling and then readjust it to a fixed size. However, there are two problems with ROI-Pooling. The first one is that ROI-Pooling is directly obtained by rounding off, so it is possible that the output after the ROI-Pooling operation is not matched with the ROI on the original image. The second is to perform a rounding operation again when converting the features corresponding to each ROI into a fixed size. Although the rounding operation has little effect on the ROI classification, since the features obtained for each ROI are not aligned with the ROI, it is influential for the pixel-level prediction target. In this paper, Mask R-CNN uses ROI Align instead of ROI-Pooling. For the first problem, ROI Align no longer performs rounding. For the second problem, bilinear interpolation is used to more accurately find the features corresponding to each block. The conclusion is that the rounding operation is no longer used in the ROI Align operation, and the features obtained for each ROI in the ROI Align operation can better align the ROI area on the original image.

Finally, the bounding-box recognition (classification and regression) and mask prediction are performed separately for each ROI by using the head network. The head structure is shown in Fig. 3.

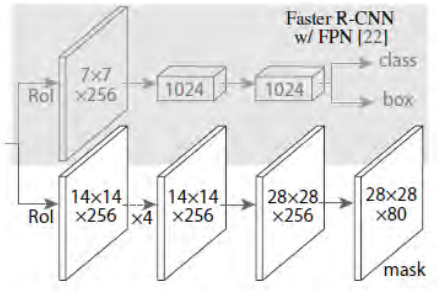

Fig. 3 Head network

The head structure can expand the output latitude of ROI Align, making the predicted mask more accurate. In the training process of the mask branch task, Mask R-CNN does not use SoftmaxLoss like FCN. Instead, we output a total of K Mask prediction charts, each corresponding to a category, and trained with the average binary cross-entropy loss. However, when training the Mask branch, for an ROI belonging to the kth category, Lmask only considers the kth mask and does not consider other

mask inputs. The effect of Mask R-CNN for instance segmentation is shown in Fig. 4. The original image is from the COCO dataset [13].

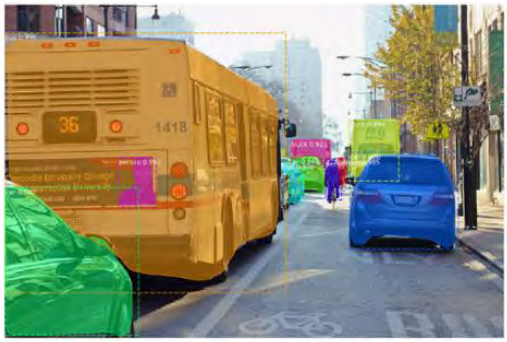

Fig. 4 Instance segmentation

2.3 Principle of Structural Light

The camera model used in this paper is Intel's Realsense SR300, and the SR300 camera is the second generation of front camera from Intel. As an important product of Intel RealSense technology, the SR300 camera integrates 3D depth and 2D lens modules to give the device a depth vision similar to the human eye. The SR300 camera integrates three lenses, including an RGB camera, an IR laser projector, and an IR camera lens. Through these three kinds of lenses, the infrared rays reflected by the objects in front of the camera are detected to infer the object depth information.

The infrared laser lens in the SR300 camera emits structured light to the object, and the structured light is reflected on the surface of the object, and the reflected structured light is received by the infrared camera. Due to the difference in distance between the infrared laser lens and the surface of the target object, the position and shape of the structured light pattern received by the infrared sensor will be deformed. According to these deformation conditions and the principle of triangulation we can calculate the depth of the object. The principle of triangulation is shown in Fig. 5.

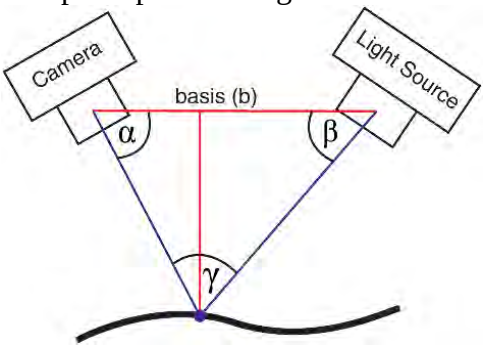

Fig. 5 Triangulation principle

However, before performing ranging, the depth image and the color image must be aligned, because the spatial coordinate system of the DEPTH image data is different from the RGB image data. The origin of the DEPTH image data is the infrared camera, and the origin of the RGB image data is RGB Camera, so there will be some error. The RGB image is shown on the left and the DEPTH image is shown on the right in Fig. 6.

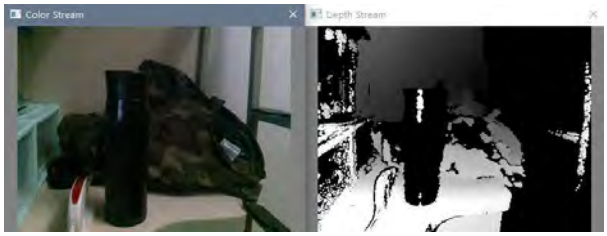

Fig. 3 RGB and DEPTH images

3. Experimental Results and Analysis

3.1 Object Recognition Experiment

The COCO dataset is an image recognition, segmentation and caption dataset. The COCO dataset is characterized by multiple objects in each image, a total of 80 object categories, more than 2 million instances, and more than 300,000 images. The COCO data set is released by Microsoft Corporation of the United States. The annotation information of the image includes not only the object position information and the object category, but also the semantic text description of the image. With the open source of COCO dataset, the field of image segmentation semantic understanding has made great progress in recent years. At the same time, COCO dataset has become a recognized "standard" dataset for performance evaluation of image semantic understanding algorithms. This paper uses the 2014 COCO data set for model training.

In this paper, the CPU of the experimental device is i7-4790k, and the GPU is GTX 1080 Ti. The official implementation of Mask R-CNN is in the framework of caffe2. In this paper, Tensorflow and Keras are used to implement, and some corresponding adjustments are made. First, the image is adjusted to the same size for training. Second, using the official 0.02 learning rate leads to a gradient explosion. It is found through experiments that using gradient clipping and reducing the learning rate can make the model converge faster. Third, some data sets provide bounding boxes, while others only provide mask. So this paper ignores the bounding box of the dataset and regenerates the bounding box for each target. The specific approach is to select the smallest box containing all the pixels of mask as the bounding box. The experimental effect diagram is shown in Fig.7.

Fig. 7 Instance segmentation effect

Table 1 Mask R-cnn performance evaluation

Implementation	AP	AP50	AP75	APS	APM	APL
Official	35.7	58.0	37.8	15.5	38.1	52.4
Implementation of this paper	36.8	54.5	40.7	19.0	43.4	54.1

For speed, the test device GPU of the author [3] is a Nvidia Tesla M40 with an average speed of 195ms for processing one picture, and the experimental device GPU of this article is a GTX 1080Ti, and its average speed is 656ms.

3.2 Structural Light Ranging Experiment

In this paper, the object recognition and positioning system is composed of Mask R-CNN model and SR300 camera, and the object ranging experiment is carried out by using the triangulation algorithm. For the object ranging experiment, the distance measurement experiment is carried out in the range of 200~1200mm. The results show that the ranging error in this range is less than 0.5%. The experimental results are shown in Fig. 8.

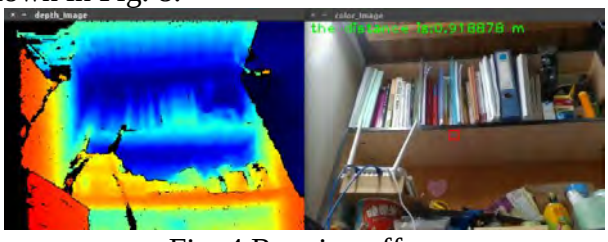

Fig. 4 Ranging effect

3.3 Experiment Analysis

It can be seen that the detection effect of Mask R-CNN on the small and medium-sized objects has been greatly improved, and the APS has increased by 3.5 percentage points, and the APM has increased by 5.3 percentage points. In the object ranging experiment within 1.2m, the systematic error does not exceed 0.5%, which satisfies the data given by Intel Corporation officially. However, in the long distance ranging experiment, the error is large. There are two reasons. The first point is due to The Mask R-CNN model is not accurate enough for far-reaching objects. The second point is that the SR300 camera is a front-facing camera with insufficient sensing depth. It can be improved with Intel's latest generation D435 camera. The minimum sensing depth of the D435 is about 0.11m and the maximum sensing depth is 10 meters.

4. Conclusion

In the indoor construction site scene, this paper proposes an object recognition and localization method based on Mask R-CNN model and structured light camera, and optimizes the parameters of Mask R-CNN model. At the same time, in the object ranging experiment within 1.2m, the data shows that the system error does not exceed 0.5%. In terms of running speed, in the case of using a GTX 1080Ti, the running speed can reach 619ms to process one picture. The experimental results show that the system can better accomplish object recognition and positioning tasks compared with the system using FCIS model and binocular camera in the dark scene of indoor construction site.

References

[1]. Li Y, Qi H, Dai J, et al. Fully Convolutional Instance-Aware Semantic Segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2359-2367.

[2]. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.

[3]. He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE, 2017: 2980-2988.

[4]. Girshick R, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]// IEEE Conference on Computer Vision & Pattern Recognition. (CVPR) Columbus, USA, 2014: 580-587.

[5]. Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[6]. Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//Advances in neural information processing systems. 2015: 91-99.

[7]. Keselman L, Woodfill J I, Grunnet-Jepsen A, et al. Intel RealSense Stereoscopic Depth Cameras[J]. 2017. https://arxiv.org/abs/1705.05548.

[8]. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Science, 2014. https://arxiv.org/abs/1409.1556.

[9]. Szegedy C, Liu N W, Jia N Y, et al. Going deeper with convolutions[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, 2015: 1-9.

[10]. Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-Resnet and the impact of residual connections on learning［J /OL］.［2016-08-23］.https: / /arXiv. org /abs /1602.07261.

[11]. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[12]. LIN T Y, DOLLAR P,GIRSHICK R,et al. Feature pyramid networks for object detection [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR).Honolulu,USA,2017:936-944.

[13]. Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//European conference on computer vision. Springer, Cham, 2014: 740-755.