Proceedings of the Fourth International Conference on Computing Methodologies and Communication (ICCMC 2020)
IEEE Xplore Part Number:CFP20K25-ART; ISBN:978-1-7281-4889-2

Deep Learning based Object Detection and Recognition Framework for the Visually-Impaired

Swapnil Bhole
Dept. of ECE
NIT Nagpur
swapnil.rajendra@gmail.com

Aniket Dhok
Dept. of ECE
NIT Nagpur
aniketdhok01@gmail.com

***Abstract*—Vision impairment or blindness is one of the top ten disabilities in humans, and unfortunately, India is home to the world's largest visually impaired population. In this study, we present a novel framework to assist the visually impaired in object detection and recognition, so that they can independently navigate, and be aware of their surroundings. The paper employs transfer learning on Single-Shot Detection (SSD) mechanism for object detection and classification, followed by recognition of human faces and currency notes, if detected, using Inception v3 model. SSD detector is trained on modified PASCAL VOC 2007 dataset, in which a new class is added, to enable the detection of currency as well. Furthermore, separate Inception v3 models are trained to recognize human faces and currency notes, thus making the framework scalable and adaptable according to the user preferences. Ultimately, the output from the framework can then be presented to the visually impaired person in audio format. Mean Accuracy and Precision (mAP) scores of standalone SSD detector of the added currency class was 67.8 percent, and testing accuracy of person and currency recognition of Inception v3 model were 92.5 and 90.2 percent respectively.**

***Index Terms*—convolutional neural network, SSD, Inception v3, transfer learning**

I. Introduction

Visually impaired people face a lot of difficulties in their lives. Recent statistics published by World Health Organization (WHO) in 2019 reveal that globally, around 2.2 billion individuals are affected by vision impairment. Detecting and recognizing common objects in the surroundings seem to be a herculean task for the visually impaired individuals. They rely either on other people, which makes the blind dependent on them, or, on their sense of touch and smell to detect objects, which is highly inaccurate and can be hazardous in some cases.

The white cane is the most popular blind navigating device. This was further improved by adding ultrasonic and IR sensors to detect obstacles in the vicinity of the visually impaired user, and provide feedback in the form of vibration or sound. Though this approach was useful for the mobility of the visually impaired user, it provided little or no information about the surroundings. For the user to have a better understanding of the surrounding, objection detection and classification, followed by recognition and audio feedback is crucial.

Neural networks, particularly, convolutional neural networks have shown promising results particularly in object detection, classification, and recognition tasks from images. In [1], the authors use a feed-forward neural network to provide speech suggestions regarding products of shopping. Real-Time smartphone-based obstacle detection and classification system is implemented in [2]. The detection process involves interest point extraction and tracking through multiscale Lucas - Kanade algorithm, background motion estimation using homographic transforms and agglomerative clustering technique, followed by classification with the help of Histogram of Oriented Gradients (HOG) descriptor into Bag of Visual Words (BoVW). A survey on Electronic Travel Aids (ETA) designed for visually impaired navigation assistance is presented in [3]. Various ETAs, their strengths, and shortcomings are discussed and compared feature-wise. It also highlights the fact that no current system incorporates all necessary features and any technology should not attempt to replace the cane stick but to complement it by proper alerting and feedback.

A deep novel architecture for visually impaired employing a late fusion of two parallel CNN's outperforms the state-of-the-art methods for activity recognition [4]. The two CNN's GoogLeNet and AlexNet complement each other in identifying different features of the same class, hence the input video is fed to both of them, and the output class scores are combined using Support Vector Machine (SVM). Yet another novel method proposed in [5] uses CNN followed by a recurrent neural network (RNN) and softmax classifier for object detection, and Hue, Saturation and Intensity (HSI) color thresholding for color recognition. An approach combining computer vision and deep learning techniques for visually impaired outdoor navigation assistant is shown in [6]. The system uses a regression-based mechanism for object tracking without a priori information, handles sudden camera movements, and exploits You Only Look Once (YOLO) for object recognition.

A smartphone app is designed for guiding visually impaired persons in [7]. It can operate in two modes: online and offline based on user network connectivity. The online mode uses Faster RCNN to generate predictions in stable conditions and YOLO for faster results. Whereas, a feature recognition module using Haar features and Histogram of Gradients (HOG) serves this purpose in offline mode. A CNN is designed for pre-trained object recognition using the ImageNet dataset [8]. A novel DLSNF (Deep-Learning-based Sensory Navigation Framework) built on the YOLO architecture is proposed in [9] for designing a sensory navigation device on top of NVIDIA Jetson TX2. SqeezeNet, a light-weight pre-

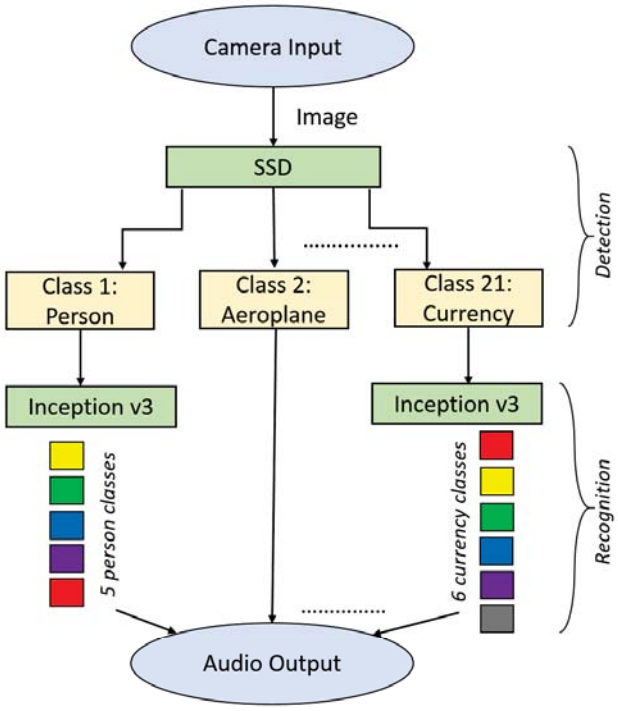

Fig. 1. A schematic of the control flow diagram of the deep learning framework

trained CNN model, achieved better performance and reduced computational latency per image [10]. SqueezeNet is improved by changing the weights of the last convolutional layer, replacing the Rectified Linear Unit (ReLU) with LeakyReLU as activation function and addition of batch normalization layer.

In this study, we propose a deep learning framework for image detection, classification and person-currency recognition. As per the authors' knowledge, no other work has integrated person and currency recognition in addition to object detection and classification for assisting the visually impaired. The input can be taken from a camera. Transfer learning is performed on the SSD-VGG16 model to predict classes of the PASCAL VOC 2007 dataset. In addition to the existing 20 classes in PASCAL VOC 2007, we add new images in the dataset incorporating currency class also, making the total predictable classes to be 21. The detected object classes are produced in the output. Also, if the detected category is either person or currency, the cropped image of the bounding box of the detected class is fed as an input to separate Inception v3 models. The Inception v3 model is trained using transfer learning using ImageNet weights. The system is currently trained to recognize 5 human faces and 6 currency note types and can be modified as per user requirements, making this approach modular, scalable and user-friendly.

The paper is organized as follows. Section II provides details about the proposed design of our deep neural network framework. Section III describes the training details. The results are presented in section IV. Section V gives an insight into the possible future extensions and finally, concludes this paper.

II. Proposed Design

The proposed system consists of four subparts namely, the camera for inputting the image into the framework, object detection and classification module, the face and currency recognition modules, and finally, audio output to the visually impaired user as shown in Fig. 1. In this work, we focus on the design of the framework, which involves object detection, classification and recognition. SSD has shown faster single shot detection results for multiple categories [11]. PASCAL VOC 2007 dataset [12] has 20 classes of objects, containing 9,963 images having 24,640 annotated objects. This dataset was modified by adding more images in the training, validation and testing sets, making the total classes to 21, the new class being currency. This modified PASCAL VOC 2007 dataset is trained using transfer learning on SSD model. Input image is resized to 300 by 300 pixels. SSD300 model was initially loaded with weights from VGG-16 model. The framework makes a bounding box around any of the trained 21 categories if it exists in the input image. Intersection over Union (IoU) method shown in Fig. 2 is used to evaluate the performance. The name of any category other than human or currency, if detected, is fed into audio output. For human or currency classes, one more stage is required which is the recognition part.

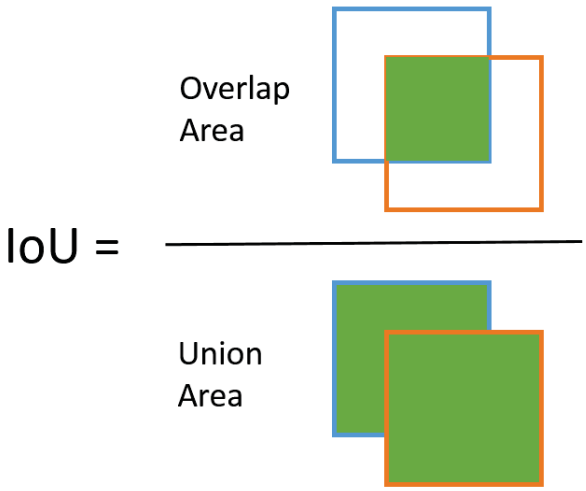

Fig. 2. Intersection over Union (IoU)

Around 15 million labelled images classified into 22,000 categories is available in the ImageNet dataset. Inception v3 model has achieved benchmarks on ILSVR 2012 classification in a modest computational cost [13]. The inception module is shown in Fig. 3. Transfer learning is performed on two instances of this model to recognize human faces and currency respectively. 5 faces are included in the human face dataset and 4 Indian and 2 Nepal currency denominations are used for

currency dataset. The Indian currency dataset used for training one of the Inception v3 model, and for SSD300 currency class was designed by us. Four human faces in dataset for the other Inception v3 model were taken from Labelled Faces in the Wild (LFW) [14], along with one human face category with 240 images added by us. Various works implemented neural networks for currency recognition [15], [16], however, there is a lack of an all in one framework combining common classes classification, human faces and currency recognition, which is achieved in this work.

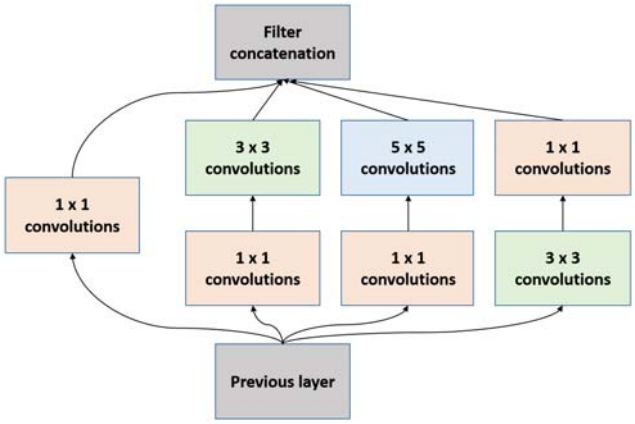

Fig. 3. Inception module

III. Training Details

Intel Core i5 CPU with 8 GB RAM and GTX 1050 Ti as the GPU was used for transfer learning on SSD300 and Inception v3 model, with Tensorflow and Keras library, coded in Python. Annotated images for 20 classes were available in the PASCAL VOC 2007 dataset. For the new class currency that we added, LabelImg Master [17] was used for generating annotations, and bounding box creations in VOC 2007 format. Fig. 4 shows a sample annotation generated by LabelImg for an image containing the currency class.

Rest of the simulation parameters are listed in table I, II, III.

TABLE I
SSD300 TRAINING PARAMETERS

Training set images	2759
Validation set images	2762
Testing set images	5204
No. of classes	21
Epochs	100
Learning Rate	0.001
Optimizer	Stochastic Gradient Descent

IV. Results

The first part of our framework, related to object detection and classification using SSD300 model, was evaluated on the modified PASCAL VOC 2007 dataset containing 5204 test images, and having 21 classes. It achieved a testing accuracy

Fig. 4. Sample figure and its annotation generated by labelImg

TABLE II
HUMAN FACE DATASET FOR INCEPTION V3

Persons	No. of images
Person 1	236
Person 2	222
Person 3	218
Person 4	240
Person 5	231

of 67.8 percent on the added currency class. This shows the model is trained successfully to recognize and classify currency class in addition to 20 other classes in the dataset. The cropped images of the classified human and currency can then be fed to separate Inception v3 models trained for the purpose. Inception v3 model trained on the person dataset achieved a testing accuracy of 92.5 percent and that trained on currency dataset achieved 90.2 percent.

TABLE III
CURRENCY DATASET FOR INCEPTION V3

Currency	No. of images
INR 100	212
INR 200	208
INR 500	264
INR 2000	252
NPR 10	212
NPR 20	212

V. Conclusion & Future Scope

A novel framework employing object detection, classification, and face and currency recognition has been presented to assist the visually impaired people. It is fairly simple, and easy to deploy, once the training part is complete. Using separate Inception models for faces and currency recognition makes it faster, user-specific and adaptable. It is one of the most generic

framework, integrating all the useful features, and will surely prove to be a great service to mankind. Future work can be done to make the face and currency recognition spoof-proof.

REFERENCES

[1] F. Jabeen, A. Muhammad, and A. M. Enriquez, "Feed forward neural network training based interactive shopping for blind," in *2015 12th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*. IEEE, 2015, pp. 1–6.

[2] R. Tapu, B. Mocanu, A. Bursuc, and T. Zaharia, "A smartphone-based obstacle detection and classification system for assisting visually impaired people," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 444–451.

[3] R. Tapu, B. Mocanu, and E. Tapu, "A survey on wearable devices used to assist the visual impaired user navigation in outdoor environments," in *2014 11th international symposium on electronics and telecommunications (ISETC)*. IEEE, 2014, pp. 1–4.

[4] J. Monteiro, J. P. Aires, R. Granada, R. C. Barros, and F. Meneguzzi, "Virtual guide dog: An application to support visually-impaired people through deep convolutional neural networks," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 2267–2274.

[5] R. Kumar and S. Meher, "A novel method for visually impaired using object recognition," in *2015 International Conference on Communications and Signal Processing (ICCSP)*. IEEE, 2015, pp. 0772–0776.

[6] R. Tapu, B. Mocanu, and T. Zaharia, "Seeing without sight-an automatic cognition system dedicated to blind and visually impaired people," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 1452–1459.

[7] B.-S. Lin, C.-C. Lee, and P.-Y. Chiang, "Simple smartphone-based guiding system for visually impaired people," *Sensors*, vol. 17, no. 6, p. 1371, 2017.

[8] K. Potdar, C. D. Pai, and S. Akolkar, "A convolutional neural network based live object recognition system as blind aid," *arXiv preprint arXiv:1811.10399*, 2018.

[9] J.-C. Ying, C.-Y. Li, G.-W. Wu, J.-X. Li, W.-J. Chen, and D.-L. Yang, "A deep learning approach to sensory navigation device for blind guidance," in *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*. IEEE, 2018, pp. 1195–1200.

[10] H. Alhichri, Y. Bazi, N. Alajlan, and B. Bin Jdira, "Helping the visually impaired see via image multi-labeling based on squeezenet cnn," *Applied Sciences*, vol. 9, no. 21, p. 4656, 2019.

[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results," http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[13] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[14] G. B. Huang, M. Mattar, H. Lee, and E. Learned-Miller, "Learning to align from scratch," in *NIPS*, 2012.

[15] N. A. Semary, S. M. Fadl, M. S. Essa, and A. F. Gad, "Currency recognition system for visually impaired: Egyptian banknote as a study case," in *2015 5th International Conference on Information & Communication Technology and Accessibility (ICTA)*. IEEE, 2015, pp. 1–6.

[16] S. Singh, S. Choudhury, K. Vishal, and C. Jawahar, "Currency recognition on mobile phones," in *2014 22nd International Conference on Pattern Recognition*. IEEE, 2014, pp. 2661–2666.

[17] Tzutalin, *LabelImg. Git code (2015)*, 2015. [Online]. Available: https://github.com/tzutalin/labelImg