Image Processing for Picking Task of Random Ordered PET Drinking Bottles

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

Takafumi Matsumaru
Graduate School of Information, Production and Systems, Waseda University, 2-7 Hibikino Wakamatsu Kitakyushu 808-0135 Fukuoka, Japan
E-mail: zhuchen@toki.waseda.jp, matsumaru@waseda.jp
www.waseda.jp

Abstract

In this research, six brands of soft drinks are decided to be picked up by a robot with a monocular RGB camera. The drinking bottles need to be located and classified with brands before being picked up. A Mask R-CNN is pretrained with COCO 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, then, the images are augmented to 1500 images per brand by using random cropping and perspective transform. The results show that the masked image can be labeled with its brand name with at least 85% accuracy in the experiment.

Keywords: Image Processing; Robotics Picking; Deep Learning; COCO Dataset

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 PET bottles by using a monocular RGB camera instead of a depth sensor.

Additionally, this research also discusses the brands' recognition under the overlapped conditions by using the Inception v3 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 Fig. 1. First, the Mask R-CNN (Regional Convolutional Neural Network) (4) is pretrained by the Microsoft COCO (Common Object in Contest) 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.

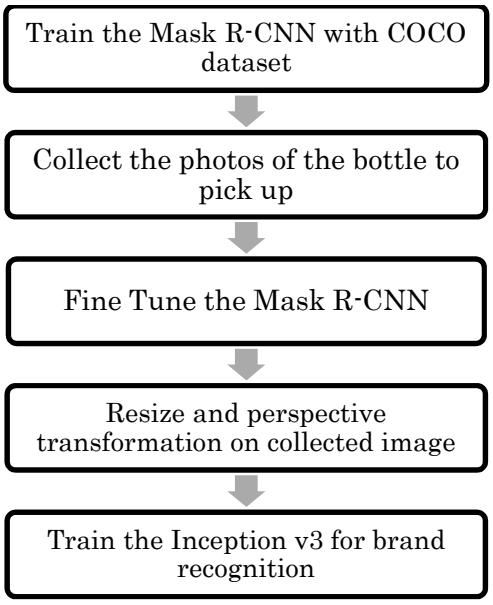

Fig. 1. The network training process

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 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 Fig. 2. First, the ROI (Region of Interest) 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.

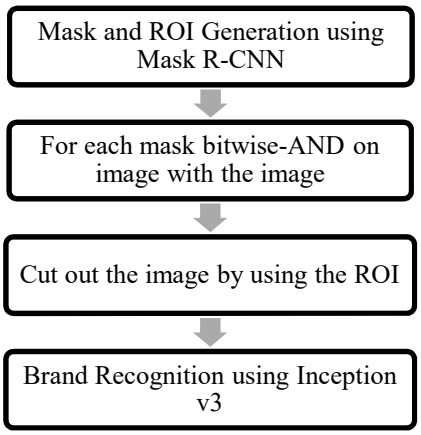

Fig. 2. Detection Process

The evaluation of the brand recognition is based on the comparison between the human and the network output, so that it can filter out the overlapped bottles that cannot be recognized by a human. Assume the number of objects detected from the image is N_d; the number of correct brands recalled by the Inception v3 is N_r; the number of which brands cannot be recognized by a human is $\overline{\mathrm{N_h}}$ in which correctly recalled by Inception v3 is $\overline{N_e}$. Then the accuracy P of the brand recognition is calculated by the following formula.

$$\mathrm{P} = \frac{N_r - \overline{N_e}}{N_d - \overline{N_h}} \times 100\% \qquad (1)$$

3. Result

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 Fig. 3.

Fig. 3. Image Segmentation for Training

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 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 1500 for each brand as shown in Fig. 4.

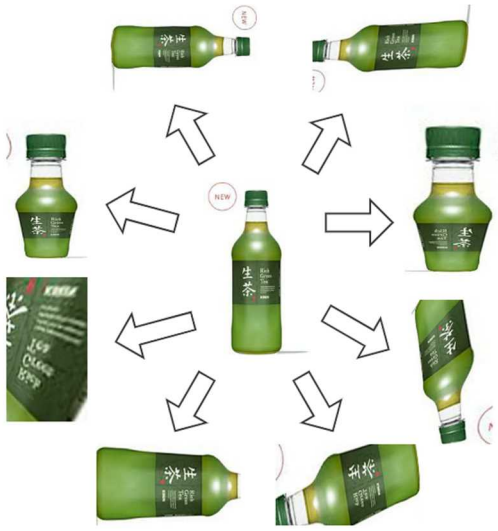

Fig. 4. Image Augmentation

The training of the Inception v3 stops, based on the accuracy that convergence to around 0.85 as shown in Fig. 5. The blue line shows the validation result of the 20% of the images which can be used to indicate the best fitting point.

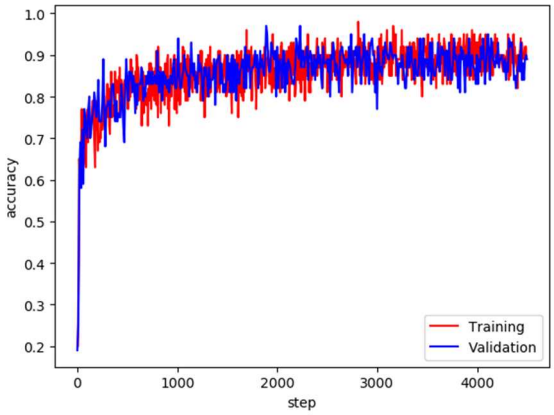

Fig. 5. Training accuracy and validation accuracy

Continuing the training steps will increase the training accuracy but not validation accuracy, which will lead to overfitting of the network. In this dataset, validation accuracy reaches the top when the training steps are 4000 steps.

3.3. *Evaluation of Mask and ROI Generation*

Fig. 6 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 Fig. 6(a). By

using the mask and ROI generated from the network shown in Fig. 6(b), the original image can be masked and cut off as shown in Fig. 6(c) and (d).

(a) Original Image (b) Mask Generation

(c) Masked Image (d) Image Cut out by ROI

Fig. 6. Mask and ROI cutting on the original image

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 Fig 6(d). These images are resized to 299x299 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 "Fig. 6. Mask and ROI cutting on the original image", 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 score under 0.6 will be seen as unacceptable result.

Table 1 Labeled result and output score

Human Labeled	Top Result	Score
Oiocha	Oiocha	0.66
Calpis	Calpis	0.94
Namacha	Namacha	0.93
Namacha	Namacha	0.56
Coca cola	Coca Cola	0.87
Irohasu	Irohasu	0.98
Afternoon Tea	Coca Cola	0.62
Namacha	Namacha	0.99
Calpis	Calpis	0.92
Namacha	Namacha	0.97
Coca Cola	Coca Cola	0.96
Coca Cola	Coca Cola	0.29
Irohasu	Irohasu	0.91
Number of correct answers		**10**

4. Conclusion & 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 80% accuracy. Based on the mask center point and the area of the mask, the robot can be guide for the picking task.

Acknowledgements

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. JJ. Bartholdi and ST. Hackman, *Warehouse & Distribution Science* (The ISyE department Georgia Institute of Technology, Atlanta, GA, 2014).
2. K. Kim, J. Kim, S. Kang, J. Kim and J. Lee, Vision-based bin picking system for industrial robotics applications, *9th International Conference on Ubiquitous Robots and Ambient Intelligence*. (Daejeon, South Korea, 2012).
3. N. Correll and K.E. Bekris et al, Analysis and Observations From the First Amazon Picking Challenge, *IEEE Transactions on Automation Science and Engineering*. **15**(1) (2016) 172-188.
4. K. He, G. Gkioxari, P. Dollár and R. Girshick, Mask R-CNN, *2017 IEEE International Conference on Computer Vision (ICCV)*, (Venice, Italy).
5. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, Rethinking the Inception Architecture for Computer Vision, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* (Las Vegas, NV, USA, 2016)