Object Detection on Video Images Based on R-FCN and GrowCut Algorithm

Kousuke Mouri, Huimin Lu, Joo Kooi Tan, and Hyoungseop Kim
Kyushu Institute of Technology, Japan

***Abstract*—Since the declining birthrate and the aging of society, there is concern about the labor shortage in Japan. There is a movement to compensate for the labor shortage by automation of factories by robots. Automation technique is wildly promoted in logistics industry, while there is few studies in objects picking. To solve this issue, we develop an image detection scheme for robotics picking from a video image. It is difficult to recognize and grasp different types of objects in robot vision field. Therefore, in the proposed method, object detection and object recognition method are proposed using Region-based Fully Convolutional Networks that is a type of object detection using deep learning. After detecting the object individually, final target object can select by applying the GrowCut algorithm. As a result, we achieve 0.6773 of the average precision and 0.6395 of Intersection over Union as the segmentation result respecively.**

Keywords—Object Detection, GrowCut, Region-based Fully Convolutional Networks.

I. INTRODUCTION

Currently, in Japan, since the declining birthrate and the aging of society progresses, there is concern about the labor shortage in the productive age population. The production age population has continued to increase after the war, reaching 87.26 million in 1995. However, it decreased to 75.28 million in 2015, it is estimated that the population will be decreased to 45.29 million in 2065 [1]. There is a trend to compensate for labor shortage due to the decrease in the productive age population by automation of factories using robots.

One of the industries in which the labor shortage is getting worse is the logistics industry. The cause of the labor shortage in the logistics industry is that the market size of ecommerce has expanded, in addition to the decrease in the productive age population due to the declining birthrate and aging population. Furthermore, due to the expansion of the market size of ecommerce, there is an increasing need for diversification of items ordered by consumers and shortening of product delivery time, so it is necessary to improve work efficiency. This paper aim to develop an automatic ordering products for various products, managing inventory, stocking inventory, picking, packing, delivering sorting, and delivering by robots. Picking is the operation of picking up shelved items and storing them exactly in the box on the line. The cause of difficulty in automating picking in the logistics industry is that it is necessary to work on a large variety of products under different conditions, rather than often working on the same type of objects repeatedly under the same conditions. Therefore, by developing a picking method of various kinds of objects by the robot, automatic picking system expect to supplement the labor shortage in the logistics industry and efficiently automate the flow from product arrival to shipment [2, 3].

In this paper, in order to automate picking operation using the robot, we developed detection and recognition of the target object and extraction method of the object region on a video image. The organization of this paper is organized as the following. Section 2 describes the proposed method of image processing for object detection and object extraction. Section 3 shows the experimental results and discussions, and Section 4 concludes this paper.

II. PROPOSED METHOD

A. Flow of Image Processing

In this paper, we perform an object detection method using RGB-D image using Kinect for Windows v2 (Kinect v2) and segmentation of the region of the target object.

As an overall flow, we first use a Kinect v2 system to acquire a RGB images and depth images. The RGB image obtained from Kinect v2 is inputted to Region-based Fully Convolutional Networks (R-FCN)[4] which is the object detection method, classification is performed after detecting the target object. After that, we divide the foreground and background region using the GrowCut method into the detected object region, and extract the detailed region of the target object. The outline of object recognition method by deep learning used in this paper shown in below.

B. R-FCN

R-FCN is an improvement over R-CNN [5] that proposed as an object detection method using a convolution neural network, and its structure shown in Fig.1. We use Region Proposal Network (RPN) for extracting object region candidate from feature map. When an image input to a convolution neural network learned beforehand using the large-scale object class recognition dataset, and a feature map is calculated. Then, the feature quantities of each area candidate are input to the classifier. Then the object class of each candidate area is predicted. Here, using a RoI pooling, we introduce a score map representing the relative position and use a Position-sensitive Score Maps to classify. Thereafter, the bounding box for the same object unified by using the non-maximal suppression.

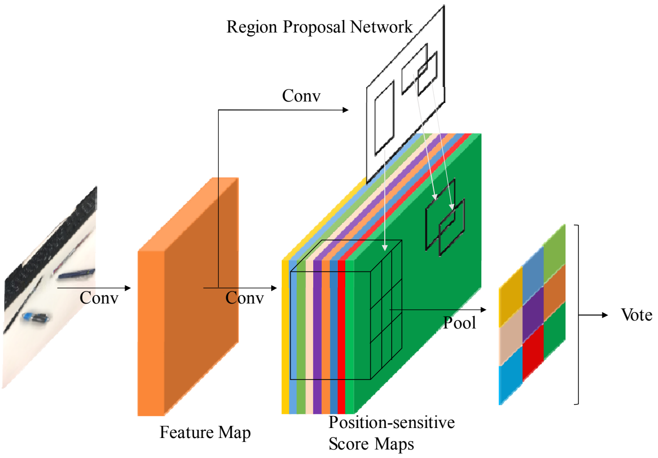


Fig.1 Structure of R-FCN.

C. Region Proposal Network

For each local region of the feature map of the convolution layer, a plurality of bounding boxes with scores of object likeness are introduced into the RPN. Therefore, a combination of a regression network that predicts the parameters of the bounding box and a classification network that predicts the presence or absence of the object is used [6].

In the regression network, prepare k anchor boxes whose shapes have been determined in advance. When predicting the bounding box, it outputs a vector of $4k$ dimension including relative position and aspect ratio from each anchor box. In the classification network, since two classes of presence or absence of an object are judged by each anchor box, $2k$ dimension is outputted.

D. Region Segmentation

The segmentation of the object area using GrowCut method performed. The GrowCut method is a method of giving seed points of the foreground and seed points of the background to the input image and using the foreground as the object of segmentation. It can compared to how the bacteria erode. Bacteria attempt to erode from neighboring pixels from pixels given seeds of labels representing foreground and background. The neighborhood pixels are protected from being eroded by bacteria. The erosion ability of bacteria is expressed as "confidence degree of label", and by using the relationship between erosion power and the defensive power of surrounding pixels, all pixels erode into either the foreground or background bacteria by generation. The defense force is the color difference between adjacent pixels, and the initial value of the erosion force is 1 for the pixel given the seed point and 0 for the other pixels, and updates the erosion power by the bacteria. In this paper, seed points automatically placed using depth information [7].

III. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this paper, the pen was detected and extracted as a target object. PASCAL Visual Object Classes 2007 (including plane, bicycle, bird, ship, bottle, bus, car, cat, chair, cow, dining table, dog, horse, motorcycle, person, potted plant, sheep, sofa) [8] was used for the learning dataset. In this paper, pen were detected by using R-FCN, then pen region is extracted by using GrowCut method.

A. Kinect for Windows v2

In this section, we describe the Kinect for Windows v2 (Kinect v2) which is the image measurement device used in this paper. The resolution of the color image of Kinect v2 is 1920 × 1080, and the resolution of the depth image is 512 × 424. In addition, tracking of human skeleton and face detection are possible by using depth information. Time of flight is used as a depth measurement method, and it is possible to measure from 500 to 4500 mm [9].

B. Performance Evaluation

For the evaluation of detection results using R-FCN, average precision rate was used. The Average Precision (AP) was given by the area under the curve of the precision-recall curve obtained from the detection result. The AP in each class was shown in Table 1. The AP was given by the lower area of the Precision-Recall curve obtained from the detection result.In addition, Intersection over Union (IoU) is used as the performance evaluation of target extraction by the GrowCut method. The IoU of the target area is shown in Table 2. IoU is obtained by the equation (1).

$$\mathrm{IoU} = n(\mathrm{A} \cap \mathrm{B})/n(\mathrm{A} \cup \mathrm{B}) \qquad (1)$$

where A is a set of correct region, B is a set of extraction results by the GrowCut method, and n is the number of pixels.

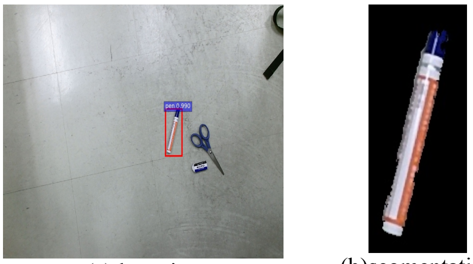
Fig.2 show the object detection results using R-FCN and the segmentation results of target area using the GrowCut method.

Table 1. Average Precision

aeroplane	0.6894	dog	0.8431
bicycle	0.7521	horse	0.8314
bird	0.6616	motorbike	0.7627
boat	0.5277	**pen**	0.7276
bottle	0.4626	person	0.7512
bus	0.7728	potted plant	0.3227
car	0.7856	sheep	0.6986
cat	0.8436	sofa	0.6615
chair	0.4387	train	0.7354
cow	0.7562	tv monitor	0.6143
dining table	0.5841		

Table 2. IoU of the target area

Experiment	IoU	Time[s]
1	0.7887	26
2	0.6354	9.3
3	0.5223	12.8
4	0.5929	14.8
5	0.6306	32.6
6	0.5482	30.6
7	0.5731	6.3
8	0.6429	26
9	0.7473	12.6
10	0.7137	20.3



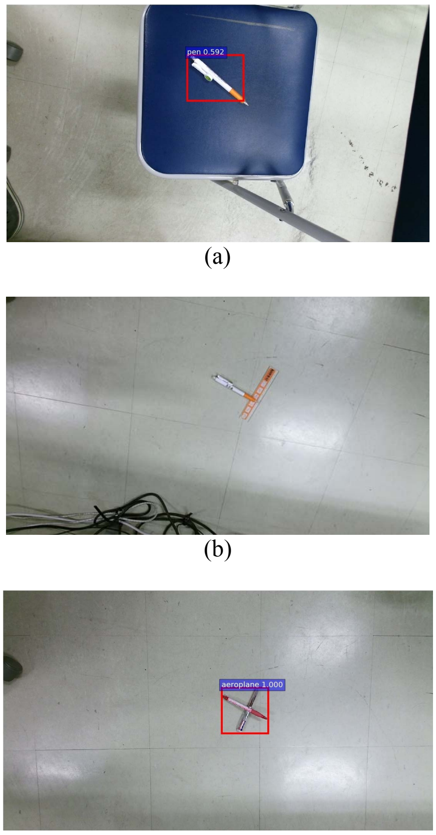
(a)detection (b)segmentation

Fig.2 Experimental result.

C. Discussion

In this paper, object detection and recognition using R-FCN and target objects extraction using GrowCut method were performed.

In object detection and recognition, the APs of each class are as shown in Table 1. The average of AP in each class obtained as 0.6773. Fig.3 shows some examples where detection and recognition were difficult. In Fig.3 (a), the bounding box does not enclose part of the target object. In Fig.3 (b), the target object could not detected correctly. Although object could detected correctly in Fig.3 (c), it is recognized as an airplane. One of these causes is considered to be a problem with learning data, and it is necessary to review the learning data set. In Fig.3 (b) and (c), overlapping objects are considered as the cause of false recognition. Subsequently, regarding segmentation of the target object, IoU is as shown in Table 2. The average value of IoU was 0.6395.



(a)

(b)

(c)

Fig.3 Example of some detection failure.

IV. CONCLUSION

In this paper, we proposed a detection and recognition of target object on a video images and extraction of object region to develop a picking operation in robot based logistics industry. As a flow of the method, detection of a target object using R-FCN, recognition of an object in a detection area, and extraction of an object area using a GrowCut method are performed. Based on our detection of object, 0.6773 of APs obtained. However, some object could not classified correctly. To overcome this problem, new convolution neural network system should be introduced. It remains our future work.

ACKNOWLEDGMENT

This work was supported by Leading Initiative for Excellent Young Researcher of Ministry of Education, Culture, Sports, Science and Technology-Japan (16809746), Grants-in-Aid for Scientific Research of JSPS (17K14694), Research Fund of The Telecommunications Advancement Foundation, Fundamental Research Developing Association for Shipbuilding and Offshore, Japan-China Scientific Cooperation Program (6171101454), International Exchange Program of National Institute of Information and Communications (FY2018), and Collaboration Program of National Institute of Informatics (FY2018).

REFERENCES

[1] National Institute of Population and Social Security Research, Population Projections for Japan : 2016 to 2065, http://www.ipss.go.jp/pp-enkoku/e/zenkoku_e2017/pp29_summary.pdf (2018/5/17)

[2] Ministry of Economy, E-Commerce market survey (Densisyo torihiki ni kansuru sijo tyosa, Japanese), http://www.meti.go.jp/policy/it_policy/statistics/outlook/h28release.pdf (2018/5/17)

[3] Mori et al., Robocon magazine (Japanese), Ohmsha, pp.6-7, (2016)

[4] J.Dai et al., "R-FCN: Object detection via region-based fully convolutional networks", NIPS, pp.379-387, (2016)

[5] R.Girshick et al., "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR, pp.580-587, (2014)

[6] S.Ren et al., "Faster R-CNN: Towards real-time object detection with region proposal networks", NIPS, vol.1, pp.91-99, (2015)

[7] V.Vezhnevets et al., "GrowCut: Interactime multi-label N-D image segmentation by cellular automata", GraphiCon, pp.150-156, (2005)

[8] M.Everingham et al., "The PASCAL Visual Object Classes (VOC) Challenge", IJCV, vol.111, no.1, pp.98-136, (2007)

[9] E. Lachat et al., "First experiences with Kinect v2 sensor for close range 3D modelling", In Proceedings of the Conference on 3D Virtual Reconstruction and Visualization of Complex Architectures, pp.93-100, (2015)