International Journal of Scientific Research & Engineering Trends
Volume 7, Issue 2, March-April-2021, ISSN (Online): 2395-566X

Blind - Sight: Object Detection with Voice Feedback

A. Annapoorani, Nerosha Senthil Kumar, Dr. V. Vidhya
Dept of Information Technology
Sri Venkateswara College of Engineering, Chennai
annu.arunachalam@gmail.com, anshu.senthilkumar23@gmail.com, hodit@svce.ac.in

***Abstract* – Computer vision deals with how computers can be made to gain high-level understanding from digital images or videos. It seeks to automate tasks that the human visual system can do. Humans glance at an image and instantly know what objects are in the image, where they are, and how they interact. An estimate of 285 million people is visually impaired worldwide, stated by WHO. The proposed Blind Sight-Object Detection with Voice Feedback is a computer vision-based application that leverages state of the art object detection techniques. These are employed to detect objects in the vicinity. You Only Look Once (YOLO): Unified, Real-Time Object Detection a new approach to object detection is deployed in this proposed work. YOLO has 75 Convolutional Neural Network (CNN). Image classification techniques are used to identify the features of the image and categorize them into their appropriate class. The COCO dataset used in this project consists of around 123,287 hand labelled images classified into 80 categories. This wide set of data is used to describe spatial relationships between objects and their location in the environment. In addition, an Indian currency recognition module is developed to identify the denominations. The text description of the recognised object will be sent to the Google Text-to-Speech API using the gTTS package. Voice feedback on the 1st frame of each second will be scheduled as an output to help the visually impaired hear what they cannot see.**

***Keywords* – Convolutional Neural Network, YOLO, COCO, gTTS**

I. INTRODUCTION

Humans can pick out objects in their line of vision in a matter of milliseconds. However, a visually challenged person is deprived of this. Blind Sight-Object Detection with Voice Feedback is to enable these visually challenged people hear what they cannot see. Earlier work in object detection has had its application in areas like optical character recognition (OCR), self-driving cars, tracking objects, face detection, face recognition, image retrieval, security and surveillance. It has also been extended to help visually challenged people to recognise the objects. However, these detections have been carried out by developing extensive neural networks like CNN and numerous hidden layers.

This decreases the speed of image retrieval and processing to a considerable amount. Other similar works are based on the local features' extraction concept. The simulation results using SFIT algorithm and key points matching showed good accuracy for detecting objects. The proposed work focuses on using The You Only Look Once (YOLO) algorithm for detecting the objects along with their position and uses the gTTS API to give voice feedback. The algorithm runs through a variation of an extremely complex Convolutional Neural Network architecture called the Darknet. Previously, classification-based models were used and only high scoring regions of the image were considered as a detection and they could be very time-consuming. Instead, YOLO is regression-based. Predicting classes and bounding boxes for the whole image is done quickly in one run of the algorithm, so that the predictions are informed by the global context in the image. In addition, an Indian Currency recognition module is included that recognises the currency and gives the audio output. The YOLO network is trained with a dataset containing approximately 500 images of each currency type. Accuracy is achieved by setting the threshold and increasing the iterations.

The rest of this paper is organized as follows: Chapter 2 presents the related work. Chapter 3 presents the proposed work. Chapter 4 presents the module description. Chapter 5 presents the experimental results for Blind-Sight: Object Detection with Voice Feedback. Chapter 6 concludes the paper.

II. RELATED WORK

Joseph Redmonet al. [1] proposed a "You Only Look Once: Unified, Real-Time Object Detection" to detect real - time objects. Frame object detection as a regression problem to spatially separated bounding boxes and associated class probabilities was proposed. A single neural network predicts bounding boxes and class probabilities directly from full images in one evaluation. Since the whole detection pipeline is a single network, it is optimized end-to-end directly on detection performance. The unified architecture is extremely fast. YOLO model

processes images in real-time at 45 frames per second. A smaller version of the network, FAST YOLO, processes an astounding 155 frames per second while still achieving double the mAP of other real-time detectors. Compared to state-of-the-art detection systems, YOLO makes more localization errors but is far less likely to predict false detections where nothing exists.

KedarPotdar et al. [3] proposed "A Convolutional Neural Network based Live Object Recognition System as Blind Aid " to perform live object recognition. Live object recognition system is used that serves as a blind aid. The act of knowing what object is in front of the blind person without touching it (by hands or using some other tool) is very difficult. In some cases, the physical contact between the person and object is dangerous, and even lethal. A Convolutional Neural Network is employed for recognition of pre-trained objects on the ImageNet dataset. A camera, aligned with the systems predetermined orientation, serves as input to a computer system, which has the object recognition Neural Network deployed to carry out real-time object detection. Output from the network can then be parsed to present to the visually impaired person either in form of audio or Braille text. Efficient weight sharing is one of the major merits of that proposed system. In addition, it uses good feature extractors to extract features. However, it cannot perform future learning which is a demerit.

J.Prakash et al.[6] proposed android based object recognition into voice input to aid visually impaired. The main features of software modules dedicated to the aid of visually impaired or blind users are mentioned. Reduce or elimination of the need for separate dedicated devices for object recognition and motion detection is the achieved. The software modules are designed for Android operating system, used in the majority of smartphones today. Principal component analysis (PCA) algorithm to recognize the object is deployed. To support real-time scanning of objects, a key frame extraction algorithm is framed that automatically retrieves high-quality frames from continuous camera video stream of mobile phones. The sequence is approximately 3 frames per second. The object is recognized then converted into text, BY text to speech application it is converted into the voice output. Easy Integration of voice command on android events like object detection system is done efficiently. However, it is hard to model long dependency using current recurrent neural networks (RNNs).

Xinyi Zhou et al. [7] elaborated on the application of deep learning in object detection. With the rapid development of deep learning, a number of research areas have achieved good results, and accompanied by the continuous improvement of convolution neural networks, computer vision has arrived at a new peak. The most popular choice of Smartphone among visually impaired users is Android based phones. Commonly, the non-operating system devices are not preferred by blind users as they do not offer special functions such as text to speech conversion. A number of dedicated devices for navigation and object recognition are in use. These wearable devices have the disadvantage that they are expensive in comparison to software. Also, the blind users are required to carry a number of gadgets and devices, each for a different purpose such as object identifiers. One of the major reasons to deploy deep learning in object detection is the accuracy of deep learning when trained with a huge amount of data. However, it is hard to model long dependency using current recurrent neural networks.

O.Karaaliet al**.** [2] proposed "Text-to-Speech Conversion with Neural Networks: A Recurrent TDNNApproach" which converts text to speech using recurrent TDNN approach. The design of a neural network that performs the phonetic-to-acoustic mapping in a speech synthesis system was described. The use of a time-domain neural network architecture limits discontinuities that occur at phone boundaries. Recurrent data input also helps smooth the output parameter tracks. Independent testing has demonstrated that the voice quality produced by this system compares favourably with speech from existing commercial text-to-speech systems.

While neural networks have been employed to handle several different text-to-speech tasks, the system is the first system to use neural networks throughout, for both linguistic and acoustic processing. The text-to-speech task was divided into three subtasks, a linguistic module mapping from text to a linguistic representation, an acoustic module mapping from the linguistic representation to speech, and a video module mapping from the linguistic representation to animated images. The linguistic module employs a letter-to-sound neural network and a post lexical neural network. The acoustic module employs a duration neural network and a phonetic neural network. The visual neural network is employed in parallel to the acoustic module to drive a talking head.

III. PROPOSED WORK

The proposed system architecture is shown in Figure 3.

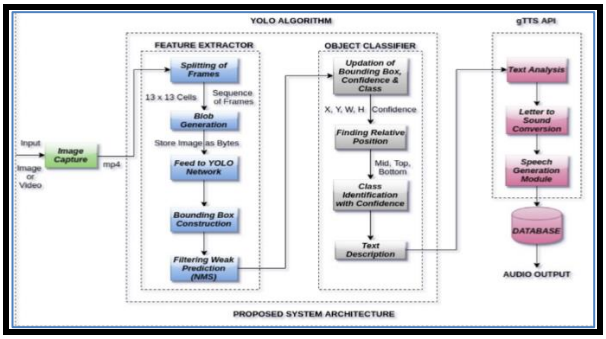

Fig.1. Proposed Blind Sight Architecture

The work in this paper includes the following:

- Input data(image or video) is fetched from the webcam that captures the object.
- Feature Extraction of the image is done in the next step, where the splitting of frames into 13x13 cells takes place along with Blob generation (stores images as bytes) and Bounding box construction.
- The third step involves Object Classification where the confidence (X, Y, W, H) is updated and the relative position is found – Mid, Top, Bottom.
- Once the features have been extracted and the objects are classified, the corresponding text description are sent to the gTTS API where the text is analysed and the letter to sound conversion takes place.
- The output is predicted correctly from the results of the classification and the object is recognised with the audio.
- The input helps in updating the database of the correct object.

IV. MODULE DESCRIPTION

The following gives a brief description about each module in the Blind-Sight: Object Detection with Voice Feedback.

1. Image Capture

The first step in the working of the Blind-Sight application is Image Capturing. Image Capturing is the process of obtaining images from a video which have to be converted into frames. Live stream is captured with a camera. OpenCV provides a very simple interface to this. A video from the camera is captured, converted it into grayscale video and displayed. To capture a video, VideoCapture() object is created. Its argument is either the device index or the name of a video file.

Device index is the number to specify which camera. Here one camera is connected and 0 (or -1) is passed. The second camera is selected by passing 1. After that, frame-by-frame capture is done. At the end release the capture is released. cap.read() returns a bool (True/False).When the frame is read correctly, it is in True state. This is verified at the end of the video by checking this return value. The method cap.isOpened()is initialized. When it is in False state the method is called using cap.open().

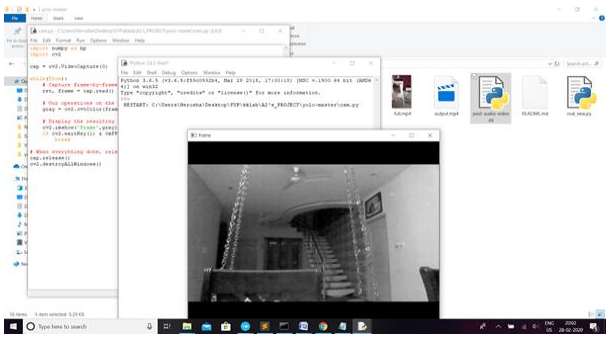

Fig.2. Image Capture from Video

2.Feature Extraction

Feature extraction is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing. A characteristic of these large data sets is a large number of variables that require a lot of computing resources to process the Splitting of Frames, Blob Generation, Feed to YOLO Network, Bounding Box Construction and Filtering Weak Prediction.YOLO uses Non-Maximal Suppression (NMS) to only keep the best bounding box. The first step in NMS is to remove all the predicted bounding boxes that have a detection probability that is less than a given NMS threshold. For example, if we set the NMS threshold to 0.6, this means that all predicted bounding boxes that have a detection probability less than 0.6 will be removed.

3.Object Classification

Object Classification is a classification problem which tends to classify different objects which could flowers, faces, fruits or any object we could imagine. Here apart from common objects, Indian Currency recognition and classification is also developed.After removing all the predicted bounding boxes that have a low detection probability, the second step in NMS, is to select the bounding boxes with the highest detection probability and eliminate all the bounding boxes whose Intersection Over Union (IOU) value is higher than a given IOU threshold. The database is updated with confidence and class.The object classification process includes the following:

- Updationof Bounding Boxes, Confidence & Class
- Finding Relative Position
- Class Identification with Confidence
- Text Description

4. Speech Synthesis

Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a speech computer or speech synthesizer, and can be implemented in software or hardware products. It is the process of generating spoken language by machine on the basis of written input. Speech synthesis involves the process of text analysis and letter to sound conversion.Text Analysis is about parsing texts in order to extract machine-readable facts from them. The purpose of Text Analysis is to create structured data out of free text content. The process can be thought of as slicing and dicing heaps of unstructured, heterogeneous documents into easy-to-manage and interpret data pieces. The annotated text is converted into voice response and gives the basic positions of the objects in the person/camera's view.

V. EXPERIMENTAL RESULTS

In this proposed work, object detection is done using YOLO Object Detection Algorithm. The location of the objects with their respective coordinates are obtained by programming. The class probability and confidence score

together help in determining the object accurately.The following is the result of object recognition. Here the system correctly identifies the objects with the following accuracies:

- Person: 0.9962
- Cell Phone: 0.9595
- Bottle: 0.9937
- Chair: 0.73

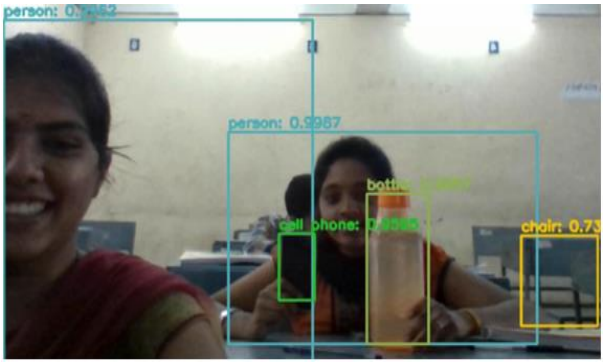

Fig.3.Prediction with Accuracy

The textual annotation is generated along with the position of the objects from the detected image. This is then used to generate the audio feedback.

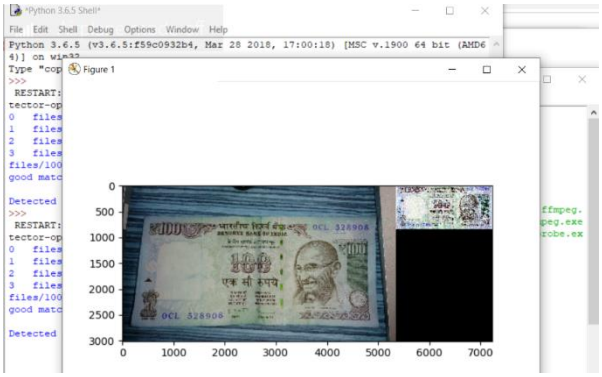

Fig.4 .Prediction with co-ordinates

Similarly, for the currency the textual annotation is generated from the detected and recognised image, and then it is sent to the speech synthesis module for voice feedback.

Fig.5.Currency Detection

```
C:\Users\annu1\Desktop\kklab\A2's_PROJECT\yolo-master>real-time-video.py
C:\Users\annu1\Anaconda3\lib\site-packages\pydub\utils.py:165: RuntimeWarning: Couldn't find ffmpeg
ng to ffmpeg, but may not work
  warn("Couldn't find ffmpeg or avconv - defaulting to ffmpeg, but may not work", RuntimeWarning)
[INFO] loading YOLO from disk...
<class 'cv2.VideoCapture'>
['mid left person']
['mid left person']
['mid center cell phone', 'bottom right person']
['mid center cell phone', 'mid left person']
['mid left person', 'mid center cell phone', 'bottom right person']
['mid left person', 'mid right person', 'mid center cell phone']
```

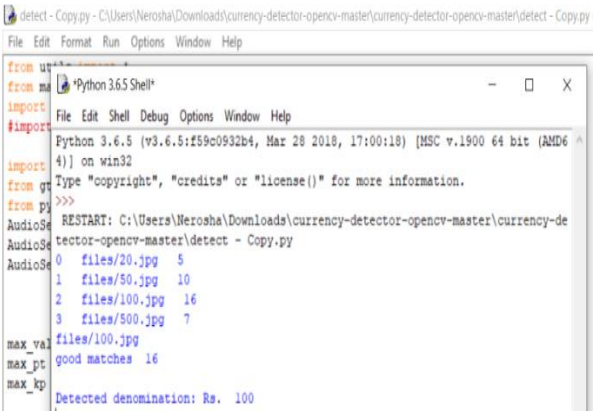

Fig.6.Textual Annotation of the Denomination

VI. CONCLUSION

This project concludes by detecting the Advancement in technology is effectively put to use employing the latest object detection techniques and combining with Google Text to Speech API. It works better than the other existing aids for visually challenged people since it uses YOLO Object Detection which is incredibly fast and can process 45 frames per second. YOLO also understands generalized object representation. This is one of the best algorithms for object detection. In the future, the application can be made as an off-line navigation device that uses 3-D sounds to provide navigation instructions to the user.

REFERENCES

[1] 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).

[2] Karaali, O., Corrigan, G., Gerson, I., & Massey, N. (1998). Text-to-speech conversion with neural networks: A recurrent TDNN approach. arXiv preprint cs/9811032.

[3] 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.

[4] Li, N., Liu, S., Liu, Y., Zhao, S., & Liu, M. (2019, July). Neural speech synthesis with transformer network. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 6706-6713).

[5] Chum, O., &Matas, J. (2005, June). Matching with PROSAC-progressive sample consensus. In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05) (Vol. 1, pp. 220-226). IEEE.

[6] 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).
[7] 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.
[8] Lee, Y., Kim, T., & Lee, S. Y. (2018). Voice imitating text-to-speech neural networks. arXiv preprint arXiv:1806.00927.
[9] Wang, L., Shi, J., Song, G., & Shen, I. F. (2007, November). Object detection combining recognition and segmentation. In Asian conference on computer vision (pp. 189-199). Springer, Berlin, Heidelberg.
[10] Wong, A., Famuori, M., Shafiee, M. J., Li, F., Chwyl, B., & Chung, J. (2019). Yolo nano: a highly compact you only look once convolutional neural network for object detection. arXiv preprint arXiv:1910.01271.
[11] 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.
[12] Choudhury, M. R. (2020, September). AProcess FOR COMPLETE AUTONOMOUS SOFTWARE DISPLAY VALIDATION AND TESTING (USING A CAR-CLUSTER). In CS & IT Conference Proceedings (Vol. 10, No. 11). CS & IT Conference Proceedings.