Robot Eye: Automatic Object Detection And Recognition Using Deep Attention Network to Assist Blind People

Ervin Yohannes
Department of Computer Science & Information Engineering
National Central University
Taoyuan, Taiwan
ervinyohannes@gmail.com

Paul Lin
Department of Computer Science Information Engineering
National Central University
Taoyuan, Taiwan
paul92035@gmail.com

Chih-Yang Lin*
Department of Electrical & Engineering
Yuan-Ze University
Taoyuan, Taiwan
andrewlin11@saturn.yzu.edu.tw

Timothy K. Shih
Department of Computer Science & Information Engineering
National Central University
Taoyuan, Taiwan
tshih@g.ncu.edu.tw

***Abstract*—Detection and Recognition is a well-known topic in computer vision that still faces many unresolved issues. One of the main contributions of this research is a method to guide blind people around an outdoor environment with the assistance of a ZED stereo camera, a camera that can calculate depth information. In this paper, we propose a deep attention network to automatically detect and recognize objects. The objects are not only limited to general people or cars, but include convenience stores and traffic lights as well, in order to help blind people cross a road and make purchases in a store. Since public datasets are limited, we also create a novel dataset with images captured by the ZED stereo camera and collected from Google Street View. When testing with images of different resolutions, our method achieves an accuracy rate of about 81%, which is better than naive YOLO v3.**

Keywords—navigation tools, deep learning, attention, detection, recognition

I. INTRODUCTION

According to the World Health Organization (WHO), there are about 285 million visually impaired people worldwide, 39 million of which are blind. As of 2020, those numbers are continuing rising [1]. In daily life, people with severe vision impairment commonly use a tactile stick to check their surrounding environment in outdoor situations. Unfortunately, this approach has some drawbacks, since users can't acknowledge objects that are far away. Researchers have recently developed useful tools for blind people, such as guide dog robots or navigation systems. However, those solutions require several days to learn the behaviors of each specific blind user and need training in specific situations [2]. In contrast, blind people are able to conduct indoor activities using intelligent assistive navigation that leverages a mobile device to help blind people recognize indoor objects. The costs for such services are lower than for outdoor activities [3]. Cloud-based technology also exists that determines the location of the blind person and sends the location information to use as a navigational system [4]. An RGB-D camera can be used to determine the state of an outdoor environment as well, for instance, to look for empty seats for blind people [5].

Recently, deep learning has become more popular than traditional methods for solving issues in daily life, especially for blind people. The development of deep learning systems can achieve high accuracy rates at low costs. Convolutional Neural Network (CNN) is one deep learning method that can be used in many vision tasks, such as helping blind people walk outside [6]. Many CNN-based methods, like Single Shot Detector (SSD) and You Only Look Once (YOLO), are used to solve detection and recognition issues [7]. Both of these are one-stage detection architectures that predict the coordinates and the classes of the objects at the same time for faster processing, and they have been implemented in many real-time systems compared with two-stage detection architectures like Faster R-CNN and Mask R-CNN [8]. Many versions of YOLO have been developed since it is easy to adopt and combine with other methods [9]. YOLO is suitable for developing systems for blind people, not only for its accuracy, but for its performance as well. In order to improve CNN-based methods for large-scale tasks, the attention mechanism has been extensively studied [10].

Object detection is an important task in computer vision that needs to be solved [15]. Although many methods have been proposed in existing research, the accuracy has not improved much [16]. In this paper, we propose automatic object detection and recognition to assist blind people in going outside. We use DarkNet-53 as the backbone, just like YOLOv3 does, but we add an attention mechanism into it. The objects for recognition consist of people, various kinds of vehicles, four types of convenience stores and traffic lights. Mean Average Precision (mAP) was used to compare the proposed method with naive YOLOv3. The contributions of our research include:

1) A proposal for automatic object detection and recognition to assist blind people

DOI 10.1109/ICPAI51961.2020.00036

2) A novel method using DarkNet-53 as a backbone with attention mechanism

3) A new dataset with an emphasis on convenience stores and traffic lights as objects

The rest of this paper is organized as follows. Section II discusses related literature, Section III details our proposed method, experimental results are discussed in Section IV, and conclusions are drawn in Section V.

II. Related Work

Navigation robots have been increasingly developed with the aim to help blind people in daily activities, since traditional navigation involves an unfamiliar environment and incurs high service costs. Joao et al. [11] created a simple robot called CaBot that helps blind people walk around inside environments, and is the same size as a suitcase that can be carried everywhere without hassle. However, in outside environments, the blind still need other people to help them with activities such as crossing the street, stopping a bus, or asking for directions to a store. Qiang et al. [12] developed a guide robot that is able to detect traffic lights and moving objects in the outside environment. The system is created using SSD and can only recognize six classes of objects. Their results show that Faster R-CNN attains higher accuracy than SSD, but that the computation time of SSD is better than that of Faster R-CNN.

Josh et al. [13] proposed a deep learning approach for a navigation system that guides blind people by using YOLO architecture to detect surrounding objects such as chairs and tables. Their system additionally provides the distance between blind people and objects. Wenbo et al. [14] have also used YOLO architecture to implement a pedestrian detection system. Their results show that the miss rate of YOLO is better than that of other methods, such as VJ, HOG, HikSvm, HogLbp, etc.

As proposed by XiaWen et al., YOLO can further be combined with color identification [17]. Their research successfully implemented YOLO for traffic light detection and reached a recognition rate as high as 100%. Many one-stage detectors like YOLO and SSD are better than two-stage detectors in terms of high computation speed; however, their accuracy does not exceed that of two-stage detectors like Fast R-CNN, Faster R-CNN and Mask R-CNN [18]. Currently, CNN-based methods are applied in complex and deep structures in order to gain wider receptive fields and higher accuracy. However, architectures with too many CNN layers require a lot of computations and may not lead to better results. To solve this, Convolutional Block Attention Module (CBAM) can be used to get the global information that is helpful in CNN. Furthermore, CBAM has only a few trainable parameters in it and can focus on the main feature representations of CNN during the training process [19][20].

III. Proposed Method

The proposed method includes general architecture and attention network.

A. General Architecture

The general architecture includes residual blocks, skipping layers, addition layer, detection layer, and attention layer as shown in Figure 1. DarkNet-53 is used as the backbone, just like in YOLOv3. In DarkNet-53, each residual block contains CNN layers, Batch Normalization, and a Leaky-ReLU activation function. In the 36th and 61st layer, skipping layers were used to prevent the vanishing gradient problem in deep learning. The addition layer is used by non-local blocks to get the global information for the whole image. There are also three detection layers to predict the objects in multi-scales.

B. Attention Network

The attention network includes max-pooling, average-pooling and shared Multi-Layer Perceptron (MLP). The architecture of the attention layer is shown in Figure 2. The input feature (F) processes max-pooling and the average-pooling layer; both features are used in a shared multi-layer perceptron to learn the representations of the objects. After adding these two outputs from MLP, a sigmoid function representing the importance of each channel from the input feature (F) is executed. The function can be multiplied with the input feature (F) to enhance or restrain itself channel-wise.

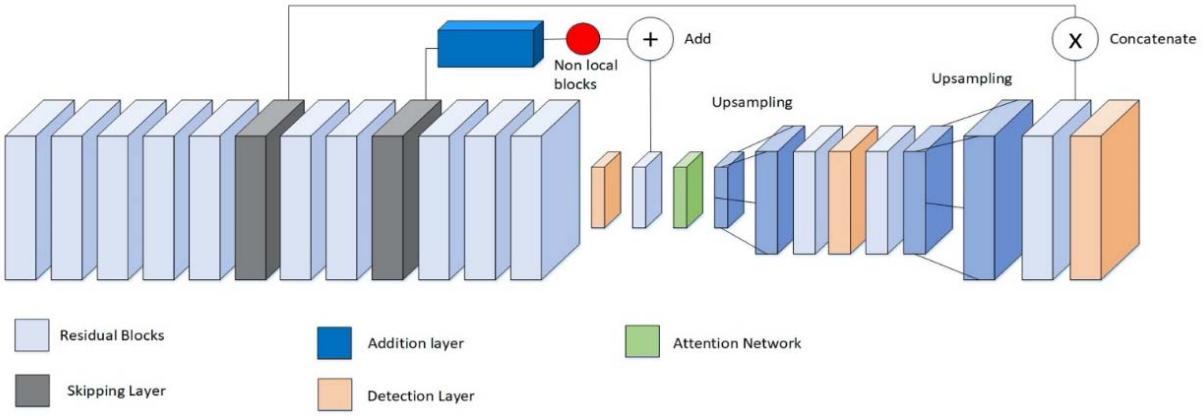


Fig. 1. General architecture containing residual blocks, skipping layer, addition layer, detection layer, and attention network.

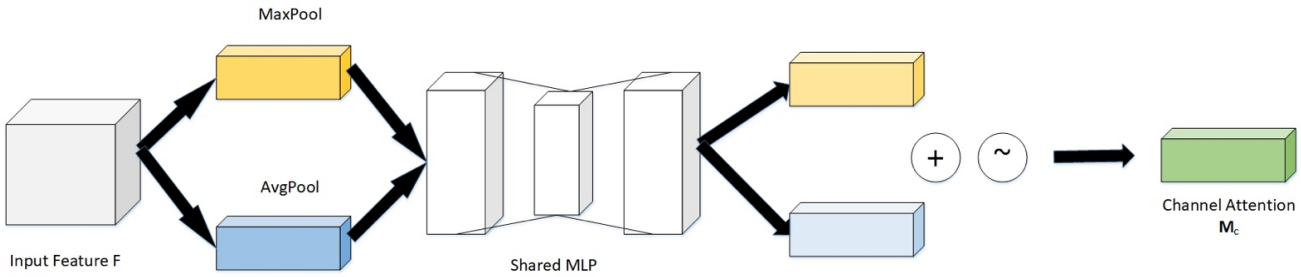


Fig. 2. Attention layer focusing on important features channel-wise.

IV. Experimental Results

There are three sections in the experimental outcomes, including datasets, training strategies, results between naïve YOLO version 3 & our proposed method, and detection & recognition results.

A. Datasets

The datasets are divided into three categories consisting of signboards, traffic lights, and public datasets. Signboard datasets are datasets with convenience stores in Taiwan, including 7-11, OkMart, FamilyMart, and Hi-Life. Our research team compiled those datasets through Google Street View in Taiwan. Second, traffic light datasets contain four classes of lights, including red traffic lights for pedestrians, green traffic lights for pedestrians, red traffic lights for vehicles and green traffic lights for vehicles. Images were taken using a ZED Stereo Camera in Taiwan streets. Third, we chose the MS COCO dataset, since it contains many images with people, cars, trucks, buses, motorcycles and bicycles. The datasets include 18,154 images for training and 2,629 images for validation. Another public datasets are PASCAL VOC datasets for testing between naïve YOLO version 3 and our proposed method. Details of the datasets are shown in Figure 3.

(a)

(b)

(c)

Fig. 3. Samples from datasets (a) Traffic light in a Taiwan street taken by our team using a ZED stereo camera (b) General object (c) Convenience store in a Taiwan street taken by Google Street View

In Figure 3, some images contain other objects (e.g. signboard datasets have both traffic lights and people), so that the model can more easily detect and recognize various objects in the image.

B. Training Strategies

The datasets should have PASCAL VOC annotation since we just using PASCAL VOC format for training. For both training, our proposed method and YOLO version 3 was implemented on the Keras and Tensorflow platform using Adam optimizer with parameter such as learning rate = 0.0001, number of epochs = 200, and batch size = 10. The machine that we used by specification of Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz, 32GB of RAM, 24GB of GPU NVIDIA GeForce GTX 1080.

C. Results between Naïve YOLO Version 3 and Our Proposed Method

We compare the results of our method with those of naive YOLOv3 in public datasets and our datasets. The results of public datasets using naïve YOLO version 3 and our proposed method are presented in Table I.

TABLE I. MEAN AVERAGE PRECISION RESULTS BETWEEN NAÏVE YOLOV3 AND OUR PROPOSED METHOD IN PASCAL VOC DATASETS

No	Results		
	Categories	***YOLOv3***	***Our proposed***
1	Person	71.76	72.33
2	Car	78.06	78.56
3	Motorcycle	80.22	80.73
4	Bicycle	75.23	75.73
5	Monitor	79.32	79.83
6	Bus	87.41	87.96
7	Aeroplane	76.34	76.85
8	Boat	81.06	81.59
9	Train	75.32	75.83
10	Bird	73.95	74.45
11	Cat	84.7	85.35
12	Cow	72.86	73.44
13	Horse	84.19	85.23
14	Sheep	88.63	89.12
15	Dog	84.75	85.36
16	Bottle	77.66	78.17
17	Chair	85.84	86.45
18	Dining table	71.89	81.88
19	Potted plant	85.63	86.12
20	Sofa	85.49	85.91
mAP		80.0155	**81.0445**

TABLE II. MEAN AVERAGE PRECISION RESULTS BETWEEN NAÏVE YOLOV3 AND OUR PROPOSED METHOD IN OUR DATASETS

No	Results		
	Categories	***YOLOv3***	***Our proposed***
1	Person	69.96	70.42
2	Car	72.40	72.66
3	Motorcycle	75.11	75.77
4	Bicycle	49.09	50.64
5	Truck	47.54	48.32
6	Bus	66.35	66.88
7	7-11	96.33	97.04
8	FamilyMart	96.67	97.46
9	OkMart	98.87	98.88
10	Hi-Life	97.56	98.05
11	v_red	93.14	93.50
12	v_green	85.36	85.48
13	p_red	88	88.63
14	p_green	92.73	92.73
mAP		80.65	**81.18**

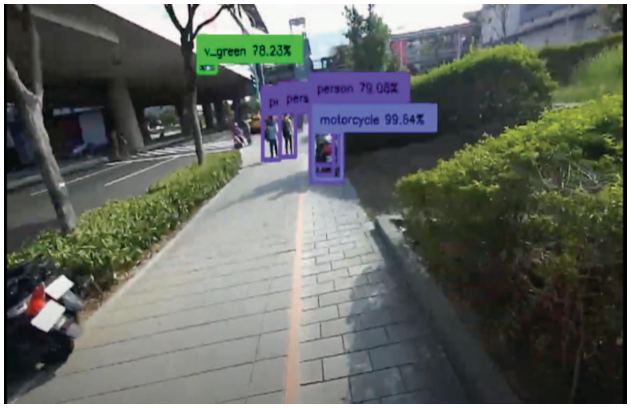
Based on the Table I, the mAP improvement about 1.02 from naïve YOLO version 3 method. We can see almost all of the classes has improvement in our proposed than naïve YOLO version 3 method. We just using PASCAL VOC dataset in the public dataset separate with MS COCO datasets for testing and we take all of the class PASCAL VOC. Meanwhile, MS COCO dataset used by mix our datasets. The results of our datasets using naïve YOLO version 3 and our proposed method can be shown in Table II.

Based on Table II, the addition of the attention mechanism achieves better results across all categories, with the exception of green traffic lights for pedestrians (p_green), which has an equivalent AP. The mAP improves by 0.53 as compared with the naïve YOLO version 3 method. We believe that addition of the attention mechanism can improvement the naïve YOLO version 3.

D. Detection and Recognition Results in our proposed method

We test our method on Taiwan streets with an image that includes many objects, most of them match our expected categories. We have two image results consist of phone camera and ZED camera. We didn't show the results of YOLO version 3 since the results is similar to our proposed method but there is different about confidence score of the results between our proposed and YOLO version 3 method. The detailed detection and recognition results are shown in Figure 4.

(a)

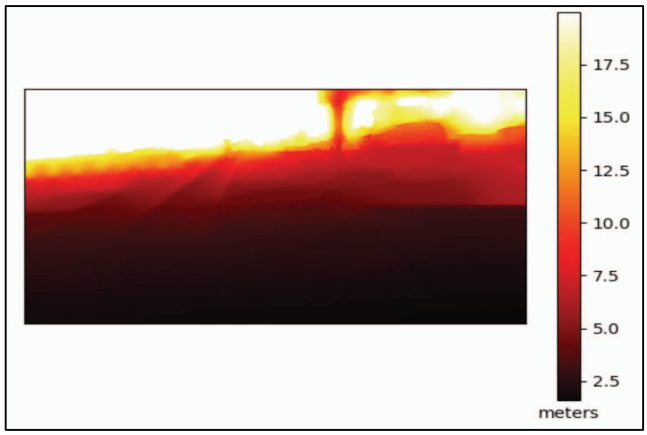


(b)

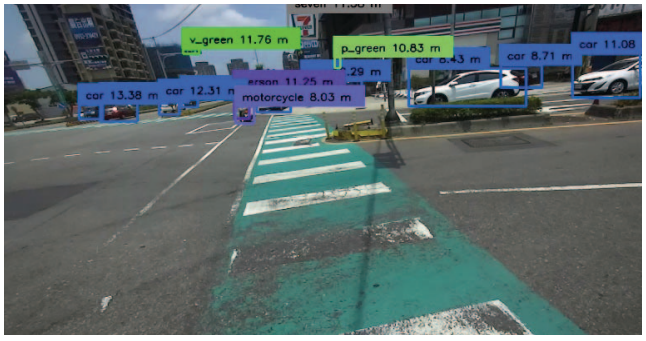
Fig. 4. Detection and recognition results (a) using a phone camera (b) using the ZED Stereo Camera.

Based on Figure 4, our proposed method can be applied to a camera with high and low resolution, which can precisely detect and recognize all of the objects. We assign different colors to each object, such as red and green to a traffic light, gray for a convenience store, and the rest for general objects. Our results show that small objects can be accurately detected and recognized (e.g. traffic lights) since we use an attention network to extract important features.

(a)



(b)



(c)

Fig. 5. Results of prediction with ZED Stereo Camera (a) RGB image of ZED Stereo Camera (b) Depth image of ZED Stereo Camera (c) Prediction with depth information

Figure 5 shows the predictions of the image produced by proposed method. The results contain the position and the depth of the objects, which can be used to notify the blind people where and how far away the objects are. With the Text-to-Speech(TTS) mechanism, the blind people know that he can walk across the road since it's green light and there is a seven convenience store over the street in this example.

V. Conclusion

Our method achieves a higher accuracy than YOLOv3 based on results with datasets compiled by our team to address the lack of published public datasets that include convenience stores and various kinds of traffic lights. Since the attention mechanism in our method can detect objects exactly, even when objects (e.g. traffic lights) are small and far away, our robot-eye system can effectively assist blind people in navigating outdoor environments.

Acknowledgment

We thank the Pervasive Artificial Intelligence Research (PAIR) Labs for their support. The Consortium is funded by the Ministry of Science and Technology.

REFERENCES

[1] P.Kumar, M.Valentina, E.Balas, A.Kumar Bhoi, andA. F.Zobaa, *Advances in Intelligent Systems and Computing 768 Cognitive Informatics and Soft Computing Proceeding of CISC 2017*. 2017.

[2] J.Zhu *et al.*, "An Edge Computing Platform of Guide-dog Robot for Visually Impaired," pp. 1–7, 2020.

[3] B.Li *et al.*, "Vision-Based Mobile Indoor Assistive Navigation Aid for Blind People," *IEEE Trans. Mob. Comput.*, vol. 18, no. 3, pp. 702–714, 2019.

[4] T. T.Khanh, T.Hoang Hai, V.Nguyen, T. D. T.Nguyen, N.Thien Thu, andE. N.Huh, "The Practice of Cloud-based Navigation System for Indoor Robot," *Proc. 2020 14th Int. Conf. Ubiquitous Inf. Manag. Commun. IMCOM 2020*, 2020.

[5] S.Kayukawa, T.Ishihara, H.Takagi, S.Morishima, andC.Asakawa, "BlindPilot: A Robotic Local Navigation System that Leads Blind People to a Landmark Object," pp. 1–9, 2020.

[6] T. K.Chuang *et al.*, "Deep trail-following robotic guide dog in pedestrian environments for people who are blind and visually impaired - Learning from virtual and real worlds," *Proc. - IEEE Int. Conf. Robot. Autom.*, pp. 5849–5855, 2018.

[7] H.Yang *et al.*, "Tender Tea Shoots Recognition and Positioning for Picking Robot Using Improved YOLO-V3 Model," *IEEE Access*, vol. 7, pp. 180998–181011, 2019.

[8] J.Ryu andS.Kim, "Chinese character detection using modified Single Shot Multibox Detector," *Int. Conf. Control. Autom. Syst.*, vol. 2018-Octob, no. Iccas, pp. 1313–1315, 2018.

[9] C.Zhao andB.Chen, "Real-Time Pedestrian Detection Based on Improved YOLO Model," *Proc. - 2019 11th Int. Conf. Intell. Human-Machine Syst. Cybern. IHMSC 2019*, vol. 2, pp. 25–28, 2019.

[10] D. Wang, F. Gao, J. Dong and S. Wang, "Change Detection in Synthetic Aperture Radar Images based on Convolutional Block Attention Module," *2019 10th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp)*, Shanghai, China, pp. 1-4, 2019.

[11] J.Guerreiro, D.Sato, S.Asakawa, H.Dong, K. M.Kitani, andC.Asakawa, "Cabot: Designing and evaluating an autonomous navigation robot for blind people," *ASSETS 2019 - 21st Int. ACM SIGACCESS Conf. Comput. Access.*, pp. 68–82, 2019.

[12] Q.Chen, Y.Chen, J.Zhu, G.DeLuca, M.Zhang, andY.Guo, "Traffic light and moving object detection for a guide-dog robot," *J. Eng.*, vol. 2020, no. 13, pp. 675–678, 2020.

[13] J. C.Ying, C. Y.Li, G. W.Wu, J. X.Li, W. J.Chen, andD. L.Yang, "A Deep Learning Approach to Sensory Navigation Device for Blind Guidance," *Proc. - 20th Int. Conf. High Perform. Comput. Commun. 16th Int. Conf. Smart City 4th Int. Conf. Data Sci. Syst. HPCC/SmartCity/DSS 2018*, pp. 1195–1200, 2019.

[14] W.Lan, J.Dang, Y.Wang, andS.Wang, "Pedestrian detection based on yolo network model," *Proc. 2018 IEEE Int. Conf. Mechatronics Autom. ICMA 2018*, pp. 1547–1551, 2018.

[15] Y. Du, M. Gao, Y. Yang, J. Zhang and Z. Yu, "A Target Detection System for Mobile Robot Based On Single Shot Multibox Detector Neural Network," *2018 IEEE 4th International Conference on Control Science and Systems Engineering (ICCSSE)*, Wuhan, China, pp. 81-85, 2018.

[16] S. Kanimozhi, G. Gayathri and T. Mala, Multiple Real-time object identification using Single shot Multi-Box detection,"*2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*, Chennai, India, pp. 1-5, 2019.

[17] X.Zhang, Z.Qiu, P.Huang, J.Hu, andJ.Luo, "Application Research of YOLO v2 Combined with Color Identification," *Proc. - 2018 Int. Conf. Cyber-Enabled Distrib. Comput. Knowl. Discov. CyberC 2018*, pp. 138–141, 2019.

[18] P.Adarsh, P.Rathi, andM.Kumar, "YOLO v3-Tiny: Object Detection and Recognition using one stage improved model," *2020 6th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2020*, pp. 687–694, 2020.

[19] L.Xie andC.Huang, "A residual network of water scene recognition based on optimized inception module and convolutional block attention module," *2019 6th Int. Conf. Syst. Informatics, ICSAI 2019*, no. Icsai, pp. 1174–1178, 2019.

[20] C.Wen, M.Hong, X.Yang, andJ.Jia, "Pulmonary nodule detection based on convolutional block attention module," *Chinese Control Conf. CCC*, vol. 2019-July, pp. 8583–8587, 2019.