IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# Real Time Vision-Based Object Detection For The Visually Impaired

**Dr.Anirban Chakrabarty [1,*] , Dr.Monalisha Ghosh [2], Bidisha Patra [3]**

[1]Department of Computer Applications, Future Institute of Engineering and Management, Makaut University, West Bengal, India

[2]Department of Computer Applications; Future Institute of Engineering and Management, Makaut University, 2 67
West Bengal, India

[3]Department of Computer Applications; Future Institute of Engineering and Management, Makaut University, West Bengal, India

***Abstract***—The most essential and significant task in computer vision is object detection whose objective is to identify all the pertinent items in the image, along with their category and position. In recent times computer vision technologies have given precise and interesting results using convolution neural network (CNN) model to identify objects. This study has investigated the application of auditory perception in comprehending visual objects, highlighting the notable parallels between the senses of sight and hearing. In this work, real-time object detection assisted with a voice based an alert system has been created with the goal of conveying the user 18 what all is in the surroundings. This type of system can support the visually challenged persons for detecting objects at a particular distance in their vicinity which will help them move around safely, without crashing into any object. Another innovative aspect of this work is that it converts into Bengali regional language after translation for aid of those people who do not understand English. Performance of the suggested model for object detection was evaluated against contemporary techniques and the outcomes of the trial demonstrated its capacity to achieve a significant accuracy level.

***Index Terms***—real-time object detection; convolution neural network; text-to-speech; voice feedback; visually impaired; deep learning.

## 1. INTRODUCTION

The objective of object detection is to enable the identification and localization of objects within an image or video. In recent times it has been widely practiced and has promising research potential in areas like face detection, pedestrian detection, and automobile detection. Object Detection is actually drawing a bounding boxes throughout the objects detected, which aids in locating where potential objects are in a particular place. Object detection is not similar as image recognition; while image recognition designates a label to an image for example an image of a cat receives the label “cat”. Any image of three cats will also receive the label “cat”. On the other hand object detection sketches a box around each cat and labels it as “cat” and it also predicts where each object is located and what label should be functional, thus more precise information about an image can be obtainable than image recognition.

Since object detection aids in the comprehension and analysis of scenes in pictures or videos, it is closely related to other related computer vision techniques like image recognition and image segmentation [1]. However, there are notable differences. While image segmentation provides a detailed pixel-by-pixel analysis of the components within a scene, image recognition merely assigns a class label to an identified object. Identifying object is distinct from these other jobs since it can locate items inside an image or video. We can then track such objects by counting them as a result.

Over the years with the faster development of technology, several research works have been carried out to solve inconveniences in everyday life, and as a result, various services for the community have been provided. Even though people with visual impairments experience an abundance of difficulties, some of the biggest challenges they encounter every day are getting information about things around them and moving safely indoors. These visually challenged persons have difficulty recognizing simple objects, be it outdoor or indoor objects.

This work aims to assist visually impaired persons who face a lot of hardship in recognizing objects. With advancements in technologies, object detection can be done using pre-trained models available. The work has utilized state of the art "You Only Look Once: Unified, Real-Time Object Detection" [5] YOLOv7 algorithm to identify the object present before the person and COCO dataset is used to train this algorithm. The YOLO algorithm is a single-shot detector that analyzes images utilizing a convolutional neural network (CNN).The objects detection has been done from real-time video taken from the webcam. The label of the object in the frame is recognized and then transformed into audio by using text to speech conversion which will be the anticipated output that can inform a blind person regarding the surrounds. This system provides users with real-time information about their surroundings in the form of voice to alert them and prevent any type of harm.

## 2. PRELIMINARIES OR RELATED WORK OR LITERATURE REVIEW

Earlier also there have been attempts to identify and analyze the problem of object detection based on deep learning techniques like the Regions with CNN features model, developed by a team at Microsoft Research in the early 20$^{th}$ century which used a combination of region proposal algorithm and Convolutional neural networks (CNNs) to detect and localize objects. In recent years, numerous researchers have employed deep learning models, particularly convolutional neural networks (CNNs), to address object detection challenges, leading to the development of various state-of-the-art object detection models[2]. Two main categories can be used to classify the object detection models: two stage detectors such as R-CNN and gated R-NN [3], Mask R-CNN [4], and one-stage detectors such as YOLO [5], SSD [6], and YOLOR [7].A two-stage detector operates in two distinct phases. Initially, it generates a set of candidate regions within the image. These regions are then processed in the second stage for object classification. The two-stage object detection methodology employs two rounds of analysis on the input image to precisely predict the presence and position of objects. In the first round, potential object locations are proposed, and in the second round, these proposals are refined to make final predictions. While the two-stage approach is generally regarded as more accurate than single-stage object detection, it is also more computationally intensive and time-consuming.

The selection of the most suitable option hinges on some specific requirements and constraints of the particular use case. In general, real-world applications are better suited for single-shot object detection, while applications that value accuracy are better served by the two-shot approach. Additionally, object detection techniques can assist individuals with special needs, such as those who are visually impaired, in comprehending the content of images.

Moreover, An investigation was carried out by Bourne et al.[8] estimated that, globally, there will be 115 million completely visually blind people by 2050. An innovative electronic talking stick has been developed to assist the visually impaired by alerting them to obstacles in their path. This system employs optical sensors to detect obstacles, and a control circuit then converts these signals into audible voice

messages. The stick is equipped with earphones on the handle to provide voice guidance to the user. The scanning device is capable of detecting obstacles within a range of one meter [9]. Eugene Evanitsky of Xerox Corporation proposed a portable aid device for the visually impaired, utilizing a cell phone to capture and analyze images. This device classifies the spatial relationships of moving objects, providing the user with environmental information through audio output [10]. Subsequently, Humberto Orozeo Cervantes developed a system that employs image analysis to describe the surroundings of visually impaired individuals. This system utilizes glasses fitted with a camera and sensor to capture images, which are transmitted to a remote computer for processing. The resulting audio output is then relayed back to the user's listening device[11]. A system for object recognition and tracking for the visually impaired, based on CNN technology using Mobile Net architecture, was proposed for its efficiency in computational resources and power consumption [12]. Another model was patented where a computer was used for analyzing the 3-D image for object analysis to produce output data and one device can provide the user with a spatial map of the physical environment through an electrical stimulation pad. [13]. Reference [14] examines the use of wearable devices and smart phones to enhance outdoor mobility for the visually impaired, offering features such as speech-based navigation, recognition of traffic signs, and communication with caregivers.

## 3. PROPOSED METHOD

A study of related work revealed that there are various existing models that do real time object detection with voice feedback but the main demerit of these models is that they use older algorithm for object detection like R-CNN, SSD, YOLOv3, YOLOv4, YOLOv5 or YOLOv6. The main drawback of these object detection algorithms is that the obtained accuracy and its real time speed does not go together, if the accuracy is good then the real time speed is slow and vice versa. The suggested approach has used YOLOv7 which is much faster and accurate because it uses a set of predefined boxes called anchor boxes and these boxes are with different aspect ratios, which are used to identify objects of different shapes, that helps to identify a broader range of objects compared to its previous versions, thus helping to reduce the number of false positive cases along with a better average precision. An outstanding feature of YOLO v7 is its exceptional computational efficiency, significantly surpassing other advanced object detection algorithms in image processing speed. Moreover, its capability to handle higher resolutions enables the detection of smaller objects with superior accuracy. Other than that, we have used Python3 for this work, the camera is initialized by using JavaScript and the camera starts capturing frames with the rate of 5 to 160 frames per second and feeds them to the algorithm. Then, the system employs YOLOv7, trained on the COCO dataset, to recognize the object placed in front of the user. A Python library called gTTS is used to transform the recognized object into spoken words. gTTS acts as a bridge between your program and Google's text-to-speech service.

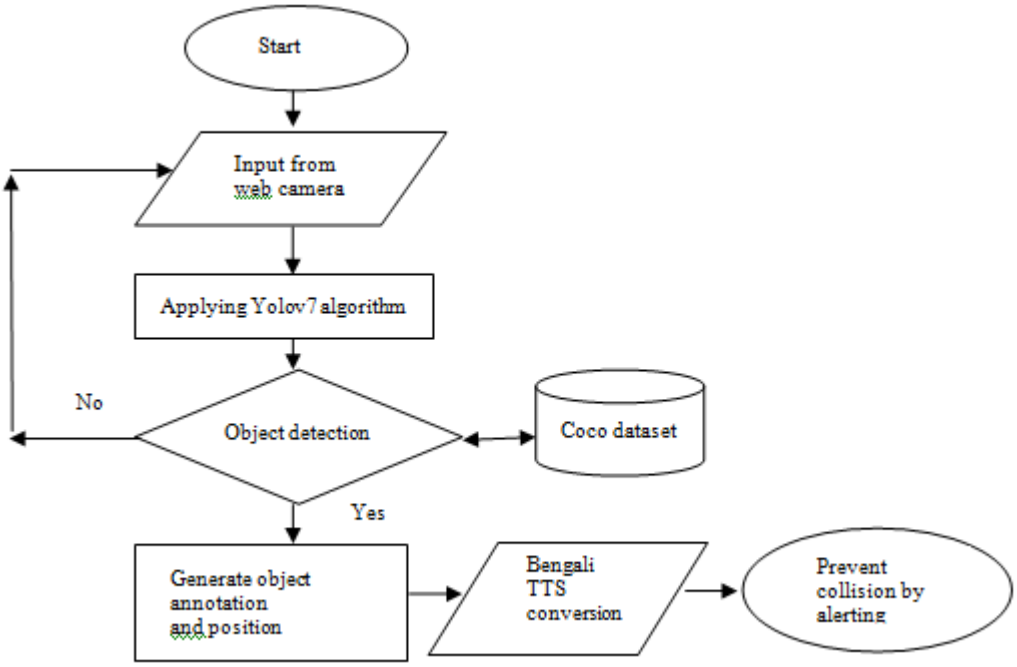


**Fig 1**: Flow chart of proposed model.

Object detectors can be made as a portable device that can be fitted or carried anywhere to detect and recognize objects. The proposed model integrated a text-to-speech engine to convert predicted object names and annotations into speech, aiding Bengali-speaking individuals with visual impairments to recognize objects from a distance. Thus the proposed system can protect the person from colliding with the objects around him hence securing him from injuries. This could be a groundbreaking way to help Bengali-speaking people who are visually impaired by combining these services with an object detection model.

## 4. Tools and Methods

### 4.1 Experimental Environment

Google Colaboratory (Colab) was selected as the platform for developing the proposed system [15]. Google's free cloud service, Colab, gives researchers access to powerful CPUs and GPUs in the cloud. This helps advance the fields of machine learning and artificial intelligence. One of the main features of Colab is that the essential Python libraries such as TensorFlow, and Matplotlib are already installed and can be imported when needed. The tools that were used are: Python 3.5, OpenCV was used to access the webcam in an efficient manner and apply object detection to each frame, Numpy 1.14. The object identified has been converted to an audio segment using gTTS. gTTS is a Python library and tool to interface with Google, translates text-to-speech API. It converts from document to spoken mp3 information that gives the spatial name of the object to the blind person in Bengali language.

### 4.2 Dataset acquisition

The Coco dataset was used in this investigation which is extensively used in object detection as it encapsulates a large number of uniform images [16]. The primary aim of this dataset is to capture realistic scenes with recognizable objects. The category id and segmentation mask of the item are among the many fields that are present in each object instance annotation. Whether an instance represents a single object or a collection of objects determines the segmentation format.. The images are associated with annotations that are stored using JSON, and the x and y coordinates of the bounding box of every image makes the dataset fit for supervised learning problems.

The COCO dataset consists of many images, annotations, and categories. The approximate sizes of the different components of COCO dataset is mentioned below.

a. Images: i. Training set: Around 1,18,000 images.
ii. Validation set: Around 5,000 images.
iii. Test set: Around 20,000 images.
b. Annotations: i. Training set: Around 84,000 annotated objects.
ii. Validation set: Around 36,000 annotated objects.
c. Categories: i. household items, person, vehicles, animals, fruits.

In order to ensure all the images were compatible with the machine learning model, they were resized to a standard dimension during preprocessing, where all images in the dataset were resized to a square format of 1024x1024 pixels by padding with zeros processed together in groups. To shorten the suggested model's training and inference times, image resizing was used. The Yolov7 model was created in the training mode, and then the Pascal VOC dataset weights were loaded into the model. All the output layers for the classification label and bounding boxes were removed and new output layers were defined and trained on the MS COCO, where the pre-trained weights from Pascal VOC dataset were not used.

### 4.3 Defining the metrics and parameters

When creating a deep learning model, it is essential to tune the right parameters because they affect the model's performance during training. Table 1 shows some of the designed object detection model's parameters, as well as the parameters that were tuned during the design phase of the object detection model.

| Parameter | Value |
|---|---|
| no_of_Classes | 8 |
| learning_rate | 0.002 |
| validation_steps | 50 |
| min_confidence | 0.5 |
| steps_per_period | 750 |

**Table 1.** Parameters used for object detection

The performance of the created object detection model was assessed using the assessment metrics mean average precision (mAP) and average precision (AP).

Average precision (AP): The AP was computed for each class by using the 10 point interpolation method, where the precision values are interpolated against 10 equally spaced recall values. The interpolated precision is the highest precision against the recall value larger than the present recall value. This is given by the formula below:

$$AP = \frac{1}{10} \sum_{r \in R} p(r) \quad (1)$$

where R indicates the 10 equally spaced recall values, and p represents the interpolated precision.

Mean average precision (mAP): The calculation of AP comprises only one class. However, in object detection, there is usually more than one class. The mAP is defined as the mean of the AP across all classes, as follows:

$$mAP = \frac{\sum_{i=1}^{n} AP_i}{n} \quad (2)$$

## 5. Results

We conducted a comprehensive assessment of our novel object detection system, focusing on its performance in real-world settings encountered by visually impaired individuals. Our evaluation utilized a dataset comprising around 118,000 images for training purpose. A diverse class of objects like humans, animals, furniture, appliances and outdoor structures has been considered for evaluation. The model is trained in such a way that it detects objects correctly with the extracted features and boundary boxes. It can detect multiple objects in one frame. The average precision and average recall values along with the accuracy of detection are displayed for different objects.


**Figure 2**. Images of real-time detected objects

| Class | Average precision | Average Recall | Class | Average precision | Average Recall |
|---|---|---|---|---|---|
| Person | 0.91 | 0.87 | Dog | 0.79 | 0.81 |
| Cell Phone | 0.95 | 0.94 | Bed | 0.89 | 0.86 |
| TV | 0.94 | 0.92 | Sofa | 0.92 | 0.93 |
| Chair | 0.94 | 0.94 | Fridge | 0.89 | 0.91 |

**Table 2.** Average precision and recall results indoor and outdoor objects


**Figure 3**. Experimental Results (% accuracy and % F1 value)

Our system delivered compelling results, boasting an impressive accuracy of 95%, coupled with average precision of 93% and average recall value of 88% for different categories of objects. These metrics highlight the efficacy of our system in object detection within real-world contexts, surpassing existing methodologies documented in the literature. A comparison of results of the proposed approach with other previous works is shown in table-3 below.

| **Reference** | **Year** | **Basic Features** | **Applied model** | **Evaluation metrics** |
|---|---|---|---|---|
| 12 | 2022 | Object identification and tracking | MobileNet | accuracy = 83.3% |
| 14 | 2021 | Crosswalk detection and navigation | OpenCV | accuracy = 84.5% |
| 17<br>**3** | 2019 | Vision assistance with wearable sensors | Custom CNN | Precision = 0.766, Recall = 0.164 |
| 18 | 2021 | Object detection and navigation | YOLOv3 | accuracy = 81% |
| 19 | 2022 | Traffic light detection and classification | OpenCV | accuracy = 87.5% |
| 21 | 2023 | Object detection and navigation | Blind's appron | N/A |
| 22 | 2022 | Animal detection and classification | SSD ResNet-50 | mAP = 93.5% |
| **Proposed approach** | 2023 | object detection using voice output | YOLOv7 | Accuracy: 92%<br>F1 value: 95% |

**Table 3.** Comparison of the proposed model with other similar works

In this object detection model, 8 different classes were identified. Using a straightforward Python program and the Google text-to-speech library, we created an audio feedback file for every class label. These files are then saved in MP3 format, with each class label corresponding to a different filename. Finally, all of the MP3 files were stored in a folder, whenever the proposed model detects an object, the corresponding MP3 file is called by the program. The corresponding MP3 file is then played back in Bengali language to the user for audio feedback. In this way, The audio outputs of the detected items are produced by the suggested system. In summary, our findings underscore the efficacy of our proposed approach in object detection for the visually impaired. However, further research is warranted to extend our system's applicability to diverse environments and objects, while also integrating seamlessly with existing assistive technologies. However, the system might find it a little difficult to accurately detect objects when there are some background disturbances or if the images are blurred.

## 6. CONCLUSIONS

This work has implemented the use of auditory senses to comprehend visual items as sense of hearing and sight sense shares a prominent similarity. In this work Voice feedback system combined with real-time object detection has been developed with the goal of conveying the user what all is in the surroundings. This system can be helpful for alerting the visually impaired persons about the nearby objects' confinement without crashing into any object. Another innovative aspect of this work is that it converts into Bengali regional language text after translation for aid of those people who do not understand English. The efficiency of the suggested object detection model was compared to earlier models and The results showed that it could reach a notable degree of accuracy. Further research can be done to implement this system in any portable device for visually impaired people to detect objects at a particular distance from them.

**Author Contributions:** Conceptualization: A.C., M.G.; methodology, A.C.; software: M.G.; validation: A.C., M.G,B.P.; formal analysis: M.G.; investigation: A.C., M.G, B.P..; resources: A.C.; data curation: M.G, B.P.; writing—original draft preparation: A.C.; writing—review and editing: A.C., M.G., B.P.; visualization: M.G..; supervision: A.C.;

**Funding:** This research received no external funding.

**Data Availability Statement:** The data used in this study is available upon reasonable request to the corresponding author.

**Acknowledgments:** The authors express their gratitude to their respective institutions for their invaluable support throughout this study.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

[1] Wafa M.Elmannai, Khaled M.Elleithy. “A Highly Accurate and Reliable Data Fusion Framework for Guiding the Visually Impaired”. *IEEE Access*: 33029-33054, 6, 2018, doi: 10.1109/ACCESS.2018.2817164.

[2] Zhao, Z.; Zheng, P.; Xu, S.; Wu, X. Object Detection With Deep Learning: A Review. IEEE Transactions on Neural Networks and Learning Systems, 3212–3232,30(11), 2019 doi: 10.1109/TNNLS.2018.2876865.

[3] Wang, J.; Hu, X. Convolutional Neural Networks with Gated Recurrent Connections. *IEEE Trans. Pattern Analysis and Machine* Intelligence, 3421–3436, 44, 2022, doi: 10.1109/TPAMI.2021.3054614

[4] He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988, doi: 10.1109/ICCV.2017.322.

[5] Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp.779–788, doi: 10.1109/CVPR.2016.91.

[6] Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Computer Vision-ECCV 2016; Leibe, B., Matas, J., Sebe, N.,Welling, M., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2016; pp. 21–37, ISBN: 978-3-319-46447-3.

[7] Wang, C.-Y. Yeh, I.-H., Liao, H.-Y.M. You Only Learn One Representation: Unified Network for Multiple Tasks. arXiv 2021, arXiv:2105.04206.

[8] Bourne, R.R.A.; Flaxman, S.R.; Braithwaite, T. Magnitude, Temporal Trends, and Projections of the Global Prevalence of Blindness and Distance and near Vision Impairment: A Systematic Review and Meta-Analysis. Lancet Glob. Health 2017, 5, e888–e897, doi: 10.1016/S2214-109X(17)30293-0.

[9] Chi-Sheng, Hsieh. ”Electronic talking stick for the blind.” U.S. Patent No. 5,097,856, 24 Mar. 1992.

[10] Evanitsky, Eugene. ”Portable blind aid device.” U.S. Patent No. 8,606,316, 10 Dec. 2013.

[11] F. Ashiq, M. Asif, M.B. Ahmad, S. Zafar, K. Masood, T. Mahmood, M.T. Mahmood, I.H. Lee, Cnn-based object recognition and tracking system to assist visually impaired people, IEEE Access 10 (2022) 819–14 834, doi: 10.1109/ACCESS.2022.3148036.

[12] Cervantes, Humberto Orozco. ”Intelligent glasses for the visually impaired.” U.S. Patent No. 9,488,833. 8 Nov. 2016.

[13] A. Bhattacharya, V.K. Asari, Wearable walking aid system to assist visually impaired persons to navigate sidewalks, in: Applied Imagery Pattern Recognition workshop, 2021, pp. 1–7. doi: 10.1109/AIPR52630.2021.9762132

[14] Google Colaboratory. Available online: https://colab.research.google.com/notebooks/intro.ipynb (accessed on 9th May 2023).

[15] https://cocodataset.org/#explore (accessed on 21st April 2023)

[16] B. Jiang, J. Yang, Z. Lv, H. Song, Wearable vision assistance system based on binocular sensors for visually impaired users, IEEE Int. Things J. 6 (2019) 1375–1383, doi:10.1109/JIOT.2018.2842229.

[17] N. Kumar, A. Jain, Smart navigation detection using deep-learning for visually impaired person, in: International Conference on Electrical Power and Energy Systems, 2021, pp. 1–5, doi:10.28945/5006.

[18] R. Khalid, M.W. Iqbal, N. Samand, M. Ishfaq, R. Rashed, S. Rafiq, Traffic light issues for visually impaired people, J. Jilin Univ. 371 (391) (2022) 3–11, doi: 10.17605/OSF.IO/JSUYB

[19] Nada Alzahrani and Heyam H. Al-Baity. “Object Recognition System for the Visually Impaired: A Deep Learning Approach using Arabic Annotation”, Electronics 2023, 12, 541. https://doi.org/10.3390/electronics12030541

[20] M. Bala, D. Vasundhara, H. Akkineni, C. Moorthy, Design, development and performance analysis of cognitive assisting aid with multi sensor fused navigation for visually impaired people, J. Big Data 10 (2023), https://doi.org/10.1186/s40537-023-00689-5.

[21] K. Manjari, M. Verma, G. Singal, N. Kumar, QAOVDetect: a novel syllogistic model with quantized and anchor optimized approach to assist visually impaired for animal detection using 3D vision, Cogn. Comput. 14 (2022), doi:10.1007/s12559-022-10020-8.