IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 4, April 2022

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.118

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 4, April 2022||

| DOI:10.15680/IJIRSET.2022.1104137 |

Real Time Object Recognition with Voice Feedback using YOLOv4-tiny Deep Learning Model

Harwant Singh Arri[1], Nirajan Khadka[2], Vishnu Chopra[3], Aman Kumar[4], Subhadeep Karmakar[5], Srijan Kumar Nayak[6]

Assistant Professor, Dept. of Computer Science and Engineering, L.P.U Phagwara, Punjab, India[1]

Student, Dept. of Computer Science and Engineering, L.P.U Phagwara, Punjab, India[2,3,4,5,6]

ABSTRACT: Object recognition is an area of computer vision that looks for meaningful objects in images and movies (by creating a bounding box). An estimate of at least 2.2 billion people is visually impaired worldwide, stated by WHO. Due to visual impairment, it restricts their movement in unfamiliar places. Here we will convert the image to text and then the text to voice in this project. Using this paper,we suggestanintelligent system that detects everyday common objects and produces voice feedback to inform them about the location of the object. First variations of YOLO (You Only Look Once) algorithm are compared and then the best one is used according to result we get it by training it on COCO dataset. After the object is detected, we calculate its spatial location then the voice output is generated by using text-to-speech(gTTS) module so that the visually impaired person will be to understand unfamiliar surroundings.

KEYWORDS- Object detection, YOLO, dataset, Deep Learning, OpenCV, python, web-application, Google Text-To-Speech, voice feedback

I. INTRODUCTION

First, youth is the future of all nations. we can build a more liveable place. As technology advances, traditional methods are being replaced by modern methods. But this evolution is what this world wants. Much research has been done to solve the inconveniences of daily life. Humans can identify objects and its locations in the matter of milliseconds. One of the biggest inconveniences that blind people experience in their daily lives is finding information about objects around them and mobility issues. Simple items are difficult for them to recognize and identifying objects of the same dimension and size is difficult. In this article, we propose an object detection using YOLO (You only looks once) algorithm to detect objects and gTTS module to convert the detected object with its location to voice as feedback. We use YOLOv4-tiny which is a lightweight and faster object detection based on YOLO-v4. Predicting classes and boundary boxes are done quickly for whole frame, so that predictions can be seen in the whole context of frame. After the object is detected and identified it then finds its spatial location in each frame.It then will then be recastinto audio segment using Google-Text-To-Speech(gTTS) module. The YOLO-v4-tiny network is trained on COCO dataset containing approximately 123,287 hand labelled images and is classified into 80 different categories.

1.1 OBJECT DETECTION

We as a human can instantly recognise an object in a matter if milliseconds but computers used to struggle with it. But since with the development of Convolution Neural Networks (CNN), this field has taken a huge leap forward in a very short amount of time. Then methods like Fast R-CNN that can not only detect the object but also can find its spatial location were introduced. But there was a limit of how fast it could detect and identify the object to be used in real time. Since then, You Only Look Once (YOLO) models which was the combination of object recognition, which is fast enough and accurate enough to be used in real time emerged. Since then,better, and faster versions of YOLO have emerged as well.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 4, April 2022||

| DOI:10.15680/IJIRSET.2022.1104137 |

1.2 YOLOv4-TINY

In this project we have used YOLOv4-tiny as our final model. YOLOv4-tiny is based on YOLOv4 method to make it have faster speed of detection. Since it has mAP of about 40.2 on COCO dataset and FPS of about 330 and BFlops of 6.9 on RTX-2070 GPU which meets the demand of a real-time application.
As its backbone network it uses CSPDarknet-53-tiny where CSPBlock module used cross stage partial network. It also used LeakyReLU as its activation function compared to Mish activation function that YOLOv4 used.
Yolov4-tiny uses feature pyramid networks to extract feature maps with different scales. It used two scales feature maps that are 13x13 and 26x26 to predict the detected object. For prediction it first adjusts the size of input to make sure that all input is of same size SxS, and every grid will use B bounding boxes to detect object. Hence it will generate SxSxB bounding boxes on the input and the generated bounding boxes covers the image.When the detected item's core falls within a grid, only the bounding boxes within the same grid will forecast the class of the detected object.
To reduce redundancy of boundary boxes in prediction, confidence threshold is proposed in a sense that if the bounding box confidence threshold is higher than the confidence thresholdgiven, then the bounding box will be keptelse it will be removed.

1.3 SPEECH SYNTHESIS

Speech Synthesis is the imitation of human speech by computers and when it used for the very purpose it will be called asspeech synthesizer. Since this project main idea is to give visually impaired people ability to see through their sense of hearing, voice feedback plays an important role. After the detection of object and its spatial location is done by YOLO algorithm and our program, it is then converted into the language a visually impaired user is most familiar with. To achieve that we use translate module of Python after that use Google Text to Speech(gTTS) to convert the translated text into an audio segment so that a visually impaired person can hear the audio and have visualization of theirsurroundings.

1.4 WEBAPP

Another focus of this project is for anyone to use this specially visually impaired persons very easily, so deploying it as a web application is the way to go. We use Heroku platform to achieve the same since it is used to deploy the project on the web network. Heroku is provides as Platform-as-a-Service (Paas) which can be referred as a cloud computing model. Heroku is a third-party service provider that hosts both hardware and software on its own infrastructure. It supports numerous programming languages such as Python, Ruby, java, NodeJS, etc.
Dyno is the name of the platform container that is used to execute and scale Heroku applications. Dynos are the basic building elements that power up a Heroku application. By deploying their software to Dynos and managing these units, developers may swiftly build and execute scaled applications. There is free availability of Dyno by Heroku platform.
Flask framework is used to build a web application. Flask is a micro-framework that provides tools, technologies, and libraries to build web-applications.

II. RELATED WORK

A Bochkovskiy et al [1] first proposed an architecture called YOLOV4 for exceptional speed and accuracy of detecting object. Then it lists several features that are said to improve CNN accuracy, after which it compares and combines them to obtain results on the COCO dataset. Then it goes on to upgrade state-of-the-arts method to model them more efficient and appropriate so that single GPU training can be accomplished.

Deeksha Janardhan et al [2] proposed a system to make object detection with voice feedback using YOLOv4 model. It first starts by listing out the advantages of yolov4 then it proposes a flowchart for its system designing on implementing it in android environment. Its main idea is to have something that outputs voice real-time based on yolov4 and to be fast enough to be of real assistance to visually impaired people.
He, K., Zhang et al.,[3] suggested a study of traditional and deep learning approaches for object detection The development of the object detection approach focuses on the model's operating principles and performance in real time and precise detection. The difficulties of object detection based on deep learning techniques. The background modelling, object recognition, and background updating operations are all part of the subtraction approach. The

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 4, April 2022||

| DOI:10.15680/IJIRSET.2022.1104137 |

backdrop subtraction method is like the frame difference approach in that it takes a frame and updates it in a timely manner. The deep learning method was used to create the object detection model. The models afterwards presented a fresh suggestion.

Xinyi Zhou et al. [4] elaborated the applications of deep networks in the field of object detection andwith rapid development of deep learning its applications has reached many research areas and have achieved state of the art results. And claims the most popular choice among visually impaired user is Android. However, there are severaldedicated devices for navigation and object recognition are in use. But its main drawback is that they are very expensive in contrast to software. One of the main benefits in using deep learning for object detection is that when trained on big dataset it gives good accuracy, but it is hard to model long dependency using current recurrent neural networks.

Joseph Redmonet al. [5] proposed a "You Only Look Once: Unified, Real-Time Object Detection" to detect real - time objects. It was the first object detection model at that time which used regression to dimensionally separate bounding boxes and its associated class probabilities was proposed as well. It used Single Shot Detector (SSD) to predict bounding boxes and class probabilities from images just from one evaluation. It is optimized end-to-end, since it is a single network model, this architecture was fast. Compared to other detector at that time, it did make more localization error but has very chance to predict negative detection where nothing exists.

Krizhevsky et al [6] used architecture called AlexNet which was the first to win ImageNet challenge. It was a breakthrough network which shifted the focus from hand engineering to DCNN. GPU and Rectified Linear Unit (RElu) were the things which paved the transition from hand crafted engineering to feature learning by CNN paradigm. After that scholar began to study various applications of CNN and deep CNN in object detection as well.

Geethapriya. S et al [7] proposed a paper to understand YOLO in detail. It compares various metrics like speed, accuracy, complexity, and other features as well with different distance estimation and other object detection algorithm. Based on all the metrics it claims that YOLO model looks at the images in a unique way than any other model. By using convolution network, it predicts bounding boxed, and all the class probabilities and it does so faster than any other algorithm. It also gives various information about different fields where YOLO model could be used as well.

III. METHODOLOGY

In the proposed system, we are using YOLO-v4-tiny and adding the ability to translate it into their native language from the existing system. Compared to YOLO-v4, the FPS have increased by eight times.

Other than that, we are using Python3, first we are loading the pretrained weights of YOLO-v4-tiny trained on COCO dataset using OpenCV Deep Neural Network (dnn) module. Then we are using same OpenCV module to take input from the camera to start capturing frames. The frames are then sent **individually to** an algorithm to identify the item. The item that has been recognized is next utilized to determine its spatial position. The text output is then translated into their local language using the translate module, and subsequently into an audio segment using the gTTS module.

The sound feedback is the primary outcome of this system which gives the location of detected object as well as the name of the object to the visually impaired person. Now they can use this to have a visualization of the surrounding around them.

To start, a visual impaired user starts by logging into the webapp. We login credentials matches with data in database (created remote database online) then only we can move forward. After success it opens the webcam. Each frame of the video is then fed into YOLO-V4-tiny model which detects the object and its spatial location. Then the text along with object and location in generated and in parallel to that first the text is translated and then the result is translated into audio feedback. Since this process is done in parallel there tends not be any lag. And since its implemented on the webapp the user specially visually impaired persons don't need to have much setup.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 4, April 2022||

| DOI:10.15680/IJIRSET.2022.1104137 |

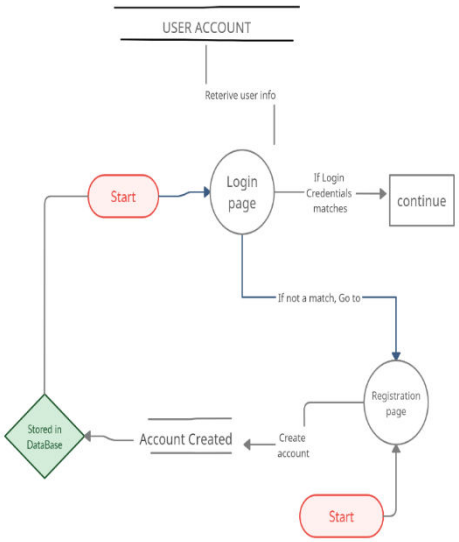

Fig 1.0 Flowchart of Webapp

First, we compare 4 different variations of YOLO model, and select the best one to use it on our case All of these models were trained on RTX 2070 GPU using COCO dataset.

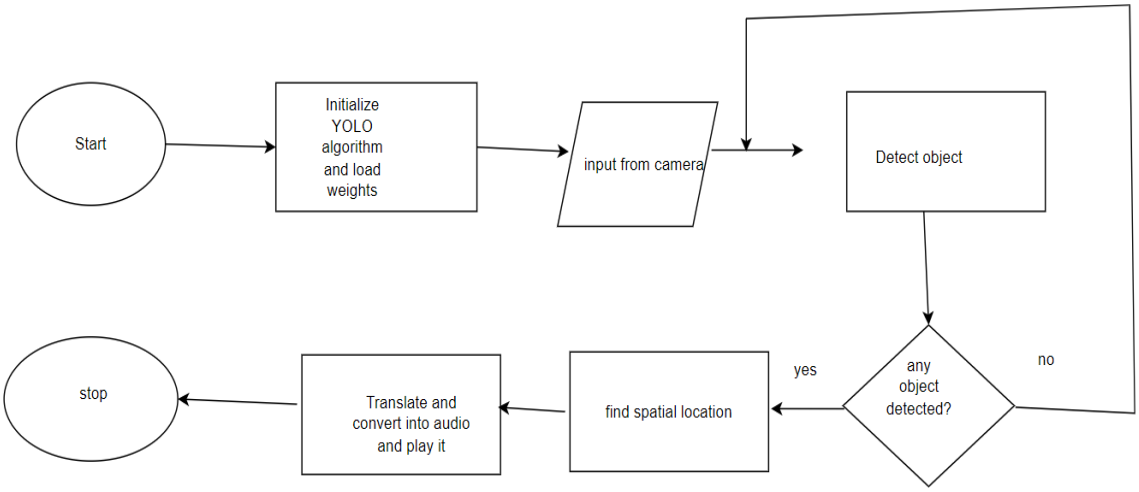

Fig 2.0 Flowchart of system design

mAP=Mean Average Precision
FPS=Frames processed per second
BFlops= Billion Floating point operation per second

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 4, April 2022||

| DOI:10.15680/IJIRSET.2022.1104137 |

Factors Analyzed	YoloV3	YOLOV3-tiny	YoloV4	Yolov4-tiny
Introduced	2018	2020	2020	2020
Number of layers	107	24	53	29
mAP	55.3	33.1	64.9	40.2
FPS	66	345	45	330
BFlops	65.9	5.6	91.1	6.9

Table 1.0 comparison of different YOLO models

From the table above we can clearly see that YOLOv4-tiny is the best one to use since we are implementing in a real-world scenario which needs high FPS and high precision. Compared to YOLOv3-tiny, YOLOv4-tiny mAP has improved by 17.6% but FPS decreased by 4.3% which is acceptable in real world because precision is more significant than FPS.

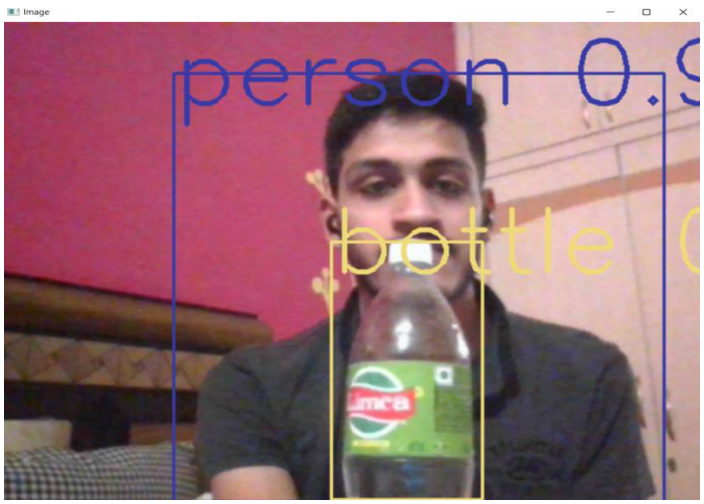

Fig 3.0 Working of the system

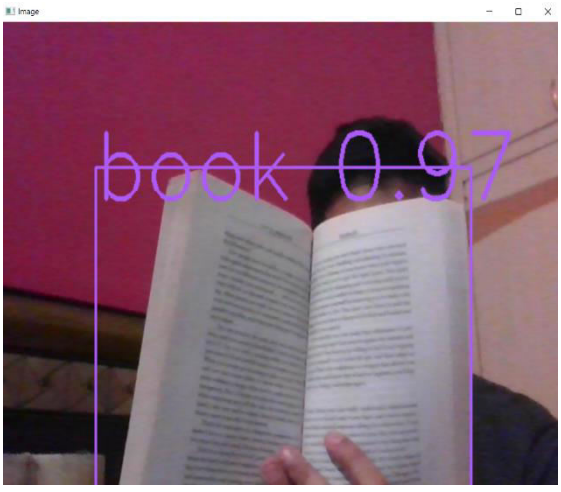

Fig 3.1 Working of the system

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 4, April 2022||

| DOI:10.15680/IJIRSET.2022.1104137 |

IV. CONCLUSION AND FUTURE SCOPE

This research paper suggests using one of the best YOLO models for real time object detection with voice feedback to perform real-time object recognition. It can be used to provide convenience in daily life of a visually impaired person. Also, it can be expected to work in other areas as well to make visually impaired person live their life in an independent way.

In our project we have compared two YOLO models (YOLOv3 and YOLOv4). And we observed that YOLOv4 is faster and efficient than YOLOv3. After testing Weights of both algorithms, YOLO_V4 seems more powerful. We know that there exist many other deep learning algorithms but attaining better accuracy is difficult with normal CPUs. But in future we will try to evaluate with more parameters. For making this project successfully working we have used ready-made dataset from google but later we can develop our own dataset from best use.

Since it is webapp, it could further be commercialized by incorporating high speed, high accuracy model and by leveraging cloud GPU with high-speed connection for all the visually impaired people to use it. Furthermore, special devices like Portable blind aid device [20], Android based object recognition [13] and even IOT and embedded device could also be used to make it more accessible. Also, it could progress even further if application can be made as navigation device that uses 3D sensors to provide better feedback.

REFERENCES

1. YOLOv4: Optimal Speed and Accuracy of Object Detection by Alexey Bochkovskiy, Chien-Yao Wang Institute of Information Science, Hong-Yuan Mark Liao Institute of Information Science Academia Sinica, Taiwan https://arxiv. org/pdf/2004.10934.pdf

2. Deeksha Janardhan, Madhuri B J, Harsha Kumar, Ketan Sahu, Suhas. S in Object Detection with Voice Feedback using YOLOv4

3. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition vol 2, no.3, pp. 770-778.

4. Zhou, X., Gong, W., Fu, W., & Du, F. (2017, May). Application of deep learning in object detection. In 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS) (pp. 631- 634). IEEE.

5. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779- 788).

6. A. Krizhevsky, I. Sutskever, G. Hinton ImageNet Classification with Deep Convolutional Neural Networks, NIPS (2012), pp. 1097-1105

7.Geethapriya. S, N. Duraimurugan, S.P. Chokkalingam, “Real-Time Object Detection with Yolo”, in 2019

8 Zraqou, Jamal S., Wissam M. Alkhadour, and Mohammad Z. Siam.” Real-Time Objects Recognition Approach for Assisting Blind People.”, 2017.

9. Synchronous Object Detection with Voice Feedback by Naman Mishra1, Deepankar Singh.International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395- 0056, p-ISSN: 2395- 0072

10. Potdar, K., Pai, C. D., &Akolkar, S. (2018). A convolutional neural network based live object recognition system as blind aid. arXiv preprint arXiv:1811.10399.

11. Jiwoong Choi, Dayoung Chun, Hyun Kim, and Hyuk-Jae Lee. Gaussian YOLOv3: An accurate and fast object detector using localization uncertainty for autonomous driving. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 502–511, 2019.

12. Huang, R., Pedoeem, J., & Chen, C. (2018, December). YOLO-LITE: a real-time object detection algorithm optimized for non-GPU computers. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 2503-2510). IEEE.

13. Prakash, M. K. D. J., Harish, P., & Deepika, M. K. (2015, March). Android based object recognition into voice input to aid visually impaired. In International Conference on Recent Trends in Engineering Science and Management (pp. 4078-4153).

14. Redmon. J, Farhadi. A, “YOLO9000: better, faster, stronger”, in 2016

15. J. Redmon, “Darknet: Open-source neural networks in c.” http://pjreddie.com/darknet/, 2013–2016.

16. Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, ARXIV, “Object Detection and Distance Estimation Tool for Blind People Using Convolutional Methods with Stereovision”, in 2019

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 4, April 2022||

| DOI:10.15680/IJIRSET.2022.1104137 |

17. Juan du, “Understanding Object detetction based on CNN Family and YOLO”, in 2018
18. Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation”, in 2014’
19.] Redmon, J., Farhadi, A. (2018) Yolov3: An incremental improvement. arXiv: Computer Vision and Pattern Recognition.
20. Evanitsky, Eugene.” Portable blind aid device.” U.S. Patent No. 8,606,316, 10 Dec. 2013.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details