As per UGC guidelines an electronic bar code is provided to seure your paper

International Journal for Modern Trends in Science and Technology, 9(05): 649-656, 2023
Copyright © 2023International Journal for Modern Trends in Science and Technology
ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0905110
Available online at: http://www.ijmtst.com/vol9issue05.html

An Application Programming Interface (API) for Entity Identification using YOLO

Dr.B.Sunil Kumar, S.K.Dinesh, S.K.Shakeel, G.Abhiram Reddy, T.Pavan Kumar

Department of Comuter Sciecne Engineering, Narayana Engineering College, Nellore, Andhra Pradesh, India

To Cite this Article
Dr.B.Sunil Kumar, S.K.Dinesh, S.K.Shakeel, G.Abhiram Reddy and T.Pavan Kumar. An Application Programming Interface (API) for Entity Identification using YOLO. International Journal for Modern Trends in Science and Technology 2023, 9(05), pp. 649-656. https://doi.org/10.46501/IJMTST0905110

Article Info
Received: 21 April 2023; Accepted: 18 May 2023; Published: 22 May 2023.

ABSTRACT

***Abstract:** This paper presents the development of an Application Programming Interface (API) for entity identification using YOLO (You Only Look Once) model. The proposed API aims to provide a convenient and efficient solution for identifying various entities in images or video streams. The system leverages the YOLO model, which is known for its real-time object detection capabilities. The API consists of several modules, including data preprocessing, model inference, post-processing, and model evaluation. The software environment includes PyTorch, CUDA Toolkit, and TensorFlow. The API allows users to input images or video streams and receive the identified entities as output. The system has been implemented and tested, demonstrating promising results in terms of accuracy and speed. The proposed API has potential applications in various domains, including surveillance, autonomous vehicles, and image analysis.*

***KEYWORDS:** API, entity identification, YOLO, object detection, data preprocessing, model inference, post-processing, model evaluation, PyTorch, CUDA Toolkit, TensorFlow, real-time, accuracy, speed, surveillance, autonomous vehicles, image analysis.*

INTRODUCTION

A human being's best basic feature is often his ability of vision. The ability of being able to see things with our eyes is regarded as a gift and is the primary factor in our day-to-day activities. A major challenge in many of the visually impaired people is they are unable to be completely independent and are limited by their vision. The visually impaired people face trouble with such activities and object recognition will be an imperative feature they can depend on a regular basis. They usually face challenges in the identification of items and movement in the surroundings especially while walking on the streets. The majority of the humans having vision witnessed at 50 years of age [3]. Limited applications have been come in market to aid the visually impaired people. However, the absence of real continuous article acknowledgment and object recognition with voice output to the "visually challenged" people is still not widely modernized. Some of these applications are centered on sensing obstacles near the user and alerting them through alarms or beeping sounds with the implementation of IoT. For different reason, quantities of gadgets are required to hold by the users. For instance, smart sticks with obstacle detector, mobile phones, navigators, etc. are required for navigation assistance. These devices are expensive and could be a hindrance to the user at times. Some of the alternatives that are widely used today to provide assistance to visually impaired

people are: Tactile signs and Braille texts that are labeled on the top of the items for the identification [4, 9]. High-tech systems such as Radio Frequency Identification Devices (RFID), barcodes or talking labels that can be used to find objects in near distance [5,9]. Normally, complete visual impairment or low vision people face difficulties in outdoors. Voyaging or strolling down in a jam-packed road may represent an incredible trouble. In general, a visually impaired person takes support from their family members or sighted friends to maneuver through unfamiliar environments as well as to avoid the obstacles. For all these obstacles, they use canes. But, a cane cannot verify the kind of object ahead of them. From their experience, blind people usually identify an object, which leads to an injury or accident if an object is not quite the same as the one anticipated. They may also come across items such as stairs, dogs, parked cars, bicycles etc. that are hard to evade when walking along a walkway. The most challenging portion of such kinds of applications is the precise recognition of the objects and their positions appropriately. To recognize any object, those applications take an image as the input and convert that image into a gray scale formatted image. After formatting into gray scale photo, they run a top- down or bottom-up inspection and extract it into a histogram, which is based on pixel color depth. A huge number of researches have been carried out in the realm of real time object recognition using Deep Learning. In various ways, the Automatic detection of objects in image as well as in video process has been executed. Few are similar to our proposal method. All these researches can be distinguished in the detection of objects in any environment. In total our proposed methodology will detect 91 categories of objects including the indoor, outdoor and electronic devices. Internationally, 1 billion individuals dislike vision disability. It incorporates a wide range of hindrance like trachoma, glaucoma, uncorrected refractive file, waterfall, age-related macular degeneration, corneal haziness, diabetic retinopathy. In this way, for making their life somewhat simple we can furnish them with the vision of a PC. We can give vision to outwardly tested individuals by object discovery and acknowledgment and by illuminating them about their environmental elements utilizing some hear-able gadget like earphones and so on. The primary goal of this work is to introduce a far reaching and near investigation of the work that has been done in the field of item location for outwardly tested individuals. We will introduce here the relative investigation of the calculations that have been utilized in existing frameworks. In the year 1985, the first research paper on a guidance system was written and published by Jack Loomis who was a Professor of Psychology at the University of California. It was based on a GPS tracking system and started the research on steering system to handle visually impaired ones [6]. In the year 2009, a South Korean team designed a prototype of Unmanned Underwater Vehicle (UUV). On that UUV, a camera and a laser beam were operated to detect any obstacle in its path. The camera captured the images and the laser marked on it. Then it was converted into grayscale images. Using the most enlightened pixel and histogram, it could detect any object in front of the Unmanned Underwater Vehicle [7]. This popularized the method of obstacle detection. In the year 2018, Gnana Bharathy presented a context of object recognition and categorization with video analytics in cloud that precisely resulted in a high performance. Here, a cascade classifier is used to automate the video stream analysis process and laid the basis for the experiment of a large kind of video analytics algorithms [8]. The paper [9] presented the chief characteristics of software components devoted to help the visually impaired or blind ones. The main objective is to less use of separate strategy for object identification and activity of movement's detection. These components are implemented for Android OS due to the major use of smart phones in day to day life. Object recognition and the motion detection module are two trainable ANN based modules. This paper described the image processing algorithms to identify the objects and detect the motion of objects. In this system, notifications are provided to the users in the form of verbal messages. In this manuscript, our main aim is to help the visually impaired people for recognizing the objects by implementing machine learning for a product. It will be very helpful for the visually impaired people to feel less visionless. The document is planned as in section 2, discussing the related work, the proposed method for object recognition module in section 3, the experimental result and conclusion and future scope in section 4 and 5.

RELATED WORK

You simply look once (YOLO) is a lively article openness assessment. Overlooking how it isn't, as of now the most cautious article straightforwardness figuring, it is a stunning choice when unsurprising unquestionable demand is required without loss of a huge heap of exactness. YOLO uses a lone CNN sort out both collecting and restricting a thing using swaying boxes. YOLOv2 outfits reliable getting ready with high precision, yet it has higher containment missteps and lower survey response than other area based pointer checks. YOLOv2 is a revived assortment of YOLO that beats the lower study response and makes the accuracy with exuberant openness. The upgrades in YOLOv2 are rapidly inspected under: The totally related layers that are submitted for expecting the cutoff box were disposed of. One pooling layer was cleared to make the spatial yield of the framework be 13 × 13 rather than 7 × 7. The yield object classes were on a significant level irrelevant since classifiers expected that yield names were completely separated. YOLO had a soft ax ability to change above scores into probabilities as maximum. YOLOv3 uses a multi-name approach. Non-prohibitive yield inscriptions can show a score that is exceptional. As opposed to using the soft max work, YOLOv3 uses free critical classifiers to enroll the likeliness of the data having a spot with a specific cutting. YOLOv3 uses shaped with cross-entropy scene for each name rather than the mean settled slip in ascertaining the framework disappointment. Keeping up a key charming ways from the soft max work lessens the check complex nature. YOLO-9000 is a typical, speedier, and more grounded assortment of YOLO. Under, brief centers are shown concerning the matter of improving it, speedier, and more grounded. Basic standard classifier: The referencing make was changed on 448 × 448 pictures instead of encouraged with 224 × 224 pictures. This helped the arrangement with performing more basic standard. This fundamental standard social affair structure gave an advancement of taking everything into account. Convolution having boxes: In "YOLOv2", obtain boxes are gotten while conveying all totally related layers. Direct zone YOLO9000 forecast zone virtuosos fundamentally indistinct with the area of the cross segment unit, which pedals the floor reality to fall some place in the degree of nothing and one. It doesn't make gauges by using the balance to the spot of relationship of the weaving box.

Joint description and consent: For setting up a tremendous degree locater, two datasets were used. a standard referencing dataset with unimaginable groupings and an ID dataset. Single-Shot Multi Box Detector: A standard Convolution network set up each little advance in turn condense the part map size and develops the significance near the more crucial layers. More conspicuous open fields are guaranteed about by the huge layers, which makes satisfactorily theoretical depiction. More subtle responsive fields are guaranteed about by the thin coat. In like way, the framework can utilize this data to expect goliath things having further coats and to expect little articles using superficial layers. The central thought was that to utilize a lone advancement for velocity and to clear zone thought. It changes the power box as displayed by the check. The final hardly any coats are committed for humbler bouncing box need, which is in like route submitted for the check of different skipping boxes. The last theory is a blend of these longings. To all the basically certain hold SSD and its plan is demonstrated.

METHODOLOGY

Object discovery and Recognition

Object discovery and object recognition are related methods for object identification; however, they differ in execution. While both are extensively used for image and video processing, object detection is said to be a subset of object recognition. Object detection and recognition are widely used in the vast range of industries from individual security to productivity in the work environment. It is applied in various fields of computer vision as well as automated vehicle systems, machine inspection, surveillance, security, and image retrieval. Generally, the non-operating system devices can not recommend the text to speech conversion function. Thus, the most mainstream decision of smart phones with visually impaired users is either Android based phones or an iPhone.

Object discovery

Object discovery: it's the phenomena in discovering the instances of objects and as well as videos where the objects are located in the specified image or video. It highlights the recognized objects with bounding boxes and their position in the frame. Object detection is a technology which relates to image processing as well as to computer vision. It defines and detects various objects, for instance, animals, vehicles and persons from videos and digital images. Object detection has the ability to categorize several objects quickly within a video or digital image. Object detection has been around for decades, however is getting increasingly evident over a scope of industries now, like never before previously. Various methods have been used to implement the object detection system. However, this paper uses deep learning technique to provide high and better accuracy for object detection of varied object categories or classes. The Object Detection Process is shown in *Fig. 1* illustrates the detection of objects in step by step starting from process of Classification, Localization and object Detection.

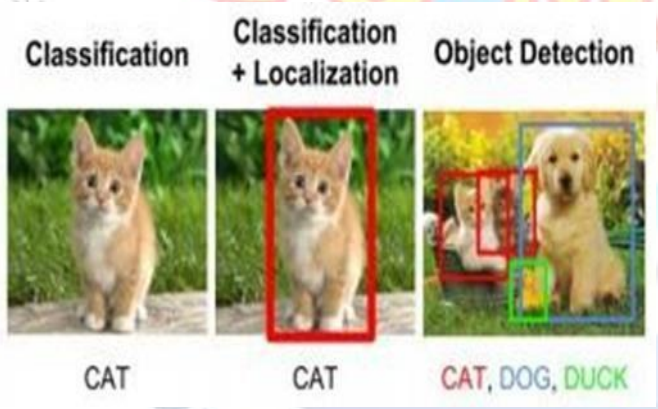

Fig. 1. Object Detection Process

YOLO Algorithm

Joseph Redmon first proposed the YOLO algorithm along with his team. He published his paper on YOLO in 2015 which was titled "You Only Look Once" Real-Time thing recognition and immediately it was a huge success. YOLO is a follows CNN. The algorithm "looks once" at the picture in the context that making predictions involves just one onward proliferation that passes in the course of the neural network. YOLO model is faster efficiently than any other method of detection of objects. YOLO's greatest advantage is their speed. This takes 45 frames per second. The model is built in such a succinct way as to familiarize its network with abstract description of items [11].

The flowchart i.e. working procedure of YOLO algorithm for real time object recognition is illustrated in a flowchart.

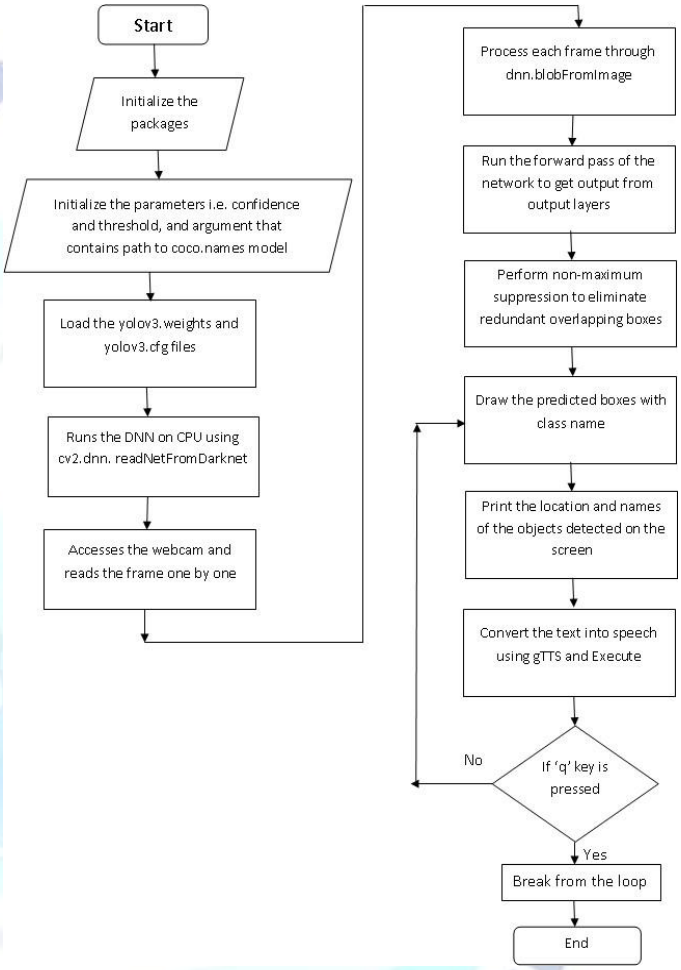

Fig. 5: Flowchart of Real Time Object Recognition using YOLO algorithm

The recent studies on implementing object recognition into the real world has been proved to be quite beneficial and efficacious. For recognizing the instances of an object or images, which belongs to an object class, is done by object detection methodologies. These methodologies frequently apply extracted features and learning algorithms for recognition process.

The main target of object detection is to localize a particular or more objects from videos or digital images. On the contrary, Object class recognition acts on categorization of objects into a specific category or a class. All objects has its own characteristics, which helps

to recognize the similar objects in other videos or images. Also makes them differentiate from other classes [17].
Object detection, detects and define objects, for instance, persons, animals, things, vehicles and so on. In order to achieve object recognition, we have to combine You Only Look Once (YOLO) architecture algorithm and COCO dataset that achieves a fast and efficient deep learning method.

Proposed Architecture

Our proposed system contains a camera, computer and headphones. The lens of the camera captures frame by frame through real time video processing, then the algorithm detects objects using Darknet module through which the model configuration file and model weight file has been identified. The program identifies them from the given You Only Look Once (YOLO) trained dataset named COCO and the output is given by the speakers or headphones, whichever is convenient and is available to the user at that moment. Here, completely we apply a deep learning-based methodology to lessen the issue of vision using object detection with a speech output in an end-to-end fashion is shown in the *Fig. 6*.

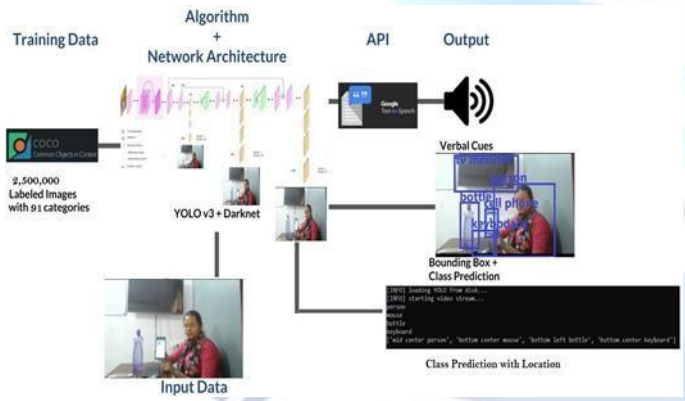

Fig. 6: Proposed architecture of object detection, object recognition and class prediction with location

YOLO is designed for complete picture processing and improves the efficiency of object detection in an unswerving way. Here, the frame detection is taken as a regression problem. The networked is focused on fresh photographs and YOLO uses the whole background during training and testing periods such that it encodes specific knowledge regarding groups and their appearances subtly. It utilizes highlights from the entire picture to all the while anticipate all bouncing boxes across all gatherings for a picture. The technique parts the information picture into an example of S x S. at the point when the focal point of a thing falls into a network cell, the matrix cell can recognize the point and decide the certainty scores for those containers. When there is nothing in that cell, at that point the certainty scores will be 0. These bouncing boxes are determined by the assessed probabilities [11].

Swift YOLO utilizes a neural network with less convolution layers, that is to say 9 out of 24 layers, and less filters in those layers. YOLO forecasts several bounding boxes per cell on the grid. After non-max suppression where each item is identified once, it then generates known items together with the bounding boxes. The algorithm outperforms other top detection algorithms [12].

In [13], Redmon and Farhadi carried out further work, YOLO900: Better, Faster, stronger. They made a range of enhancements to the YOLO discovery system includes the identification of over 9000 types of items by mutually pick up discovery and categorization. In paper [17], the identical researchers published an added paper on their development through YOLO advancing much further YOLOv3: An Incremental Improvement. This version has been used in the proposed methodology.
As mentioned earlier, the yolov3 makes detection at three scales, it down samples the input images by 32, 16, and 8. For the first 81 layers, the image is down sampled by the stride of 32 of the 81st layer. Regarding our image of 480*480, the resulting feature map would be 480/32, which is 15*15. Yolo4 has more models than yolo3.

Speech Output using GTTS API

The speech output module, that uses the speakers or headphone or a Bluetooth ear piece, has been added specifically to alert or simply let the user know about the surrounding objects around or in front of them. In addition to that, Speech Output would be more beneficial and faster at navigation especially when the user is walking in the streets. Hence, visually impaired individuals will take required preventive steps or stop for a short while till the objects ahead of them move out.

This module works by using Google's Text-To-Speech (GTTS) API, which is widely used in android

smartphones. It is a screen reader program that Google has built for Android operating system. With the help of several languages such as German, French, Tamil, Hindi, English and many more, GTTS powers functions to read out loud the text on the screen. All these languages support in GTTS API. It was released on 6 November 2013. The speech is delivered in either fast audio speed or slow speed. Nevertheless, as of the most recent update, the voice of the created audio cannot be altered. This API has the best advantage of sounding very natural [14].

Table 1: Implementation details about algorithm, dataset and parameter

Dataset	Parameter	Algorithm	Layer Size	Training Time
COCO	AveragePrecision	SVM	6	915.63 sec
COCO	AveragePrecision, Recognition Rate	CNN	8	961.12 sec
COCO	AveragePrecision	YOLO2	12	912.755 sec
COCO	AveragePrecision	YOLO3	14	906.45 sec

RESULTS AND DISCUSSION

4.1 Dataset

The experimental result uses COCO dataset. The dataset includes 91 familiar of objects, 82 of which have more than 5,000 named instances. In total, there are 2,500,000 instances. The dataset is as well considerably bigger than PASCAL VOC dataset and SUN datasets [15].
In [15] utilized a joint training on both the datasets i.e. ImageNet as well as COCO dataset to train YOLO9000. The outcome is a YOLO model, named as YOLO9000, which will predict the detected object categories without a labeled data [15].

The COCO dataset consists of 91 labels such as:
Kitchen and dining objects including spoons, knives, forks, cups, wine glasses etc. Animals including sheep, cows, horses, birds, dogs, cats, etc. Stop signs and fire hydrants, Cars and trucks, Airplanes, Bicycles and People.

The items in the dataset are classified using per instance segmentations to assist in accurate localization of items. There are presently two versions of COCO datasets for classified and segmented pictures. The photos in the dataset were collected from everyday scenes offering relevant information. In everyday scenario, many objects or things may be located within the same frame and every item ought to be recognized as a diverse entity and segmented correctly. The COCO dataset contains the naming and segmentation of the items present within the photos. We took this dataset and created our object finding and detection system for impaired people [17].

4.2. Results

With Open CV, we develop our deep learning-based real-time object detector that needs efficient access to our webcam/video stream and application of object detection to each frame. Initially, we have a tendency to capture a frame from the stream and resizing the frame. After that, with the help of DNN module, the frame is converted to a blob. For heavy lifting: set the blob into our Neural Network model as the input and feed the input through the network that gives us our detections. We've identified objects in the input frame at this point. At that point, we see confidence values and decide if we ought to draw a box and label around the thing.

We begin with iteration over our detections, considering that it is possible to identify multiple objects in one image. Always apply a check to the confidence associated with each detection. This confidence value is referred to as probability. When the confidence is high (i.e. above the threshold), the prediction will be displayed in the terminal; a colored displayed in the terminal; a colored bounding box as well as the prediction will be outlined on the image with text.

The model runs smoothly in 4 GB and above RAM operating system. A the end, the model will generate the outputs in various ways such as display in command prompt, in audio mp3 format and in a bounding box with label. At first, the model will generate the output in

command prompt which displays the output class label name with location as shown in *Fig. 7*.

```
[INFO] loading YOLO from disk...
[INFO] starting video stream...
person
mouse
bottle
keyboard
['mid center person', 'bottom center mouse', 'bottom left bottle', 'bottom center keyboard']
```

Fig. 7: Snippet of an output generated by our model on first detection

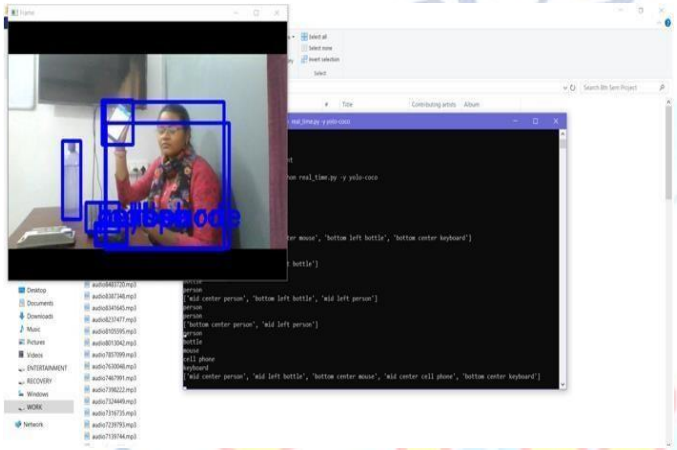

Fig. 8: Output series shown on both Frame and Command Prompt

The series of output displayed in both frame as well as in command prompt. The frame output is represented in bounding box with class label is shown in *Fig. 8. For Fig. 7 and 8*, we took screenshots randomly while running the program. *Fig. 7* was taken at the beginning of the webcam, *Fig. 8* was taken after the camera started adjusting, and finally *Fig. 9* is the correct output. The speech module respectively reads the names of the objects with their position displayed in the command prompt aloud.

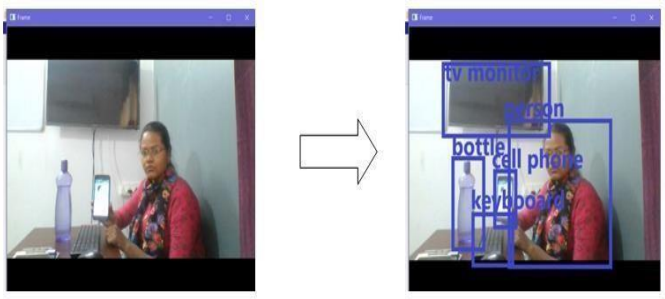

Fig. 9: Estimated Output with the given input

The speech output is faster with fast internet connection. The voice could be clearly heard with normal speakers as well as the headphones. The accuracy of the object detection module is 90%. Depends on the speed of computer, the final output may process around 16-18 Frames Per second (FPS) and the final output is a deep learning-based object detector.

Table 2: Results Comparisons of various algorithms

Algorithm	Model Size	Frames/Sec	Map(Precision)	Layer Size
CNN	57.3	6	44.67	6
SVM	234.86	11	43.84	8
YOLO2	202.3M	14	44.59	12
YOLO3	245.78 M	18	46.85	14

CONCLUSION

In a wide range of industries, the future of object detection has enormous opportunities. Video processing and object recognition algorithms are proposed based on available resources and dedicated to visually challenged users. People with visual impairments face major challenges when walking around and avoiding obstacles in their daily life. In day-to-day tasks, this program can support visually impaired and blind users. The experiment results of CNN, SVM, YOLO2 and YOLO3 were compared by means of precision and generating frames per sec. Yolo3 precision is 46.8% as better than other and by generated frames Yolo3 was outperformed with maximum 18 per sec. It will lessen the issue of movement and object recognition with a compact solution without the need to bring any additional devices dedicated to it.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] Mayur Rahul, Namita Tiwari, Rati Shukla, Devvrat Tyagi and Vikash Yadav (2022), A New Hybrid Approach for Efficient Emotion Recognition using Deep Learning. IJEER 10(1), 18-22. DOI: 10.37391/IJEER.100103.

[2] World Health Organization, Blindness and Vision Impairment, World Health Report, Geneva, 2019.

https://www.who.int/news-room/factsheets/detail/blindness-and-visual-impairment, October 2019.

[3] Matusiak, K., Skulimowski, P., Strurnillo, P., Object recognition in a mobile phone application for visually impaired users, International Conference on Human System Interactions, 2013.

[4] Dionisi, A., Sardini, E., Serpelloni, M., Wearable object detection system for the blind, IEEE International Instrumentation and Measurement Technology Conference Proceedings, pp. 1255-1258, 2012.

[5] Jack M. Loomis, Digital map and navigation system for the visually impaired, Department of Psychology, University of California, Santa Barbara, 1985.

[6] Muljowidodo, K., Mochammad A. Rasyid, SaptoAdi, N., Agus Budiyono, Vision based distance measurement system using single laser pointer design for under water vehicle, Indian Journal of Marine science, Vol. 38, No. 3 pp. 324-331, 2009.

[7] Neha Bari, Nilesh Kamble, Parnavi Tamhankar, Android based object recognition and motion detection to aid visually impaired, International Journal of Advances in Computer Science and Technology, Vol. 3, No.10, pp. 462-466, 2014.

[8] Jason Yip, Object Detection with Voice Feedback YOLO v3 + gTTS, https://towardsdatascience.com/object-detection-with-voice-feedbackyolo-v3-gtts-6ec732dca91.

[9] Samkit Shah, CNN based Auto-Assistance System as a Boon for Directing Visually Impaired Person, 3rd International Conference on Trends in Electronics and Informatics, 2019.

[10] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, You Only Look Once: Unified, Real-Time Object Detection, IEEE Conference on Computer Vision and Pattern Recognition, pp. 779-788, 2016.

[11] ODSC-Open Data Science, Overview of the YOLO Object Detection Algorithm, https://www.medium.com/@ODSC/overview-of-the-yoloobject-detection-algorithm-7b52a745d3e0.

[12] Joseph Redmon, Ali Farhadi, YOLO9000: Better, Faster, Stronger, IEEE Conference on Computer Vision and Pattern Recognition, pp. 65176525, 2017.

[13] gTTS, https://pypi.org/project/gTTS/.

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, Lawrence Zitnick, C., Microsoft COCO: Common objects in Context, European conference on Computer Vision, pp. 740-755, 2014.

[15] Amikelive Technology Blog, What object categories/labels are in COCO dataset?, https://tech.amikelive.com/node-718/what-object-categories-labels-are-in-coco-dataset/, 2018.

[16] Abdul Vahab, Maruti S Naik, Prasanna G Raikar, Prasad S R, Applications of Object Detection System, International Research Journal of Engineering and Technology, Vol. 6, No. 4, pp. 4186-4192, 2019.

[17] Christian Szegedy, Alexander Toshev, Dumitru Erhan, Deep Neural Networks for object detection, Advances in Neural Information Processing Systems 26, 2013.