Computers, Materials & Continua
DOI:10.32604/cmc.2022.020827
Article

Tech Science Press

Deep Learning Based Audio Assistive System for Visually Impaired People

S. Kiruthika Devi* **and C. N. Subalalitha**

Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, 603203, India
*Corresponding Author: S. Kiruthika Devi. Email: kiruthis2@srmist.edu.in
Received: 10 June 2021; Accepted: 05 August 2021

Abstract: Vision impairment is a latent problem that affects numerous people across the globe. Technological advancements, particularly the rise of computer processing abilities like Deep Learning (DL) models and emergence of wearables pave a way for assisting visually-impaired persons. The models developed earlier specifically for visually-impaired people work effectually on single object detection in unconstrained environment. But, in real-time scenarios, these systems are inconsistent in providing effective guidance for visually-impaired people. In addition to object detection, extra information about the location of objects in the scene is essential for visually-impaired people. Keeping this in mind, the current research work presents an Efficient Object Detection Model with Audio Assistive System (EODM-AAS) using DL-based YOLO v3 model for visually-impaired people. The aim of the research article is to construct a model that can provide a detailed description of the objects around visually-impaired people. The presented model involves a DL-based YOLO v3 model for multi-label object detection. Besides, the presented model determines the position of object in the scene and finally generates an audio signal to notify the visually-impaired people. In order to validate the detection performance of the presented method, a detailed simulation analysis was conducted on four datasets. The simulation results established that the presented model produces effectual outcome over existing methods.

Keywords: Deep learning; visually impaired people; object detection; YOLO v3

1 Introduction

In recent times, Artificial Intelligence (AI) models started yielding better outcomes in terms of voice-rich virtual candidates like Siri and Alexa [1], independent vehicles (Tesla), robotics (car manufacturing), and automated conversion (Google translator). In line with this, AI-based solutions have been introduced in assistive techniques, especially in guiding visually-impaired or blind people. Mostly, the systems mentioned above overcome the independent navigation issues with the help of portable assistive tools such as infrared sensors, ultrasound sensors, Radio Frequency Identification (RFID), Bluetooth Low Energy (BLE) beacon and cameras. Followed

by, in autonomous direction, visually-impaired people require some other assistive models too. For example, computer vision methods are to be unified with Machine Learning (ML) model to provide moderate solutions for the above-defined problem. For instance, a computer vision module is projected to examine the currency with the help of Speeded-Up Robust Features (SURF) [2]. Also, the system is capable of recognizing US currencies with true recognition rate and false recognition rate. Alternatively, visually-impaired users can shop in departmental stores through prediction whereas barcode analysis provides the visually-impaired individuals with details about the product through voice communication.

Chen et al. [3] presented a model to guide the visually-impaired people to analyze and go through the content. In this prediction model, the candidate regions are initially predicted with a text of special statistical features. Followed by, commercial Optical Character Recognition (OCR) software is applied to examine the content present inside the candidate regions. Alternatively, the application is based on travel assistant model that predicts and examines the public transportation modes. Further, the system predicts the text writings on buses and stations and inform the visually-impaired users regarding station names, numbers, bus numbers and target location, etc. Moreover, Jia et al. [4] reported an issue in identifying the staircases within buildings and the model informs the user only if they are nearby 5 m of the staircase. It depends upon iterative preemptive Random Sample Consensus (RANSAC) method to predict the number of steps in the staircase. However, the performance of the model was determined based on the calculation of doors within the buildings and by examining general as well as stable properties of doors like edges and corners. Consequently, a model was used for predicting restroom signage based on Scale-Invariant Feature Transform (SIFT) characteristics. Object prediction and analysis are highly studied problems in computer vision applications.

The object prediction models [5], developed earlier, were constructed on the basis of extracting hand-engineered attributes, before implementing a classification method. Later, the reformation of Neural Networks (NN) in 2012 and the advent of advanced architectures such as Convolutional Neural Network (CNN), Region CNN (RCNN), You Only Look Once (YOLO), Single Shot Multi-Box Detector (SDD), pyramid system, and Retina-Net networks have simplified the process. In spite of high efficiency, the maximum processing costs make it difficult to implement on wearable devices. As a result, visually-impaired users make use of portable devices to predict objects. Hence, the researchers have come up with a cost-effective and efficient solution that can predict several objects. However, detecting the accurate location of objects is still a challenge.

In this background, the current research article presents an Efficient Object Detection Model with Audio Assistive System (EODM-AAS) using DL-based YOLO v3 model for visually-impaired people. The aim of the research article is to develop a model that can generate a comprehensive description of the objects around visually-impaired people. The presented model includes a YOLO v3 model for multi-label object detection. Also, the presented model computes the position of object in the scene and lastly, it creates an audio signal to inform the visually-impaired persons. In order to validate the detection performance of the presented model, a comprehensive simulation analysis was conducted on four datasets namely, David3, Human4, Subway and Hall Monitor.

Rest of the paper is organized as follows: Section 2 presents a review of state-of-the-art techniques for object detection and classification for assisting visually-impaired people. Section 3 describes the proposed EODM-AAS model and its implementation details. This is tailed by Section 4 in which the experimental analysis of the proposed model on four different datasets

and comparison with other models are discussed. Finally, Section 5 contains the conclusion and future enhancement of the work.

2 Related Works

The challenges involved in object classification include dynamic modifications in natural scenarios and visual features of the objects (color, shape and size). If effective models are planned to be deployed for object classification, it should consider the scenarios from unseen conditions and correlate the object features as well. In this section, various state-of-the-art models for object prediction and classification, exclusively meant for helping the visually-impaired people, have been discussed. Lin et al. [6] employed FRCNN and YOLO models for real-time object detection to help the visually-impaired people. These models alert the users about the objects around them. In this work, the researchers classified the detected object as either 'normal object' or 'emergency object' according to the class identified and relative of the object from visually-impaired person. Accordingly, the visually-impaired persons are alerted. Furthermore, developers have extended the work using bone conduction headphones so that the users can listen the audio signals.

Lakshmanan et al. [7] presented a system to help the visually-impaired users by providing instructions to them with which they can perform a collision-free navigation. This model applies a prototype with Kinect sensor attached to the walking stick that predicts the velocity of moving objects using the estimated depth map. Huang et al. [8] proposed a novel approach to predict static as well as dynamic objects with the help of depth information generated by connected component model, designed based on Microsoft Kinect sensor. Static classes such as rising stairs and steep stairs are identified whereas dynamic classes are detected as dynamic and are not further classified. Poggi et al. [9] applied DL method in the classification of objects with the help of CNN.

Vlaminck et al. [10] utilized RGBD camera tracking method to localize the objects with the help of color and depth information. After that, object classification is carried out by obtaining geometrical features of the object. The developers focused on three classes such as staircases, room-walls and doors. Vlaminck et al. [11] applied 3D-sensors to detect the objects in indoor environment. The researchers focused on detecting four classes such as steps, wall, door and bumpy surface of the floor.

Hoang et al. [12] applied mobile Kinect bounded on a user's body. It predicts both static as well as dynamic objects and inform the same to visually-impaired people. Moreover, it predicts the people with the help of Kindest SDK sensor and considers the depth of image as input. Further, static objects such as ground and wall are forecasted using plane segmentation in this research work. A modern ultrasonic garment prototype was presented in literature [13]. It is a real-time adaptive object classifier which applies acoustic echolocations to extract the features of objects in navigational path of visually-impaired users.

Takizawa et al. [14] presented an object recognition model with the help of computer vision technique like edge prediction that can guide the visually-impaired users to identify the type of object. Mandhala et al. [15] proposed machine learning-based solution named clustering technique to classify the multi-class object. Bhole et al. [16], used deep learning techniques such as Single Shot Detector (SSD) and Inception V3 to classify bank concurrency notes in real-time environment. Vaidya et al. [17] presented an image processing method with machine learning approach to classify the multiclass objects.

When reviewing the state-of-the-art techniques proposed so far to assist the visually-impaired people, most of the models incorporated several sensors to detect the objects. Sensor-based

detection techniques have their own setbacks too such as high cost, power consumption and limited accuracy. These drawbacks are experienced when it comes to object detection with respect to distance from the visually-impaired persons during their navigation. A few research works has focused on the prediction of multi-class objects using machine learning models. Those models are heavier in terms of computation and memory resource that may not be suitable for embedding with real-time assistive tools. Recent advancements in computer vision era and deep learning play an important role in the field of object detection. Though several deep learning algorithms have been proposed for object detection applications, it is still challenging to localize the object rather than recognizing it. Hence, an efficient deep learning model is the need of the hour which should be able to locate multiple–objects, classify the multi-class objects found in the scene, should be a light weight model and should attain maximum accuracy in real-time environment at minimum time. The one stage object detector i.e., You Only Look Once (YOLO) [18] algorithm is the suitable algorithm to meet the requirement.

The key aim of this proposed approach is to guide the visually-impaired users in unknown places by directing them through vocal messages regarding the position of object identified in the scene and its class name. For objection detection and classification, Yolo v3 is used due to its agility in predicting real-time objects [18]. The next section describes the working process of the proposed EODM-AAS model.

3 The Proposed EODM-AAS Model

Fig. 1 shows the workflow of the proposed EODM-AAS model. Initially, the input video undergoes frame conversion process during when the entire video is segregated into a set of frames. Then, object detection process takes place using YOLO v3 model to identify the set of objects in the frame. Followed by, the position of the object in the scene is determined. At last, an audio signal is generated using Python package called pyttsx to notify the visually-impaired people effectually.

3.1 Object Detection

Primarily, every frame undergoes YOLO v3-based object detection process to identify and classify multiple objects in the frame. YOLO v3, with an input image size of $416 * 416$ pixels, has been used in current study. Besides, YOLO v3 is generally trained on COCO dataset comprised of a total of 80 objects. However, in this work, YOLO v3 model was used to predict 30 different object classes connected with visually-impaired persons. YOLOv3 is a 3rd generation product of YOLO semantic segmentation model [19]. It accomplishes both classification and regression tasks by detecting the classes of objects and its location. Hence YOLO v3 is highly suitable for assisting visually-impaired persons. YOLOv3 follows the same procedure of classification and regression alike its previous versions such as YOLO v1 and v2 and YOLO9000, the variants of YOLO family. YOLO v3 imbibes most of the elements from v1 and v2. In addition to that, it also makes use of Darknet 53 [20], with convolution layer, and Resnet connections to eliminate the issue of diminishing gradients. In prediction state, FPN (Feature Pyramid Network) applies three scale feature maps in which minimum feature maps offer semantic details whereas maximum feature maps offer fine-grained details. Then, instead of SoftMax, independent multinomial logistic classification is applied in YOLOv3 structure while binary cross-entropy loss for class prediction is applied during training stage.

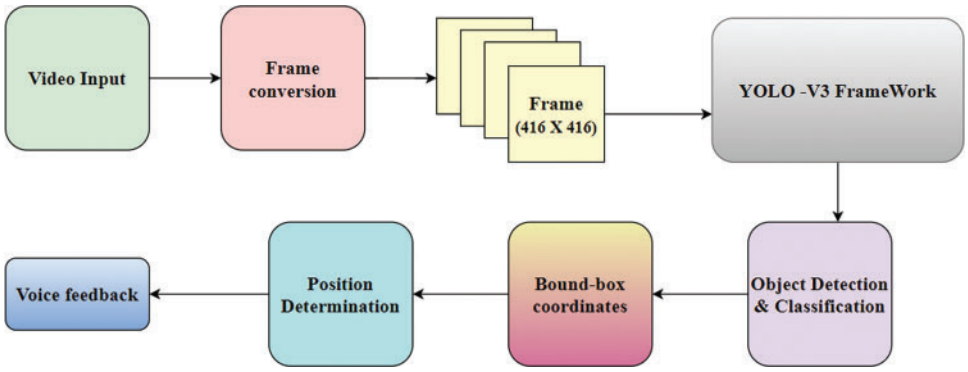

Figure 1: Work flow of EODM-AAS model

YOLOv3 has darknet-53 feature-extraction system and YOLO prediction layer which incur lower processing costs and can be applied in embedded device platforms. Also, the actual input of darknet is 416 × 416 pixels. Hence, the prediction of multi-scale targets can be accomplished by generating pixel feature maps. The density of pixel grid results in limited down-sampling of the iterations which in turn activates the prediction of small targets. Fig. 2 shows the architecture of YOLO v3 [19].

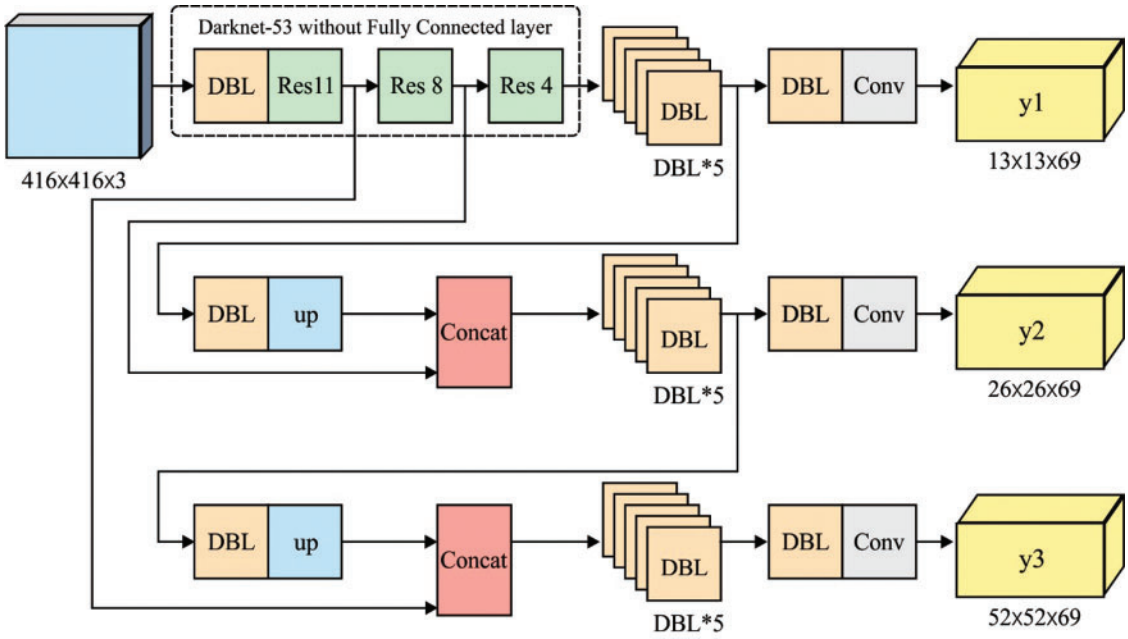

Figure 2: Architecture of YOLO v3

YOLOv3 model assumes object prediction as a regression problem. It forecasts class probabilities as well as bounding box offsets, from complete images, using single feed forward CNN. It intends to eliminate region proposal simulation, feature resampling and summarization at every step in a single network so as to make an end-to-end prediction approach. This method classifies input image as tiny grid cells. When an intermediate portion of an object comes under a grid cell,

then the grid cell is answerable for object prediction [21]. A grid cell detects the location details of B bounding boxes and estimates the objectness values, equivalent to the bounding boxes, using Eq. (1).

$$C_i^j = P_{i,j}(Object) * IOU_{pred}^{truth} \tag{1}$$

where C_i^j implies the objectness value of jth bounding box in ith grid cell. $P_{i,j}$(Object) refers to a function of object. IOU_{pred}^{truth} indicates Intersection Over Union (IOU) from the predicted box as well as ground truth box. Also, YOLOv3 scheme applies binary cross-entropy of the examined objectness values as well as truth objectness, as the portion of loss function is depicted as follows.

$$E_1 = \sum_{i=0}^{S^2} \sum_{j=0}^{B} W_{ij}^{obj} \left[\hat{C}_i^j \log\left(C_i^j\right) - \left(1 - \hat{C}_i^j\right) \log\left(1 - C_i^j\right) \right] \tag{2}$$

where, S^2 defines the count of grid cells, and B denotes the count of bounding boxes. C_i^j and $\hat{C}_i^j$ refer to the examined objectness value as well as truth objectness value, correspondingly. The location of the bounding box depends upon four predictions such as t_x, t_y, t_w, t_h, when considering (c_x, c_y) as the offset of grid cell, directed from top left corner of the image. The middle portion of the bounding boxes is referred to as offset from top left corner of the image using (b_x, b_y) and is determined as given below:

$$\begin{aligned} b_x &= \sigma(t_x) + c_x \\ b_y &= \sigma(t_y) + c_y \end{aligned} \tag{3}$$

where σ denotes a sigmoid function. The width and height of the detected bounding box are evaluated by function given below.

$$\begin{aligned} b_w &= p_w e^{t_w} \\ b_h &= p_h e^{t_h} \end{aligned} \tag{4}$$

whereby p_w, p_h define the width and height of a bounding box prior to what is accomplished by dimensional clustering. Ground truth box is composed of four attributes (g_x, g_y, g_w and g_h) that corresponds to the detected attributes such as b_x, b_y, t_w and t_h. According to Eqs. (3) and (4), the true values of $\hat{t}_x$, $\hat{t}_y, \hat{t}_w$, and $\hat{t}_h$ are determined using the Eq. (5):

$$\begin{aligned} \sigma(\hat{t}_x) &= g_x - c_x \\ \sigma\left(\hat{t}_y\right) &= g_y - c_y \\ \hat{t}_w &= \log(g_w / p_w) \\ \hat{t}_h &= \log(g_h / p_h) \end{aligned} \tag{5}$$

YOLOv3 model applies square error of coordinate examination as single portion of the loss function. It is illustrated as follows

$$E_2 = \sum_{i=0}^{S^2} \sum_{j=0}^{B} W_{ij}^{obj} \left[\left(\sigma\left(t_x\right)_i^j - \sigma\left(\hat{t}_x\right)_i^j \right)^2 + \left(\sigma\left(t_y\right)_i^j - \sigma\left(\hat{t}_y\right)_i^j \right)^2 \right]$$

$$+ \sum_{i=0}^{S^2} \sum_{j=0}^{B} W_{ij}^{obj} [((t_w)_i^j - (\hat{t}_w)_i^j)^2 + (t_h)_i^j - (\hat{t}_h)_i^j)^2] \quad (6)$$

3.2 Position Determination

Once YOLO v3 model identifies the objects in the frame, the next step is to determine the objects' position. For that, every frame is segregated into 3-row X 3-column grid cell as shown in Fig. 3. The whole image is divided into three positions such as top, center and bottom as row-wise and left, center and right as column-wise. After that, the central location of every bounding box is calculated based on box coordinates such as x, y, width(w) and height (h).

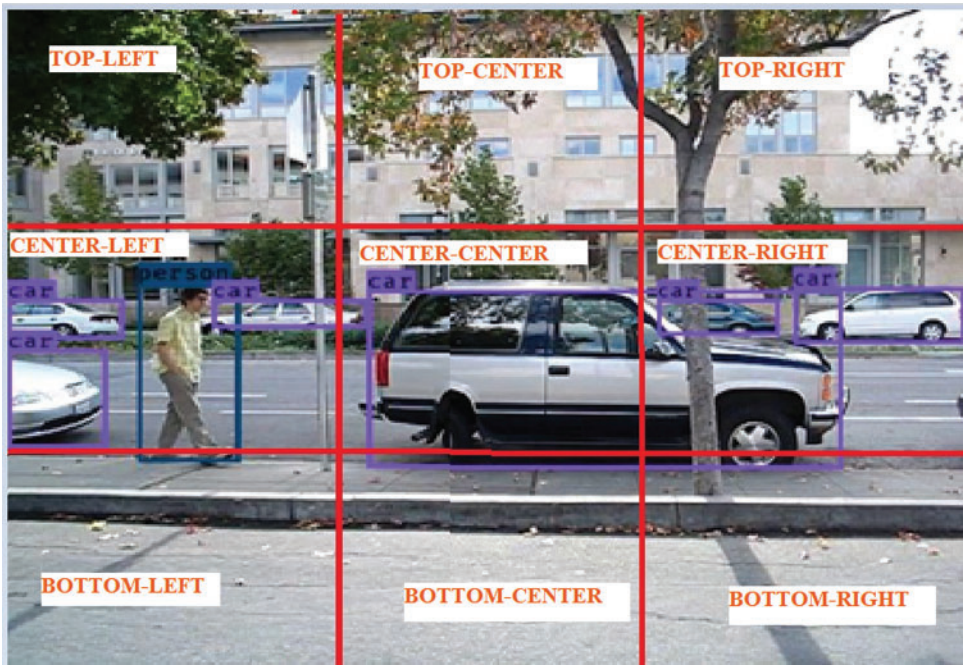

Figure 3: Object position determination

3.3 Audio Signal Generation

In audio signal generation stage, both detected object and its position in the frame are converted into an audio signal with the help of pyttsx Python library. Being a cross-platform text-to-speech conversion library, it is independent of the platform. Furthermore, a major benefit of this library is that it works offline as well. The python code snippet of pyttsx usage is given below.

```
Pseudocode Location Finding ()
center_x = round((2 * x + w) / 2)
center_y = round((2 * y + h) / 2)
if center_x <= W / 3:
W_pos = "left"
elif center_x <= (W / 3 * 2):
W_pos = "center"
else:
W_pos = "right"
if center_y <= H / 3:
H_pos = "top"
elif center_y <= (H / 3 * 2):
H_pos = "center"
else:
H_pos = "bottom"
```

```
pip3 install pyttsx3.
import pyttsx
engine = pyttsx.init()
engine.say("Your Message")
engine.runAndWait()
```

4 Experimental Results Analysis

This section discusses about the results of the detailed experimentation conducted upon EODM-AAS model when using four datasets namely, David3, Human4, Subway, and Hall Monitor [22]. The first dataset has a total of 252 frames while the second one has 667 frames and third and fourth ones have 176 and 300 frames respectively. Few details related to the dataset are given in Tab. 1 and some of the sample test images are shown in Fig. 4.

Table 1: Dataset descriptions

Dataset	Number of frames
David3	252
Human4	667
Subway	176
Hall monitor	300

Fig. 5 shows the results of the qualitative visualization analysis attained by the presented EODM-AAS model on the applied David3 dataset. Fig. 5a depicts the input image whereas the output image is shown in Fig. 5b. The figure infers that the proposed EODM-AAS model detected the objects as 'car' and 'person'.

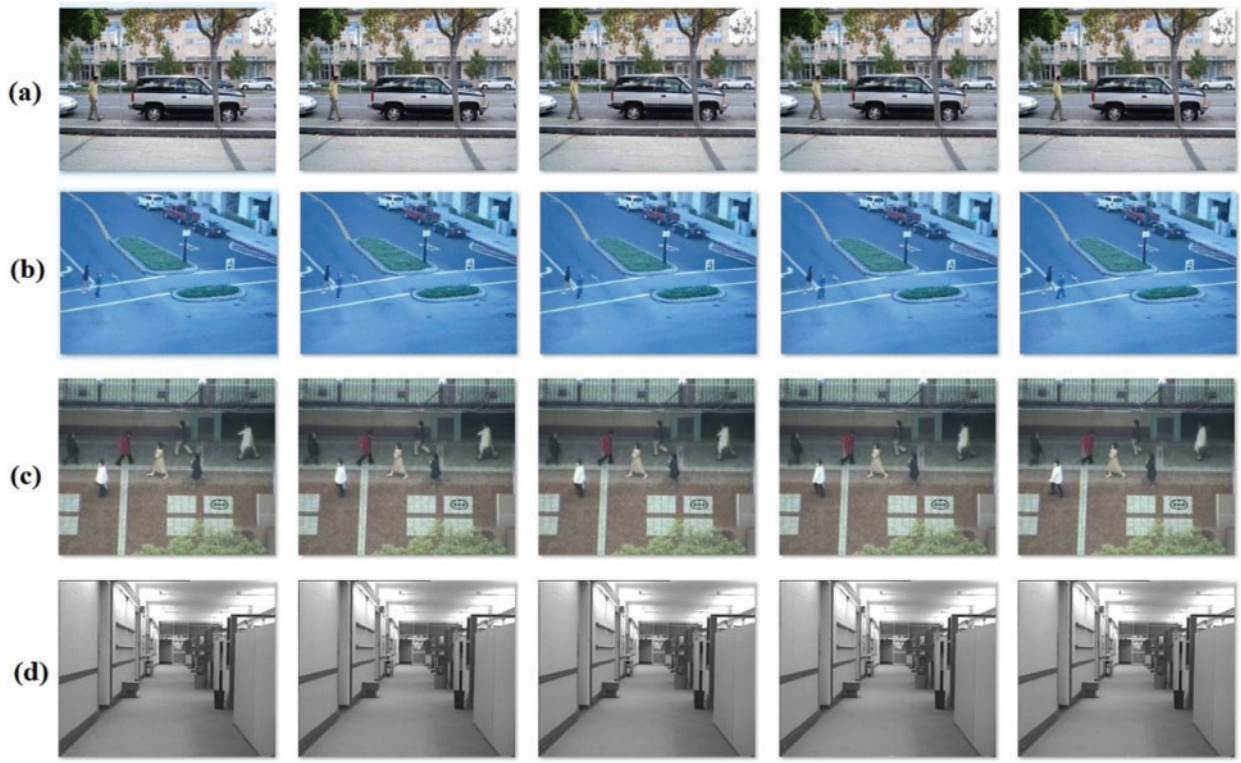

Figure 4: Sample frames of dataset a) David3 b) Human4 c) Subway d) Hall monitor

Fig. 6 shows the results of qualitative visualization analysis of the projected EODM-AAS model on the applied Subway dataset. Fig. 6a showcases the input image while the output image is illustrated in Fig. 6b. The figure infers that the EODM-AAS model identified the object, 'person' proficiently.

Fig. 7 portrays the results of qualitative visualization analysis of the proposed EODM-AAS model on the applied Hall monitor dataset. Fig. 7a depicts the input image while the output image is shown in Fig. 7b. The figure denotes that EODM-AAS model identified the object, 'suitcase' correctly.

Figure 5: Visualization analysis results of EODM-AAS model on David3 dataset (a) Original image, (b) Output image

(a) (b)

Figure 6: Visualization analysis results of EODM-AAS model on subway dataset (a) Original image, (b) Output image

(a) (b)

Figure 7: Visualization analysis results of EODM-AAS model on hall monitor dataset (a) Original image, (b) Output image

Fig. 8 clearly visualizes the results obtained by the proposed EODM-AAS model on David3 dataset. From the figure, it is clear that the presented EODM-AAS model detected the objects such as 'car' and 'person' along with position details such as "on your left center person"' and "on your center-center car".

Tab. 2 and Fig. 9 demonstrate the classification results accomplished by EODM-AAS model. The experimental values infer the effectual detection of the proposed EODM-AAS model on all the applied datasets under sensitivity and specificity metrics as given in the Eq. (7) and (8).

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{7}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{8}$$

For instance, on test David3 dataset, EODM-AAS model depicted effective detection results with sensitivity and specificity being 98.16% and 94.56% respectively.

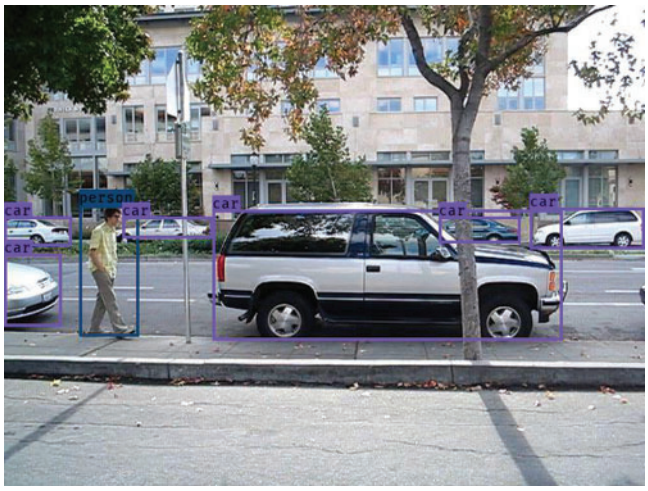

Figure 8: Audio assistive system: i) On your left center person found ii) On your center-center car found

Table 2: Analysis results of the proposed EODM-AAS model

Dataset	Sensitivity	Specificity	Average
David3	98.16	94.56	96.36
Human4	98.19	93.12	95.66
Subway	97.98	93.76	95.87
Hall monitor	98.34	93.54	95.94
Average	98.17	93.75	95.96

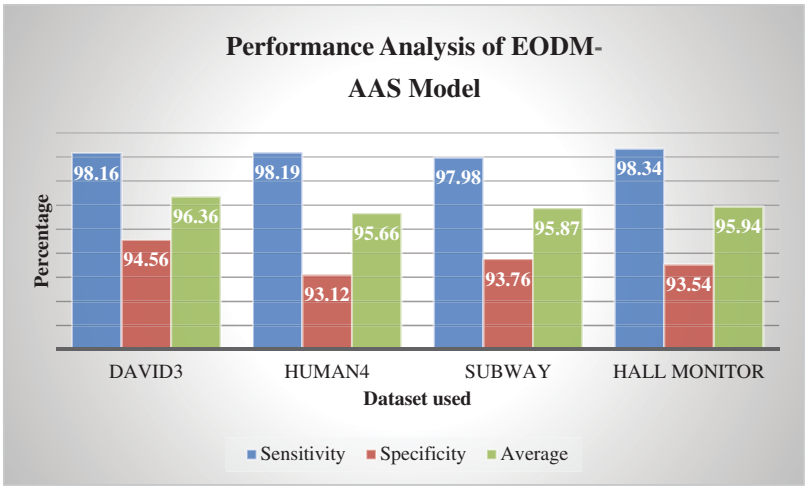

Figure 9: Analysis results of EODM-AAS model

Moreover, the Pre-trained ResNet model yielded reasonable results with a sensitivity of 89.54% due to skip connections among the layers. However, it is still a heavier model. In addition, Convolutional SVM Net and Fine-tuning SqueezeNet models demonstrated acceptable results i.e., sensitivity values such as 93.64% and 96.05% respectively while the prediction speed can be increased in spite of its small size.

Also, the Pre-trained VGG16 and Fusion using OWA models exhibited closer sensitivity values of 97.2% and 97.66% respectively. But VGG model, in spite its accuracy, took more time for training due to its heavy architecture. However, the presented EODM-AAS model displayed a better performance compared to all other methods and obtained a high sensitivity of 98.17%. It is highly suitable for real-time object detection, because of light weight structure and high prediction accuracy as shown in Tab. 3.

Table 3: Comparison of the proposed EODM-AAS model with existing techniques [23,24]

Methods	Sensitivity	Specificity	Accuracy
SURF + GPR	77.72	99.28	89.46
EDCS + GPR	70.00	90.12	80.66
MR random projection	77.18	91.41	84.90
Pre-trained GoogLeNet	83.63	96.86	90.85
Pre-trained ResNet	89.54	96.38	93.56
Convolutional SVM Net	93.64	92.17	93.50
Fine-tuning SqueezeNet	96.05	89.14	93.19
Pre-trained VGG16	97.20	86.70	92.55
Fusion using OWA	97.66	89.86	94.36
Deep-MLP	82.00	89.00	86.10
Proposed EODM-AAS	98.17	93.75	96.45

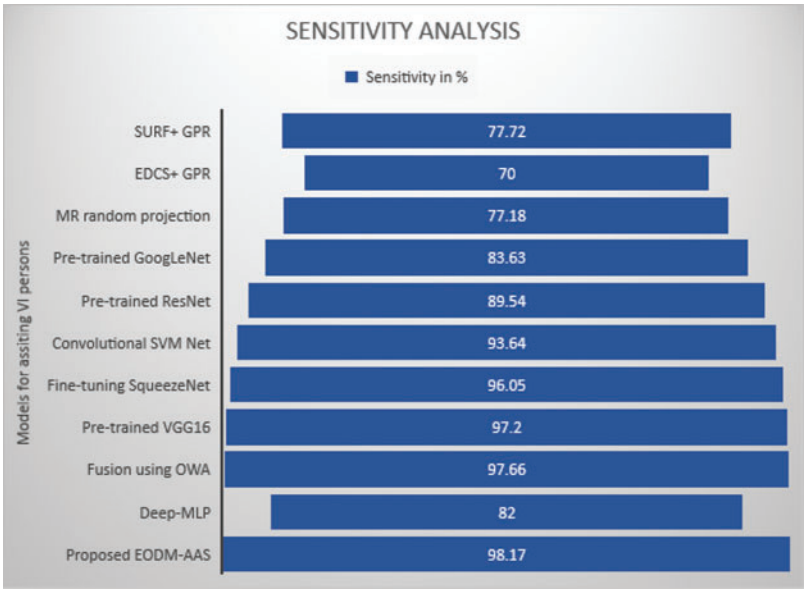

Figure 10: Comparative analysis of EODM-AAS model in terms of sensitivity

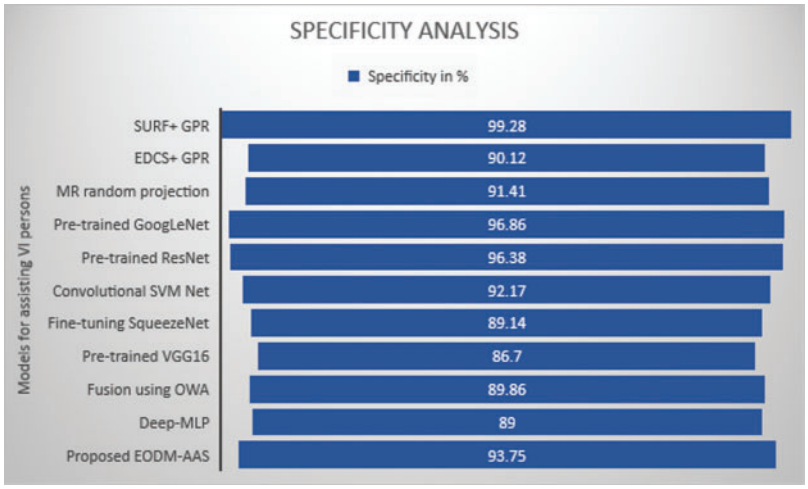

Figure 11: Comparative analysis of EODM-AAS model in terms of specificity

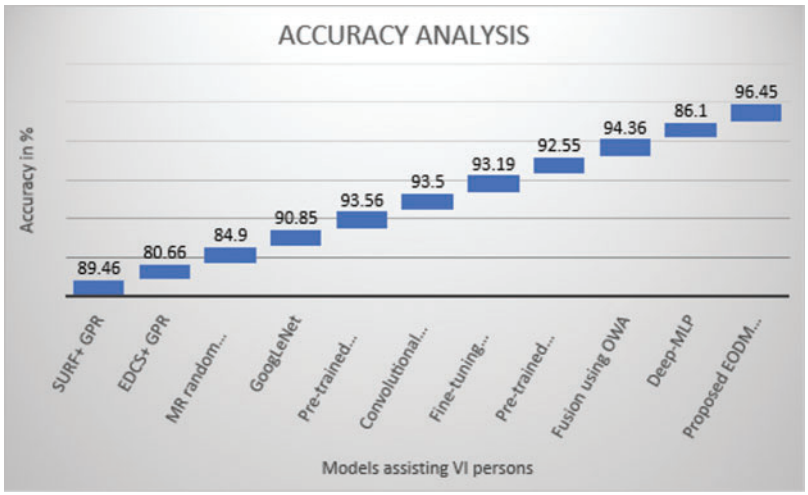

Figure 12: Comparative analysis of EODM-AAS model in terms of accuracy

Figs. 10–12 showcase the specificity, sensitivity and accuracy analyses results of the proposed EODM-AAS model against existing methods respectively. From the above mentioned results of the analysis, it is evident that the proposed EODM-AAS technique is an effective tool over other techniques, thanks to the incorporation of YOLOv3 model in it.

5 Conclusion

The current research article introduced an effective DL-based YOLO v3 model to perform object detection process so as to assist visually-impaired people. The aim of the research article is to derive a model that can provide a detailed description of the objects around visually-impaired people. The input video is initially transformed into a set of frames. Every frame undergoes YOLO v3-based object detection process to identify and classify multiple objects in the frame. Once the YOLO v3 model identifies the objects in the frame, the next step is to determine the position of the object in the frame such as, on your left, on your right, on your center, etc. In the last stage, the detected object and its position in the frame are converted into an audio signal using

pyttsx tool. In order to investigate the detection performance of the presented model, a detailed simulation analysis was performed on four datasets. The simulation outcomes inferred that the proposed method achieved better performance compared to that of the existing methods. As a part of future scope, the presented model can be implemented in real-time environment.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] M. B. Hoy, "Alexa, siri, cortana, and more: An introduction to voice assistants," *Medical Reference Services Quarterly*, vol. 37, no. 1, pp. 81–88, 2018.

[2] F. M. Hasanuzzaman, X. Yang and Y. Tian, "Robust and effective component-based banknote recognition for the blind," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 1021–1030, 2012.

[3] X. Chen and A. L. Yuille, "Detecting and reading text in natural scenes," in *Proc. of the 2004 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, 2004. CVPR 2004*, Washington, DC, USA, vol. 2, pp. 366–373, 2004.

[4] T. Jia, J. Tang, W. Lik, D. Lui and W. H. Li, "Plane-based detection of staircases using inverse depth," in *Proc. of Australasian Conf. on Robotics and Automation*, New Zealand, Victoria University of Wellington, pp. 1–10, 2012.

[5] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. of the 2001 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, CVPR 2001*, Kauai, HI, USA1, pp. 511–518, 2001.

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

[7] R. Lakshmanan and R. Senthilnathan, "Depth map based reactive planning to aid in navigation for visually challenged," in *2016 IEEE International Conf. on Engineering and Technology (ICETECH)*, Coimbatore, India, pp. 1229–1234, 2016.

[8] H. C. Huang, C. T. Hsieh and C. H. Yeh, "An indoor obstacle detection system using depth information and region growth," *Sensors*, vol. 15, no. 10, pp. 27116–27141, 2015.

[9] M. Poggi and S. Mattoccia, "A wearable mobility aid for the visually impaired based on embedded 3D vision and deep learning," in *2016 IEEE Symp. on Computers and Communication (ISCC)*, Messina, Italy, pp. 208–213, 2016.

[10] M. Vlaminck, L. H. Quang, H. V. Nam, H. Vu, P. Veelaert *et al.,* "Indoor assistance for visually impaired people using a RGB-D camera," in *2016 IEEE Southwest Symp. on Image Analysis and Interpretation (SSIAI)*, Santa Fe, NM, pp. 161–164, 2016.

[11] M. Vlaminck, L. Jovanov, P. V. Hese, B. Goossens, W. Philips *et al.,* "Obstacle detection for pedestrians with a visual impairment based on 3D imaging," in *2013 Int. Conf. on 3D Imaging*, Liege, Belgium, pp. 1–7, 2013.

[12] V. N. Hoang, T. H. Nguyen, T. L. Le, T. H. Tran, T. P. Vuong *et al.,* "Obstacle detection and warning system for visually impaired people based on electrode matrix and mobile kinect," *Vietnam Journal of Computer Science*, vol. 4, no. 2, pp. 71–83, 2017.

[13] D. Y. K. Sampath and G. D. S. P. Wimalarathne, "Obstacle classification through acoustic echolocation," in *2015 Int. Conf. on Estimation, Detection and Information Fusion (ICEDIF)*, Harbin, China, pp. 1–7, 2015.

[14] H. Takizawa, S. Yamaguchi, M. Aoyagi, N. Ezaki and S. Mizuno, "Kinect cane: Object recognition aids for the visually impaired," in *2013 6th Int. Conf. on Human System Interactions (HSI)*, Sopot, Poland, pp. 473–478, 2013.

[15] V. N. Mandhala, D. Bhattacharyya, B. Vamsi and N. T. Rao, "Object detection using machine learning for visually impaired people," *International Journal of Current Research and Review*, vol. 12, no. 20, pp. 157–167, 2020.

[16] S. Bhole and A. Dhok, "Deep learning based object detection and recognition framework for the visually-impaired," in *2020 Fourth Int. Conf. on Computing Methodologies and Communication (ICCMC)*, Erode, India, pp. 725–728, 2020.

[17] S. Vaidya, N. Shah, N. Shah and R. Shankarmani, "Real-time object detection for visually challenged people," in *2020 4th Int. Conf. on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, pp. 311–316, 2020.

[18] P. Zhang, Y. Zhong and X. Li, "SlimYOLOv3: Narrower, faster and better for real-time uav applications," in *2019 IEEE/CVF Int. Conf. on Computer Vision Workshop (ICCVW), Seoul, Korea (South)*, pp. 37–45, 2019.

[19] Q. Wang, S. Bi, M. Sun, Y. Wang, D. Wang *et al.,* "Deep learning approach to peripheral leukocyte recognition," *PLOS One*, vol. 14, no. 6, pp. e0218808, 2019.

[20] J. R. Darknet, "Open source neural networks in c," 2016. [Online]. Available: http://pjreddie.com/darknet/.

[21] L. Zhao and S. Li, "Object detection algorithm based on improved YOLOv3," *Electronics*, vol. 9, no. 3, pp. 537, 2020.

[22] "Dataset," 2021. [Online]. Available: http://cvlab.hanyang.ac.kr/trackerbenchmark/datasets.html. (Accessed on Feb 10, 2021)

[23] H. Alhichri, Y. Bazi and N. Alajlan, "Assisting the visually impaired in multi-object scene description using OWA-based fusion of CNN models," *Arabian Journal for Science and Engineering*, vol. 45, no. 12, pp. 10511–10527, 2020.

[24] S. K. Jarraya, W. S. Al-Shehri and M. S. Ali, "Deep multi-layer perceptron-based obstacle classification method from partial visual information: Application to the assistance of visually impaired people," *IEEE Access*, vol. 8, pp. 26612–26622, 2020.