Visually Impaired Person Assistance Based on Tensor FlowLite Technology

Nethravathi B[1], Srinivasa H P[2], Hithesh Kumar P[3], Amulya S[4], Bhoomika S[5], Banashree S Dalawai[6], Chakshu Manjunath[7]

Department of Information Science and Engineering, JSS Academy of Technical Education,
Bangalore-60, Karnataka, India[1, 4, 5, 6, 7]

Department of Computer Science and Engineering, T. John Institute of Technology, Bangalore-83, Karnataka, India[2]

Department of Computer Science and Engineering, JSS Academy of Technical Education, Bangalore-60, Karnataka, India[3]

***Abstract*—The most exciting thing about computer visualization is to detect a Real time object application system. This is abundantly used in many areas. With the more increase of development of deep learning such as self-driving cars, robots, safety tracking, and guiding visually impaired people, many algorithms have improved to find the relationship between video analysis and images analysis. Entire algorithms behave uniquely in the network architecture, and they have the same goal of detecting numerous objects in a composite image. It is very important to use our technology to train visually impaired people whenever they need them, as they are visually impaired and limit the movement of people in unknown places. This paper offers an application system that will identify all the possible day-to-day objects of our surroundings, and on the other side, it promotes speech feedback to the person about the sudden as well as far objects around them. This project was advanced using two different algorithms: Yolo and Yolo-v3, tested to the same criteria to measure its accuracy and performance. The SSD_MobileNet model is used in Yolo Tensor Flow and the Darknet model is used in Yolo_v3. Speech feedback: A Python library incorporated to convert statements to speech-to-speech. Both algorithms are analyzed using a web camera in a variety of circumstances to measure the correctness of the algorithm in every aspect.**

***Keywords*—Tensor flow; SSD; Yolo; Yolo_v3; gifts; deep learning**

I. Introduction

Human optical coordination is extremely precise and can handle multitasking even with unconscious notice. If you have a large amount of data, you need a more perfect system to appropriately detect and identify multiple objects at once. You can use better algorithms to train your computer to recognize multiple objects in an image with high precision. Object recognition is the utmost difficult application of computer visualization because this requires a comprehensive understanding of the image. This system was developed using a laptop, webcam, and Bluetooth headphones. TensorFlow is used to develop intelligent object detection algorithms. Models of IoT, embedded, and mobile devices are run by using tensor flow lite. It consists of low latency and low binary size, making it easy to design devices at the edge of your network. This advances potential, connectivity, and confidentiality. An unlimited length of speeches can be read by using google translate text to speech API which is a python library interface of google text to speech (gtts).

1) Deep-Learning: Deep learning comes under a branch of machine learning, a multi-tiered semantic network. These try to mimic the behavior of the human brain, permitting it to learn from the vast amount of data, even distant beyond its capabilities [1]. Single-layer semantic networks can provide limited estimates, but additional hidden layers help improve placement and accuracy.

2) Tensor flow lite: A version of TensorFlow is tensor flow lite. Tensor flow lite is used because it can run TensorFlow models on IoT, embedded, and mobile devices. TensorFlow Lite has low latency and a small binary size, which makes it easy to design devices on the network. It advances expectancy, confidentiality, and connectivity. Google Text-to-Speech (gtts), a Python library interface that uses the Google-Translate Text to Speech API. This can read language objects of infinite length. The system accurately detects objects up to 12-13 feet away.

3) Tensor flow framework: It is an open-source Machine learning framework used by many applications developed by Google using Google datasets. TensorFlow develops datasets for many applications, has a flexible architecture that is written in Python or C++, and has various available frameworks such as Keras, Theano,Caffe, and Torch. Among these, Tensor Flow provides robust machine learning and easy model building. TensorFlow Lite is used for embedded mobile IoT models. Tensor flow lite has two important components as listed below:

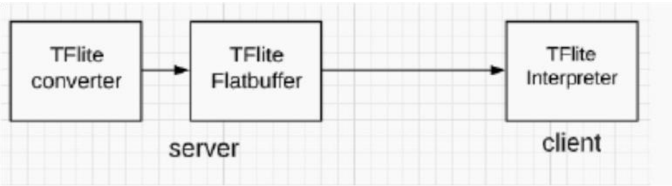


Fig. 1. Tensor Flow Components.

- Interpreter
- Converter

The interpreter is one which makes calculations based on the input data which have been given, and the Converter creates a flat buffer file (see Fig. 1). This file is given by the client's device. This uses the TensorFlow Lite Interpreter file internally. TensorFlow lite supports APIs in numerous languages such as Objective-C, C ++ and Python Java, Swift. Model optimization tools of tensor flow lite reduce the size of the model and improve performance by not losing its accuracy. TensorFlow Lite features an effective flat buffer model format optimized for small size and portability. Flat Buffer is an effective cross-platform serialization library for Rust, C #, Java, JavaScript C ++, and Typescript. Storage-efficient Flat Buffer trains algorithms to accelerate datasets, classify the data, and predict results precisely. When the input data is entered into the model, the weights are adjusted until the model fits properly. To avoid over fitting in the model cross-validation process is done.

4) Yolo V3: Yolo v3 You Look Only Once is the algorithm proposed by Redmond compared to the method used in the detection algorithms before YOLO V3. This is also useful for discovery bind boxes and class dividers that instantly capture the potential of a class (see Fig. 2). YOLO V3 takes a very different approach to object detection, providing up-to-date results that are superior to other real-time, large margin detection algorithms. Each of these N grids is responsible for content ingestion and retrieval. In parallel, these grids predict binding box bindings in terms of cell bindings [11]. This process handles both cell detection and detection from a single image. This saves a lot of computational effort because multiple cells predict the same by predicting different merge boxes. Duplicate forecasts are generated.

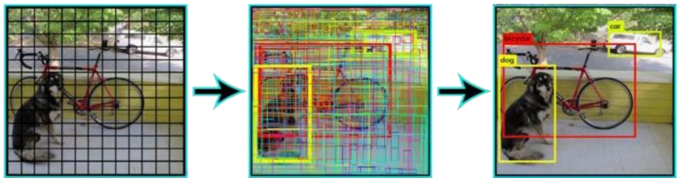


Fig. 2. Working of Yolo V3.

5) CNN: Convolution Neural Network is a concept of deep-learning algorithm that can capture an image and assign a value (readable weight and bias) to the various elements /elements in the image and be able to distinguish one from the other. The pre-processing required for a convolution network is very low when compared to other partition Algorithms. Earlier times, filters were used to be made by hand with adequate training convolution networks that can read these filters and symbols.

The organization of the paper starts with the literature survey, where total 12 papers were analyzed with merits and demits. Methodology is explained with system architecture and use case diagram, followed by results and discussion.

II. Literature Survey

Literature review aimed to look at various learning procedures and see if they could be applied to our use case. In [1], it has been proposed that humans use their eyes and mind to see and perceive the playing field and surrounding objects. In blind people, these talents are misplaced or impaired to varying degrees. Their eyes cannot perform the duty of sight. This paper is about designing and implementing a portable utility that helps people accurately determine their distance from them without seeing their environment. The proposed device is equipped with a CNN-based real-time data acquisition technology known as YOLO (You Look Only Once) and a separate digital camera on a Raspberry Pi board. The system additionally measures inference space and logically promises these statistics to the visually impaired. In [2], a geometry-based approach evaluates the entire 3D model. This way there is no need to isolate (or split) the object of interest. This system does not need to isolate the object of interest behind the surrounding. Compared to other methods, the proposed system is tested with real data collected from places such as kitchens or restaurants. The work proposed in this review meets the requirements of high accuracy, execution time, and compliance with the VIP criteria. An analytical data set has been published. In [3], this program does not need to isolate the object of interest behind the surrounding compared to other methods, the proposed system is tested with real data collected from places such as kitchens or restaurants. The work proposed in this review meets the requirements of high accuracy, execution time, and compliance with the VIP criteria. An analytical data set has been published. This system predicts his or her distance for approximate accuracy level. In [4], humans use their eyes and mind to see and perceive the playing field and surrounding objects. In blind people, these talents are misplaced or impaired to varying degrees. Their eyes cannot perform the duty of sight. This article is about designing and implementing a portable utility that helps people accurately determine their distance from them without seeing their environment. The proposed device is equipped with a CNN-based real-time data acquisition technology known as YOLO and a separate digital camera on a Raspberry Pi board. The gadget additionally measures inference space and logically promises these statistics to the visually impaired. In [5], in most cases, ongoing support is needed in almost all situations, especially in daily work. Other issues include difficulty recognizing people and detecting obstacles. To calculate this avoidance, this combines many of the available technologies and is integrated into one device with many features that you can use, for blind people. It takes less time to train the model. In [6], its functionality is easily compromised by creating complex training that incorporates many small image elements with more quality content object finders and scene separators, standard acquisition properties, and other modifications, improving acquisition function. Briefly review specific activities, including key discovery, facial recognition, and pedestrian detection. In [7], here the system gathers all information on motionless cameras to detect moving objects in digital videos. The system is developed based on the visual flow rate and usage as well as the combination. Object limitations are set using blob analysis; to obtain the audios, a central filter is used and the objects which are needed are removed using thresholding algorithms for natural operation. The accuracy is less than 65%. For this purpose, they suggest that 3D counterparts that are not specific to the most widely used 2D point detectors are insufficient, and suggest another option to support these points of interest, developing a data-based awareness algorithm with a spatial window temporarily. It supports only the static data set and performance is less than 65% K-Means. An algorithm with auto encoders is used. In [9], the experimental data, they achieved higher error rates of 1 and 5 of 37.5% and 17.0%, which is much better than the prior art. Some of these are surveyed by a completely linked layer with a maximum level of integration and a final 1000 Softmax. This used unsaturated neurons to speed up training and more efficient GPUs for convolutional performance. In the experimental data, they achieved higher error rates of 1 and 5 of 37.5% and 17.0%, which is much better than the prior art. A top-five test score of 15.3%, compared to 26.2% obtained by the top runners-up. In [10] this work, they introduced the Regional Proposal Network (RPN) which shares the features of complete image conversion and acquisition network that allows most cost-effective regional proposals. Regional Proposal Network is an entirely synchronized network that predicts object parameters and resistance scores in each area. RPN trained completely to produce high-quality regional proposals, which is used from Fast R-CNN for recognition. The RPN shows the integrated network where it is headed. Accuracy is high but large number of datasets need to be trained, so it consumes more time and space and computation power. The code has been made public. In [11], test results show the 7-bounds of conversion, the system can provide excellent real-time performance. This shows a 7 layer YOLO with an 11x11 grid cell that can accurately detect small people and vehicles. This provides accuracy and real-time performance and can only identify the person and car. In [12-15], use background-for-back classification methods to extract motion content and produce embedded videos. Then obtained directional vectors based on silhouette computations. They also used the flexible aspect of human movement to make smooth decisions over time and minimize errors in job perception. Our monocular method tolerates moderate viewing changes and can be used for both advanced and side views for most functions. This method does not do any small sample but instead is effective for all small windows. This improves the overall accuracy of the entire detector, it is not supported for other compound scoring functions. By using kernels, and by using this approach they demonstrated the great benefits of working on the three datasets which are publicly available. Surprisingly, they showed that a single solid HOG filter can exceed the modern partially crippled component. The detailed analysis of existing work is depicted in Table I.

TABLE I. THE DETAILED ANALYSIS OF EXISTING WORK

Reference paper No.	Special Features	Merits	Demerits	Conclusion and Future Work
[1]	The proposed MedGlasses system includes a couple of wearable glasses, (AI)- primarily based intelligence drug tablet box, a cellular tool app, and a cloud-primarily based facts control platform. The results shown that recognition accuracy of 95.1% can be achieved.	The improves the safety of medication for the visually impaired people	1.It is limited to only medicines. 2.It is not suitable for other objects.	This system efficiently moderates the problem of drug connections that are caused by taking wrong drugs, hence it decreases the cost of medical management and provides visually impaired patients with a safe medicinal facility.
[2]	This proposes a different framework for the detection of the objects in our daily activities and this system provides its related information which existing such as size, and security path for grasping on a surface. It contains pipelines which are the combinations of a series of point cloud representation and table plane detection, objects detection, and the full model estimation through a robust system.	It has met the requirements Of high precision, performing time, and correctness.	It requires an enormous amount of data to train. Accuracy is not up to the mark	In this framework, they considered the advantages of deep learning (e.g. RCNN, YOLO) that it can be an efficient method for exploration tasks, and the geometry-based approach Model This approach does not require separation or segmentation of the object of interest from the background.
[3]	The proposed system used YOLO methodology and a system with a single camera which is attached to a Raspberry Pi board also estimates the distance of the object and detects it and later sends this data to the blind person the method of audio feedback.	It gives accuracy above 98.8%.	We need large pieces of data to train the model.	This system concluded that it can detect the objects with its distance the accuracy of 98.8%
[4]	This application was developed using the features of the Python and OpenCV libraries, and this eventually ported to the Raspberry Pi3 Model and B+ platform. It has high Processing as well as includes the Extended identification of far distance objects accurately	It consumes less time to train the model.	1.It is expensive. 2.Its processing speed is very less.	This proposed system addresses various optimum distances and objects with less value of image source measurement
[5]	This is a voice assistance system to guide visually blind people in their daily activities. The device used is the various combination of technologies and consolidates all of these into a solitary multipurpose device that is used by the blind people	It consumes less time to train the model.	It is expensive. Its processing speed is very less.	This system discusses the different systems and algorithms for training models to achieve the accuracy of the detection.
[6]	This model behaves differently in network construction, learning strategies, optimization features, etc. it also	Improves the detection	specifies the object. It is not suitable to	It concludes several detection tasks including salient object detection, face

	provides an overview of the deep learning-based object detection environment.	performance further.	detect all the objects.	recognization, traffic signal recognization, etc.
[7]	Here the system gathers all information on motionless cameras to detect moving objects in digital videos. The system is developed based on the visual flow rate and usage as well as the combination. To obtain the audios, a central filter is used and the objects which are needed are removed using thresholding algorithms for natural operation.	Applied many algorithms to detect the object and trained using many algorithms	This is used to detect and track moving objects but accuracy is less than 65%.	It is concluded that the proposed system will accurately detect and track the correct moving objects in moving video.
[8]	In this work it has been developed an allowance of ideas of spatiotemporal case Due to this reason, it showed the direct 3D counterparts to frequently used 2D interest point detection	K-Means Algorithm with autoencoders	It supports only the static data set and performance is less than 65%.	Devised the recognition algorithm related to spatiotemporal data It resulted in recognization based on the variety of the dataset with a large amount of data
[9]	In the experimental data, they achieved higher error rates of 1 and 5 of 37.5% and 17.0%, which is much better than the prior art. Some of these are surveyed by a completely linked layer with a maximum level of integration and a final 1000 Softmax.	achieved in between the top 5 error rates of 37.5% and 17.0%	It needs more amount of data and it consumes more time to train the model.	Used the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the other system error rate.
[10]	used Region Proposal Network which shares full-image convolutional features with the detection network, so it enables nearly cost-free region proposals. The RPN is trained end- tend to create high-quality region proposals, which are used by Fast R-CNN for the detection of an object	Precisely identifies the object that is needed	Accuracy is a high but large number of datasets needed.	Simple, useful, and the code can be freely available Code is open-source with public availability.
[11]	This system used the methodology of YOLO and seven layers of convulsion network(CNN) The various size of grid cells are used for the processing of the object detection.	This system can provide excellent detection accuracy and real-time object detection	Trained only for less dataset.	The visual quality assessment using real-world images that shows the 7layer of YOLO with 11x11 grid cells correctly detects people and small vehicles
[12]	This system does not perform any of the sub-functions but it minimizes all of the MMOD which are used to progress any object identification method which is straightforward in the parameters, such as HOG or bag of visual word models.	MMOD optimizes the accuracy of the entire object detectors.	It does not support any complex functions, that uses the kernel	This approach shows significant performance improvement on three public datasets. Surprisingly it shows, that a single hard HOG filter can perform a deformable part model.

III. Methodology

A. Modules

1) Image Capturing: Proposed System consists of a camera, placed in the classroom to capture all the students. From these captured image frames, using the opencv and system will detect the student's face in the captured image using Haar cascade face detection technique. 2. Training phase: In the training phase we are applying CNN algorithm.3. Object recognition: i) Find object location: Object localization from the input video stream can be identified. Image classification from object detection, which goes through a ConvNet that results in a vector of features fed to a softmax to classify the object. Neural network have a few more output units that encompass a bounding box. In particular, we add four more numbers, which identify the x and y coordinates of the upper left corner and the height and width of the box (bx, by, bh, bw). ii) Find object movement: The Python's built in deque datatype to efficiently store the past N points the object has been detected and tracked at. Libraries imutils also used by collection of OpenCV and Python convenience functions. By checking the X, Y co-ordinate values and tracking them, the object movement can be detected. This process is further explained in Fig. 3, a use-case diagram.

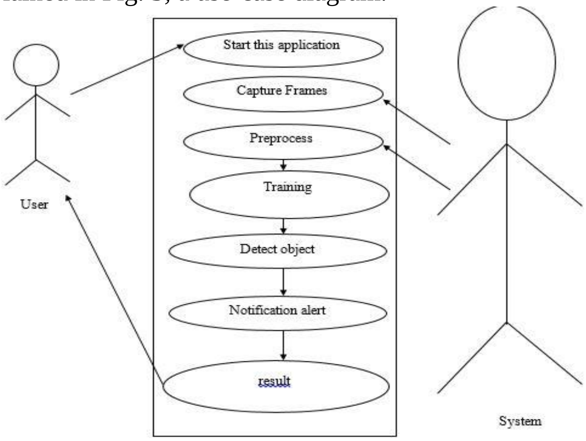


Fig. 3. Use-case Diagram.

2) System architecture: The architecture of Object detection system for blind people is described in Fig. 4. In this system input is real-time object through a camera, the captured image is stored for pre-processing, then the object is identified through the captured image, Then the calculation of distance of the objects from the person and generates the audio signal for the identified objects, transfers the audio signal in the form of audio feedback. When user starts the application, the system captures the frame, the object capture within the frame undergoes for pre-processing in the system. After preprocessing, the system is trained to detect object, the object within the frame is detected, if multiple object is found within the frame, then the distance between those objects are also detected. The notification alert of those detected object is sent as a result to the user. So, the user can be aware of any obstacles in their path. By using this application even the blind people live their lives independently.

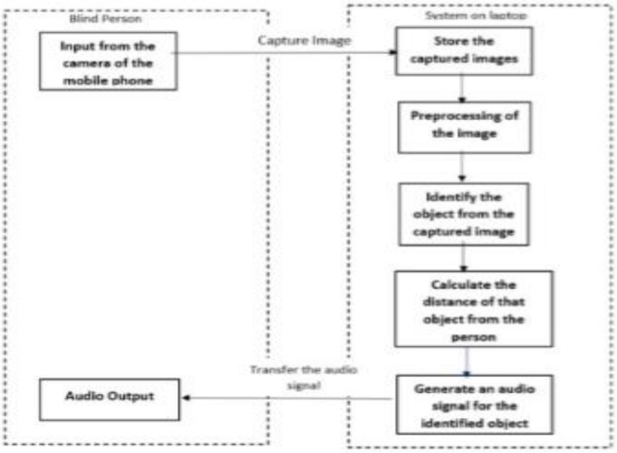
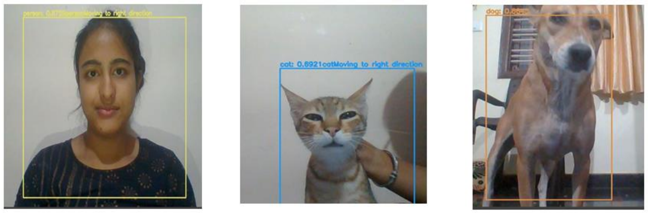

Blind Person
System on laptop
Input from the camera of the mobile phone
Capture Image
Store the captured images
Preprocessing of the image
Identify the object from the captured image
Calculate the distance of that object from the person
Generate an audio signal for the identified object
Transfer the audio signal
Audio Output

Fig. 4. System Architecture.

IV. Results and Discussions

In the above Fig. 5(a) shows that the person is captured in a frame the system detects the object within the frame then it sends audio feedback as person and also even in the form of text in the figure the person is moved to right so the system sends the audio and text feedback has person is move to right. The object is captured within the frame of 0.8769 pixel. Similarly in the Fig. 5(b) the object known as cat is captured within the frame of 0.6921 pixel and in the Fig. 5(c) the object dog is captured within the frame of 0.8645 pixel.

When multiple objects are detected within the frame the frame is clustered for each object then the system calculates the distance between the each object. In the above Fig. 6, objects such as person, cat and cell phone are captured within the frame. The system calculates the distance between each object and sends the result in the form of audio and text in console feedback.



(a) (b) (c)

Fig. 5. (a) Person (b) Cat (c) Dog.

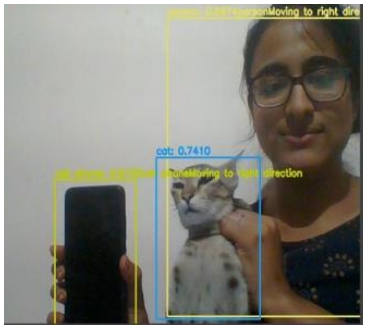


Fig. 6. Multiple Object Detection.

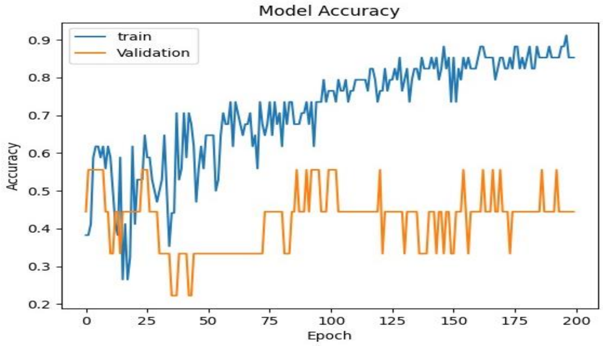


Fig. 7. The Graph of Model Accuracy.

The above Fig. 7 shows the graph of model accuracy in which x-axis is represented as epoch involved in training and y-axis is represented as accuracy. The graph shows the accuracy for 200 epoch in this dataset that has been divided as train and validation data. This graph shows the overall model accuracy of our implementation, that as the epoch increases accuracy increases.

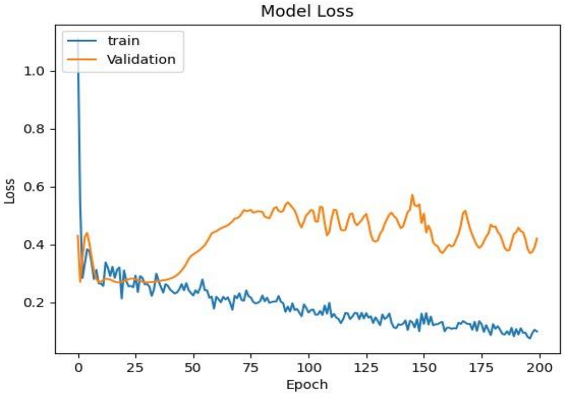


Fig. 8. The Graph of Model Loss.

The above Fig. 8 shows the graph of model loss in which x-axis is represented as epoch involved in training and y-axis is represented as loss. The graph shows the loss for 200 epoch in this dataset has been divided as train and validation data this graph shows the overall model loss of the implementation that as the epoch increases, loss reduced.

V. Conclusion

Developed a desktop application using python, the system automatically detects the objects and makes the voice alert to blind people. Object detection is used to search objects in the real world in images of the world, often found in blind scenes. It depends on the location and cameras that are used to detect the object. This architecture is proven and is reliable for blind people. Since the use of tensor flow lite, its response time is quick and even it requires small power, making it suitable for transferable applications. There are many things in the data set, and all of these are basic things that anyone needed in their daily activities. The biggest attraction of this work is that the output signal is in the voice listening format and blind people make use of headphones which makes it easy to guide visually impaired persons, so the overall accuracy rate is 98% which is better than previous implementations. Thus it allows users to become self-reliant without having to seek help.

References

[1] W. Chang, L. Chen, C. Hsu, J. Chen, T. Yang, and C. Lin, "MedGlasses: A Wearable Smart-Glasses-Based Drug Pill Recognition System Using Deep Learning for Visually Impaired Chronic Patients," in IEEE Access, vol. 8, pp. 17013-17024, 2020.

[2] V. Le, H. Vu and T. T. Nguyen, "A Frame-work assisting the Visually Impaired People: Common Object Detection and Pose Estimation in Surrounding Environment," 2018 5th NAFOSTED Conference on Information and Computer Science (NICS), Ho Chi Minh City, 2018, pp. 216-221s.

[3] S. Duman, A. Elewi and Z. Yetgin, "Design and Implementation of an Embedded Real-Time System for Guiding Visually Impaired Individuals," 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, Turkey, 2019, pp. 1-5.

[4] L. Tepelea, I. Bucio, C. Grava, I. Gavrilut and A. Gacsádi, "A Vision Module for Visually Impaired People by Using Raspberry PI Platform," 2019 15th International Conference on Engineering of Modern Electric Systems (EMES), Oradea, Romania, 2019, pp. 209-212.

[5] I. Joe Louis Paul, S. Sasirekha, S. Mohanavalli, C. Jayashree, P. MoohanaPriya and K. Monika, "Smart Eye for Visually Impaired-An aid to help the blind people," 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), Chennai, India, 2019, pp. 1-5.

[6] P. Vyavahare and S. Habeeb, "Assistant for Visually Impaired using Computer Vision," 2018 1st International Conference on Advanced Research in Engineering Sciences (ARES), Dubai, United Arab Emirates, 2018, pp. 1-7.

[7] O. Stephen, D. Mishra, and M. Sain, "Real-Time object detection and multilingual speech synthesis," 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, 2019, pp. 1-3.

[8] M. A. Khan Shishir, S. Rashid Fahim, F. M. Habib, and T. Farah, "Eye Assistant: Using a mobile application to help the visually impaired," 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), Dhaka, Bangladesh, 2019, pp. 1-4.

[9] C. Jayawardena, B. K. Balasuriya, N. P. Lokuhettiarachchi, and A. R. M. D. N. Ranasinghe, "Intelligent Platform for Visually Impaired Children for Learning Indoor and Outdoor Objects," TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), Kochi, India, 2019, pp. 2572-2577.

[10] S. Pehlivan, M. Unay, and A. Akan, "Designing an Obstacle Detection and Alerting System for Visually Impaired People on Side". Medical Technologies Congress (TIPTEKNO), 2019.

[11] M. Putra, Z. Yussof, K. Lim, and S. Salim "Convolutional neural network for the person and car detection using Yolo framework", Journal of Telecommunication, Electronic and Computer Engineering, Vol.10, PP 67-71.

[12] M. Singh, A. Basu and M. K. Mandal, "Human Activity Recognition Based on Silhouette Directionality," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 18, no. 9, pp. 1280-1292, Sept. 2008, doi: 10.1109/TCSVT.2008.928888.

[13] Ghaith Al-refai and Mohammed Al-refai, "Road Object Detection using Yolov3 and Kitti Dataset" International Journal of Advanced Computer Science and Applications(IJACSA), 11(8), 2020.

[14] Mohana and HV Ravish Aradhya, "Object Detection and Tracking using Deep Learning and Artificial Intelligence for Video Surveillance Applications" International Journal of Advanced Computer Science and Applications(IJACSA), 10(12), 2019.

[15] Hafsa Ouchra and Abdessamad Belangour, "Object Detection Approaches in Images: A Weighted Scoring Model based Comparative Study" International Journal of Advanced Computer Science and Applications(IJACSA), 12(8), 2021.