Journal of Robotics, Networking and Artificial Life
Vol. **6**(**1**); June (2019), *pp.* 38–41
DOI: https://doi.org/10.2991/jrnal.k.190531.008; ISSN 2405-9021; eISSN 2352-6386
https://www.atlantis-press.com/journals/jrnal

Research Article

Image Processing for Picking Task of Random Ordered PET Drinking Bottles

Chen Zhu*, Takafumi Matsumaru

Graduate School of Information, Production and Systems, Waseda University, 2-7 Hibikino, Wakamatsu Kitakyushu, Fukuoka 808-0135, Japan

ARTICLE INFO

Article History

Received 09 November 2018
Accepted 19 November 2018

Keywords

Image processing
robotics picking
deep learning
COCO dataset

ABSTRACT

In this research, six brands of soft drinks are decided to be picked up by a robot with a monocular Red Green Blue (RGB) camera. The drinking bottles need to be located and classified with brands before being picked up. The Mask Regional Convolutional Neural Network (R-CNN), a mask generation network improved from Faster R-CNN, is trained with common object in contest datasets to detect and generate the mask on the bottles in the image. The Inception v3 is selected for the brand classification task. Around 200 images are taken or found at first; then, the images are augmented to 1500 images per brands by using random cropping and perspective transform. The result shows that the masked image can be labeled with its brand name with at least 85% accuracy in the experiment.

© 2019 *The Authors.* Published by Atlantis Press SARL.
This is an open access article distributed under the CC BY-NC 4.0 license (http://creativecommons.org/licenses/by-nc/4.0/).

1. INTRODUCTION

Under the lower birth rate and aging society, the cost of human labor is becoming higher. In a warehouse, the picking task for goods sorting takes more than half of the total cost [1]. During the festival and special events, the drinks are randomly put in a big box or a cooler box with water and ice. The existing picking robot can hardly process the overlapping of the objects without modeling or the same objects [2]. In this paper, the image processing for the robot picking task is discussed.

1.1. Related Works

Random picking is a challenging problem in the robotics and computer vision fields. The aim of this task is to pick up objects which are manipulated under structured layout by using a robot arm's end-tip effector. Bin picking was studied when Amazon started the picking challenge. By using a 3D image sensor, the position and pose of the object to be picked up are calculated [3].

On the other hand, for the industrial random picking robots, FANUC, YASKAWA, etc. have developed the bin picking robot by using the structured light or binocular camera.

1.2. Contributions

In some special application such as bottles being put in the ice water, a normal 3D sensor cannot get the correct depth information. In this paper, a deep-learning-based image processing method is purposed to detect and segment the randomly ordered Polyethylene Terephthalate (PET) bottles by using a monocular Red Green Blue (RGB) camera instead of a depth sensor.

Additionally, this research also discusses the brands' recognition under the overlapped conditions by using the Inception v3 [4] without the knowledge on the target.

2. METHODOLOGY

In this research, random piled up drinking bottles of different sizes and brands are required to be picked up. The bottle is not limited to one type of bottle, so deep learning-based detection method are used to solve this problem. The whole process is divided into two stages: the training stage and the detection stage. The training stage is to train the network in order to get the corresponding kernel and bias value. The detection stage is to detect and generate a mask on each bottle and find out the brands of the bottle.

2.1. Network Training Stage

The network training is divided into five steps as shown in Figure 1. First, the Mask Regional Convolutional Neural Network (R-CNN) [5] is pretrained by the Microsoft Common Object in Contest (COCO) dataset. The COCO dataset has a large number of images with labels and segmentation lines. To prevent overfitting, the Mask R-CNN is trained with all 80 classes of COCO dataset. Second, around 200 photos are taken or found for each brand of bottle. Next, the dataset of bottles is used for fine tune the Mask R-CNN. Then, all the images are augmented with random cutting and perspective transform to increase the dataset size to 1500 brands per brands. Finally, the augmented images are used for training for brand recognition.

*Corresponding author. Email: zhuchen@toki.waseda.jp

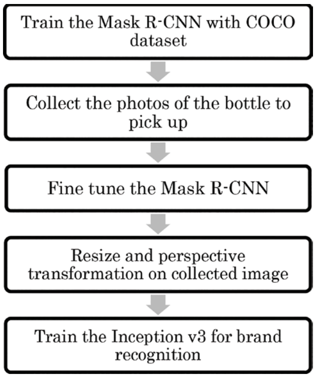
Train the Mask R-CNN with COCO dataset
Collect the photos of the bottle to pick up
Fine tune the Mask R-CNN
Resize and perspective transformation on collected image
Train the Inception v3 for brand recognition

Figure 1 | The network training process.

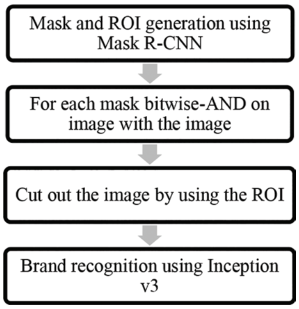
Mask and ROI generation using Mask R-CNN
For each mask bitwise-AND on image with the image
Cut out the image by using the ROI
Brand recognition using Inception v3

Figure 2 | Detection process.

The training of Mask R-CNN takes 160 epochs in total, where 40 epochs for the classification head, and 120 epochs for the ResNet-101 backbone.

Around 80% of the images are randomly selected for the training, and the remaining 20% of images are used for validation. The training is performed with learning rate 0.01 and the training is stopped when the validation accuracy no longer rising along with the training accuracy.

2.2. Detection Stage

The detection stage contains four steps as shown in Figure 2. First, the Region of Interest (ROI) box and the mask are needed to be generated by using Mask R-CNN. Next, the mask is bitwise-AND with the original image. Then, by using the ROI generated in the first step, a bottle is cut out from the image with a black background. Finally, for each image with only one bottle visible are sent to the Inception v3 network for brand recognition.

The output of the Inception v3 is a vector with six elements that indicate the confidence of each class. The evaluation of the brand recognition is based on the comparison between the ground truth, human labeled and the highest confidence of network output, so that it is possible to see the accuracy of brand recognition by using Inception v3 compare to human beings.

Assuming the set of objects with size N is S. The total image of objects detected from the image is $S = S_{\text{human}} \cup S_{\text{inception}}$ and images cannot be classified by Inception v3 or human is S_{failed} and S_e as shown in Figure 3. The accuracy of brand recognition P is calculated by formula (1).

$$\begin{aligned} P_{\text{inception}} &= \frac{N_{\text{inception}}}{N} \times 100\% \\ P_{\text{human}} &= \frac{N_{\text{human}}}{N} \times 100\% \end{aligned} \tag{1}$$

3. RESULTS

Based on the method mentioned in the previous section, the experiment is performed. To run the different network on the same machine, a library called "Protocol Buffer" is used as data exchange.

3.1. Training of Mask R-CNN

The bottles are the primary detection target in this research. However, the number of images that can be used to retrain the whole network is limited. The COCO dataset comes with 80 classes for object detection plus 1 class for background. So, all 81 classes are used for training the whole network at first. Then the bottles taken from the test subject are labeled with a class name and a mask as shown in Figure 4.

The training rate in this step is set to 0.01 and only training the mask and classification parts in the network.

3.2. Training for Brand Recognition

The brand recognition is implemented by the Inception v3. Retraining the whole network will cause too much time and easy to get overfitted. So, the initial weight of Inception v3 is transferred from the object recognition network. In this research, six kinds of

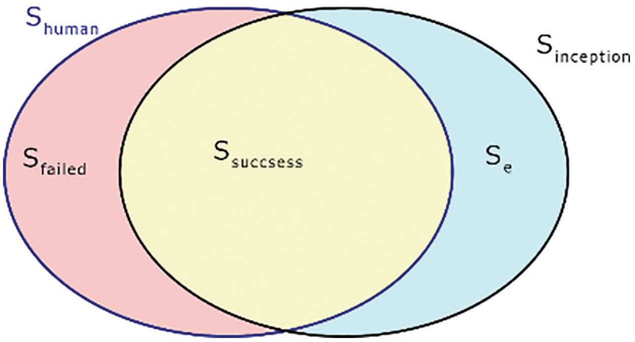

Figure 3 | Brand validation result set.

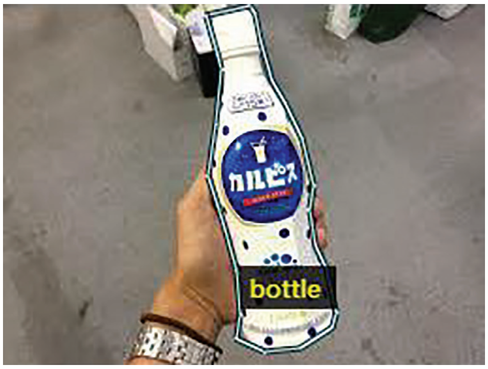

Figure 4 | Image segmentation for training.

drinking in the Japan market including Oiocha, Coca-Cola, Calpis, Afternoon tea, Irohasu, and Nama cha are selected as test subject. For each brand, around 150–200 images are collected from the Internet or taken directly. Then, the images are processed randomly with cropping, perspective transform, rotation and zooming to increase the number of images up to 2000 for each brand as shown in Figure 5.

During the training stage, 20% of all the images in the dataset is selected as the validation dataset. The accuracy of the training is record on the end of each batch. The training of the Inception v3 stops, based on the accuracy that convergence to around 0.85 as shown in Figure 6. The validation accuracy stops increase after around 3500 training steps.

3.3. Evaluation of Mask and ROI Generation

Figure 7 shows the result of the mask and ROI generation. The evaluation is based on the real image taken from a normal monocular camera as shown in Figure 7a. By using the mask and ROI generated from the network shown in Figure 7b, the original image can be masked and cut off as shown in Figures 7c and 7d.

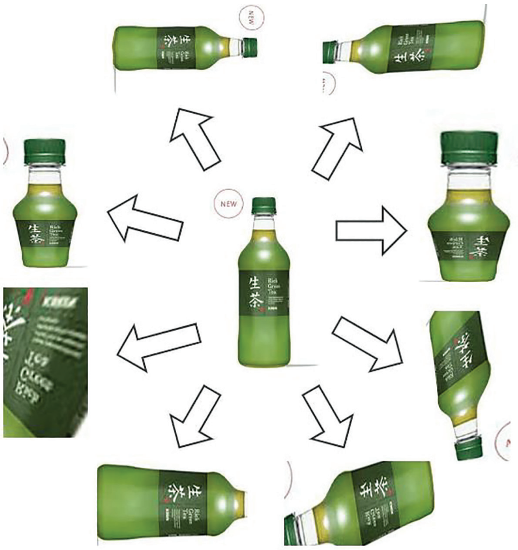

Figure 5 | Image augmentation.

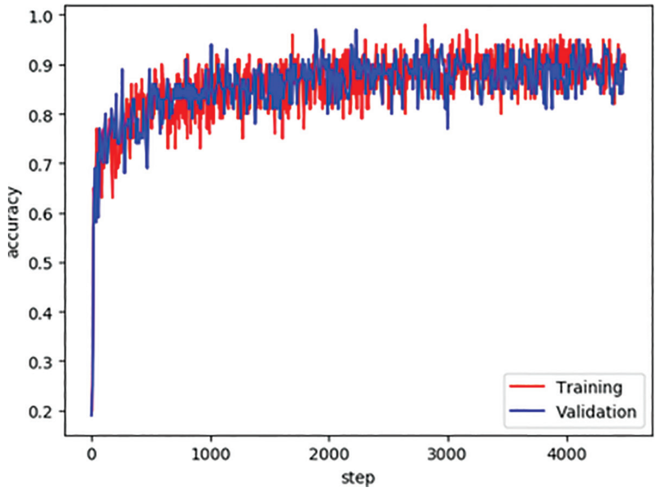

Figure 6 | Training and validation accuracy.

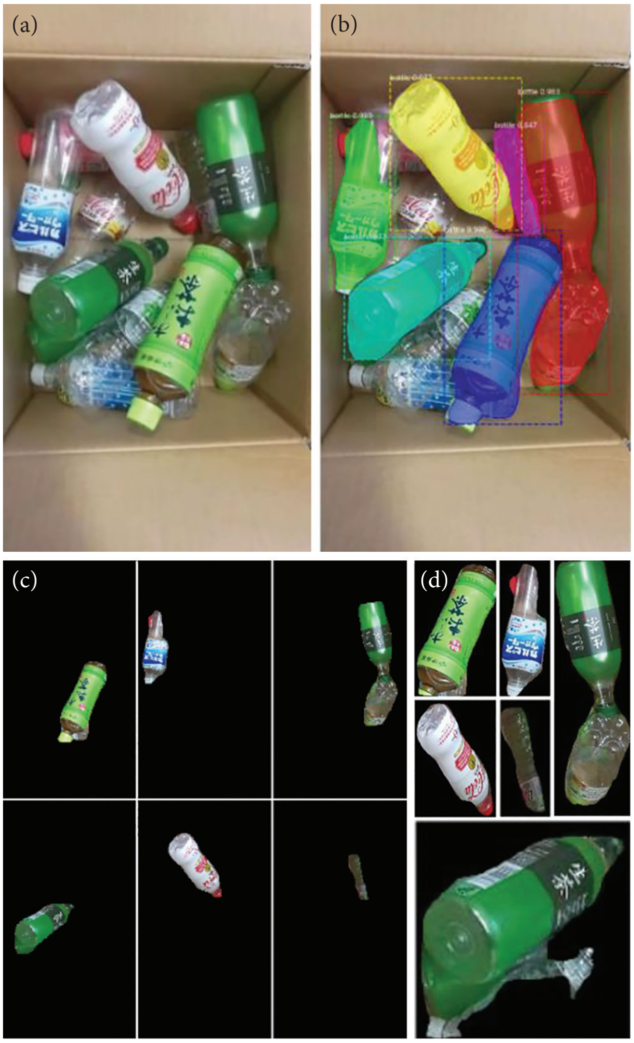

Figure 7 | Mask and ROI cutting on the original image. (a) Original image. (b) Mask generation. (c) Masked image. (d) Image cut out by ROI.

As shown in the result, bottles with color can be correctly detected from the image. However, bottles with a transparent appearance have a lower detection rate.

3.4. Evaluation of the Brand Recognition

The brand recognition is based on the image cut off from the Figure 7d. These images are resized to 299 × 299 and sent to Inception v3 for the brand recognition one-by-one. The output with the highest score is selected as the result. Here, we select one more group of test data besides the images in Figure 7, and the result of the brand recognition is shown in Table 1.

As the result shows, although the network gives out all the correct result, the score of output is not very satisfying in some cases, because the confidence under 0.6 will be unacceptable result.

As the data in Table 1 shows, after filtering out the confidence lower than 0.6, the total accuracy is around 85% which is like human labeled.

Table 1 | Labeled result and output confidence

Ground truth	Machine labeled	Machine confidence	Human labeled
Oiocha	Oiocha	0.995	Oiocha
Calpis	Calpis	0.996	Calpis
Coca-Cola	Coca-Cola	0.977	Coca-Cola
Namacha	Namacha	0.984	Namacha
Namacha	Namacha	0.971	Namacha
Coca-Cola	Coca-Cola	0.900	Coca-Cola
Irohasu	Irohasu	0.994	Irohasu
Afternoon tea	Afternoon tea	0.557	Afternoon tea
Namacha	Namacha	0.986	Namacha
Calpis	Calpis	0.995	Calpis
Namacha	Namacha	0.999	Namacha
Coca-Cola	Coca-Cola	0.882	Coca-Cola
Coca-Cola	Coca-Cola	0.986	Coca-Cola
Irohasu	Irohasu	0.693	Unknown
Coca-Cola	Irohasu	0.989	Unknown
Number of correct answers		13 of 15 result	
Total accuracy		86%	

The Inception v3 can partially treat with the transparent objects in the image.

4. CONCLUSION AND DISCUSSION

The combination of the Mask R-CNN and Inception v3 can detect and recognize the brand of the bottles with overlapping in at least 85% accuracy which gives a near result compare to human beings.

CONFLICTS OF INTEREST

There is no conflicts of interest.

ACKNOWLEDGMENTS

This research is supported by Japan Society for The Promotion of Science (KAKENHI-PROJECT-17K06277), to which we would like to express our sincere gratitude.

REFERENCES

[1] J.J. Bartholdi, S.T. Hackman, Warehouse & Distribution Science, The ISyE department Georgia Institute of Technology, Atlanta, GA, 2014.

[2] K. Kim, J. Kim, S. Kang, J. Kim, J. Lee, Vision-based bin picking system for industrial robotics applications, 2012 9th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), IEEE, Daejeon, South Korea, 2012, pp. 515–516.

[3] N. Correll, K.E. Bekris, D. Berenson, O. Brock, A. Causo, K. Hauser, et al., Analysis and observations from the first amazon picking challenge, IEEE Transactions on Automation Science and Engineering, IEEE, NY, USA, 2018, pp. 172–188.

[4] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA, 2016, pp. 2818–2826.

[5] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, Venice, Italy, 2017, pp. 2980–2988.

Authors Introduction

Mr. Chen Zhu

He is currently studying mechatronics Master course at Graduate School of Information, Production and Systems of Waseda University. He is a member of IEEE. His main research interest is robotics, image processing and deep learning.

Prof. Dr. Takafumi Matsumaru

He received the M.S. and PhD degrees in mechanical engineering from Waseda University, Tokyo, Japan, in 1987 and 1998, respectively. He is a Full Professor with Graduate School of Information, Production and Systems, Waseda University, where he is in charge of courses in robotics, mechatronics and bioengineering. His research interests include human-robot interaction.