A Quick Data Generation Method for Training Object Detection Algorithms in Home Environments

Yuma Yoshimoto*†, Muhammad Farhan Mustafa‡, Wan Zuha Wan Hasan‡, and Hakaru Tamukoh†
*JSPS Research Fellow, Japan
†Kyushu Institute of Technology, Japan
‡Universiti Putra Malaysia, Malaysia

***Abstract*—Deep neural networks are the mainstream of object detection algorithms. The required data are scene images and annotation data for training. Here we propose a new model for training object detection algorithms. Data in our method are quickly generated by a four-step procedure: (1) videos acquisition of objects, (2) saving of video frames as scene images, (3) generation of annotation data from the detection results of You Only Look Once 9000 (YOLOv2), which inputs scene images, and (4) data augmentation. In a comparison experiment, the proposed method generated data 10 times faster than conventional methods. We then trained YOLOv2 on the data generated by the proposed method, and evaluated the effectiveness of the proposed method. The training increased the Intersection-over-Unin measure of YOLOv2, confirming the effectiveness of training by the proposed method.**

I. Introduction

Recently, service robots that support the daily-life activities of people in their home environments are attracting much attention. Service robots require algorithms that detect objects from scene images. Mainstream object detection methods are based on convolutional neural networks (CNNs) [1]. CNNs are trained on the data of scene images and their corresponding annotation data. Among the many published datasets for object detection are Common Objects in Context Dataset (COCO Dataset), and the PASCAL Visual Object Classes Dataset [2] [3]. When the object data are unavailable, datasets for recognizing the target objects must be prepared. For instance, service robots must recognize the objects in their individual home environments, Dataset generation conventionally proceeds by (1) taking images of the objects and (2) manually constructing the annotation data of the images. This method requires a long time and the efforts of many persons. Moreoever, reparing human and temporal resources in the home environment is a difficult task. Therefore, fast data generation methods are required.

This paper proposes a quick data generation method that produces scene images from object videos. The annotation data are produced from the scene images by an object detection algorithm pre-trained on another dataset.

II. Related Works

A. Object Detection Algorithms

Object detection algorithmes provide the object names and coordinates of scene images. A detection result is shown in Fig. 1. Objects in the image are delineated with bounding boxes (BBs) labeled with text. The BB data are the object coordinates, and the text data are the object names. CNN has become the mainstream of object detection methods. For training a CNN, the dataset must contain the following information:

- A scene image including the objects to be recognized
- The corresponding annotation data (names and BB coordinates of the objects).

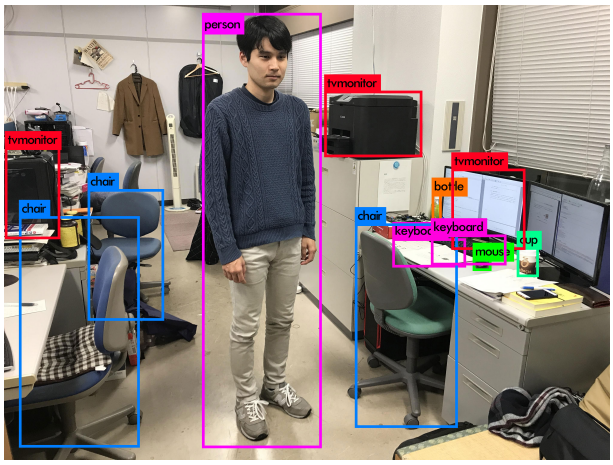

Fig. 1. Object Detection

B. You Only Look Once 9000

You Only Look Once 9000 (YOLOv2) is a CNN model for object detection proposed by Joseph Redmon *et al.* [4]. YOLOv2 achieves high-accuracy detection (76.8 mAP when tested on the VOC 2007 dataset) at high speed. YOLOv2 performs faster than Single Shot MultiBox Detector, another high-speed object detection algorithm that processes 512×512 images at 19 fps [5] (in contrast, YOLOv2 can process 544×544 images at 40 fps).

III. Proposed Method

This paper proposes a quick data generation method for training CNN models. As mentioned above, training CNN models for object detection requires scene images and annotation data. The steps of the proposed method, namely, scene-image generation, annotation data generation, and data

TABLE I
EXPERIMENTAL SETUP

CPU	Intel(R) Core i7-8750H
Memory	32GB
GPU	NVIDIA GeForce GTX 1080 8GB
OS	Ubuntu16.04
Language	Python 2.7.6

augmentation, are detailed in the following subsections A, B and C respectively.

A. Scene Images Generation Step

Scene images are generated as follows.
(1) The objects to be detected (target objects) are video-recorded under the following conditions.

- The background is a solid color.
- Every frame includes only one target object.
- The height and width of the target object are half the height and width of the scene image.

(2) Every video frame is saved as an image file.

B. Annotation Data Generation Step

Annotation data include the object name and BB data of every scene image. The annotation data are generated as follows:
(1) The pre-trained YOLOv2 loads the scene images, detects the target objects, and outputs the BB data.
(2) The name and BB data of the target object are saved as the annotation data of that object.

C. Data Augmentation Step

Finally, the data are augmented as follows.
(1) Each scene image is processed through a series of filters (see Figure 2).

- 4 change contrast filter
- Blur filter
- Histogram equalization filter
- Add Gaussian noise
- Add salt and pepper noise
- 5 color filters

(2) Fourteen reverse images are generated from the original image and saved. A further 13 images are generated by stage (1), giving 27 images from a single original image.
(3) The annotation data of the 27 images are generated and saved.

IV. EXPERIMENT

The data generation speeds of the proposed method and two conventional methods were compared in an experiment using YOLOv2. The effectiveness of the data generated by the proposed method was also evaluated. For this purpose, we trained YOLOv2 on the data, and evaluated the accuracy of YOLOv2.

The experiment environment is described in Table I.

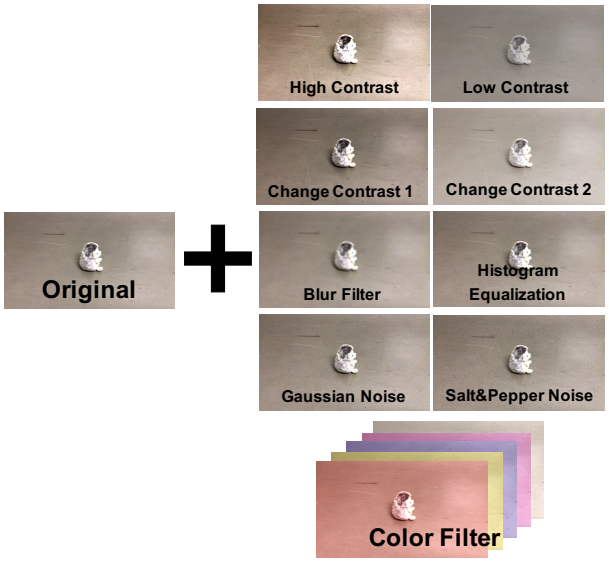

Fig. 2. Data Augmentation Methods

TABLE II
SUMMARY OF PROPOSED METHOD AND CONVENTIONAL METHODS

	Proposed Method	Conventional Method 1	Conventional Method 2
Images preparing methods	Follow III-A	Taken by person	Taken by person
Number of objects in each image	1	3	1
Generating methods of annotation data	Follow III-B	Generated by person	Generated by person
Augmentation methods	Follow III-C	Follow III-C	Follow III-C

A. Experiment 1

1) Experimental Method: In this experiment, we compared the data generation speeds of the proposed and conventional methods. Figure 3 shows the three objects investigated in the experiment, and Table II describes the data generation methods.

The data generation speeds of the proposed and conventional methods were compared for one object in one image. The speeds were calculated from the data generation time (t), the number of generated data (n), and the number of objects in one data (o) as follows:

$$speed = \frac{t}{n \times o} \quad (1)$$

Conventional Method 1 acquires the objects as photographs (see Fig. 4 (a)). Each image contains three objects in this method, but only one object in Conventional Method 2 (Fig. 4 (b)). Furthermore, in both Conventional Method 1 and Conventional Method 2, the annotation data are manually generated by a person. In all methods, data are augmented by the algorithm described in III-C.

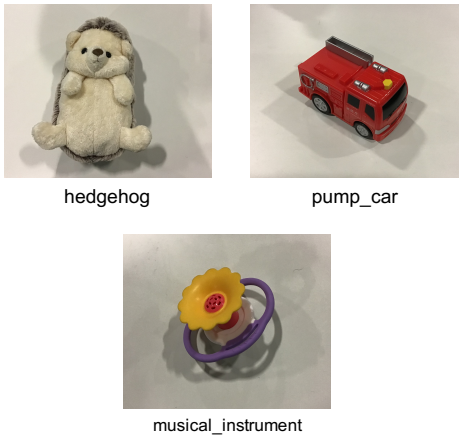

Fig. 3. Objects

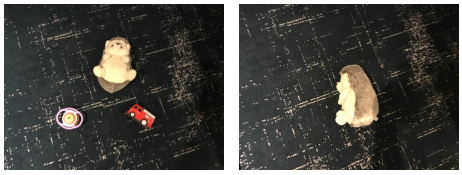

Fig. 4. Image data generated by Conventional Method 1, 2

2) Experimental Results: Table III shows the experimental results. The proposed method was 10 times faster than conventional methods 1 and 2.

B. Experiment 2

1) Experimental Method: This experiment checks the effectiveness of the data generated by the proposed method. YOLOv2 was trained on the data generated by the proposed method in experiment 1, and the detection accuracy was determined. The experimental method proceeded as follows.
(1) Initialize YOLOv2 using the weight file from darknetl9, which is trained on the COCO dataset.
(2) Fine-tune YOLOv2 on the data generated by the proposed method.
(3) Evaluate the accuracy by the Intersection-over-Union (IoU) measure.

2) Experimental Result: Figure 5 shows the average IoUs of the objects. These results confirm the training effectiveness of the data generated by the proposed method.

TABLE III
SPEED COMPARISONS OF THE PROPOSED AND CONVENTIONAL METHODS

Methods	Proposed Method	Conventional Method 1	Conventional Method 2
Data Generation Time (t)	13 min 2 sec	10 min 25 sec	8 min 34 sec
Number of Data (n)	36428	560	1680
Number of Objects in One Data (o)	1	3	1
Speed	0.021 sec	0.372 sec	0.306 sec

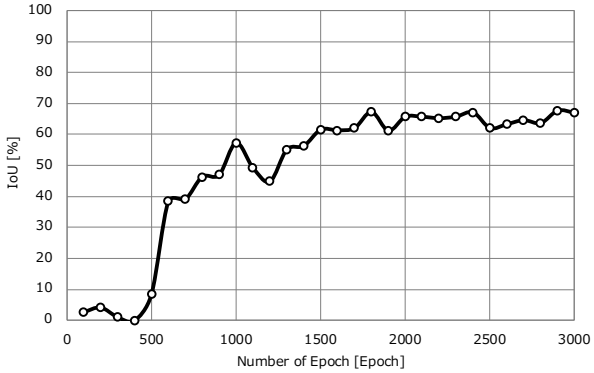

Fig. 5. IoU of training in the proposed method

TABLE IV
ACCURACY COMPARISON OF THE DATA GENERATED BY EACH METHOD AFTER 5 MINUTES

	hedgehog	pump_car	musical_instrument	average
Proposed Method	0.639	0.719	0.609	0.655
Method 1	0.364	0.358	0.737	0.486
Method 2	0.281	0.493	0.065	0.288

C. Experiment 3

1) Experimental Method: Finally, we compared the recognition accuracy of the object recognition algorithm trained by the three methods. The object recognition algorithm was trained three times, once each on the datasets generated by the three dataset generation methods. Each training was performed on the training data prepared in five minutes by the evaluated method. Finally, the numbers of training data were calculated from the "Speed" values in Table III as follows:

$$number\_of\_data[images] = \frac{300sec}{Speed} \quad (2)$$

The training was performed as described in experiment 2. The training number was 2000 epochs.

2) Experimental Result: The accuracies of the trained algorithms were evaluated on the three objects shown in Fig. 3. The experimental results, listed in Table IV, confirm that the proposed method achieved the highest accuracy among the three methods.

V. CONCLUSIONS

This paper proposed a quick data generation method for training object detection algorithms in home environments. The method generates the dataset as follows.

- The target objects are video-recorded.
- Each frame in the videos is saved as an image.
- The objects in the image are detected by the pre-trained YOLOv2, which outputs the BB data. The object names and BB data are saved as the annotation data.
- The datasets are augmented.

The experimental results confirmed that: (1) data generation by the proposed method is much faster than manual data

generation by a person; (2) the method is effective for training object detection algorithms; (3) the method achieves higher accuracy than other methods on training data generated within the same training time.

In future work, the algorithm trained by the proposed method will be implemented in service robots.

ACKNOWLEDGMENT

This research was supported by JSPS KAKENHI Grant Numbers 19J11593 and 17H01798.

REFERENCES

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," Proceedings of the IEEE, vol.86, no.11, pp.2278–2324, 1998.

[2] T. Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick,and P. Dollár, "Microsoft COCO: Common Objects in Context," arXiv:1405.0312, 2014.

[3] M. Everingham, S. M. A. Eslami, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge: A Retrospective," International Journal of Computer Vision, vol.111, num.1, pp.98-136, January 2015.

[4] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," arXiv:1612.08242, 2016.

[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector," arXiv:1512.02325, 2015.