IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

OBJECT DETECTION WITH AUDIO FEEDBACK

[1] D. Venkat Sampath Siva Ram Prakash, [2] D. Lakshman Sai, [3] M. Saraswathi.

[1] Student, [2] Student, [3]Associate professor

Department of Computer Science and Engineering,

SCSVMV, Kanchipuram, India

ABSTRACT:

Since edge computing allows for the implementation of more image-processing tasks on the decentralized network layer of the surveillance system, smart video surveillance at the edge has recently become popular in the development of security apps. As a result, when the camera network system is expanded to a greater extent, many security applications, including behavior identification and prediction, employee safety, perimeter intrusion detection, and destruction limitation, can minimize their latency or even process in real time. One of the difficult uses of computer vision is object recognition, which has been extensively used in a variety of fields, such as autonomous vehicles, robotics, surveillance, guiding visually impaired people, etc. Numerous algorithms were advancing the connection between video analysis and image comprehension as deep learning advanced quickly. With various network architectures, each of these algorithms accomplishes the same task of finding multiple objects within a complicated image. It is crucial to use our technologies and train them to assist blind people whenever they need it because the lack of visual impairment restricts movement in an unfamiliar environment. Technically, one of the most important steps in the execution of these apps is human detection. Deep learning techniques have been extensively used on edge devices to detect human objects due to their high detection rates. However, it is difficult to use these techniques for real-time applications on edge devices with limited resources because of their high computation costs. from the You Only Look Once (YOLO).

Keywords:

Image edge detection, Video surveillance, Real-time system, Safety, Security, Task Analysis, Testing

1 INTRODUCTION

Humans almost by birth are trained by their parents to categorize various persons as children self is one person. A person’s visual System is very accurate and precise that can handle multi-tasks even with a less conscious mind. When there are large data then we need a more accurate system to correctly recognize and localize multiple persons simultaneously. Here machine comes into existence, we can train our computers with the help of better algorithms to detect multiple persons within the image with high accuracy and preciseness. Person Detection is the most challenging application of computer vision as it requires a complete understanding of images. In another word, the person tracker tries to find the presence of persons within multiple frames and assigns labels to each person. There might be many problems faced by the tracker in terms of complex images, Loss of information, n, and transformation other the 3D world2D to 2D images. To achieve good accurate person detection, we should not only focus on classifying persons but also on locating the positions of different persons that may vary from image to image. It is very important to develop the most effective real-time persons tracking algorithm which is a challenging task. Deep learning since 2012 is working on these kinds of problems and has revolutionized the domain of computer vision. This paper aims to test the performance of both algorithms in different situations in real-time using a webcam and is made primarily for visually impaired people. Blind peoples have to rely on someone who can guide them or on their physical touch which is sometimes very risky also. Daily navigation of blind people in unfamiliar environments could be a frightening task without the help of some intelligent systems. The key concern behind this contribution is to investigate the possibility of expanding the counts of persons at one go to expand the support given to visually impaired people. Some common limitations of the previous techniques are less accuracy, complexity in the scene, lighting, etc. To overcome all those challenges two algorithms are analyzed on all possible grounds and from every perspective to achieve good accuracy.

A definition is both important and difficult because the everyday word "Object detection" is a notoriously fluid term in meaning. Object detection is one of the most challenging concepts to define in surveillance. There are different definitions of object detection in the scientific literature. In everyday surveillance, Object detection is any relatively brief conscious experience characterized by intense mental activity and a high degree of pleasure or displeasure. Scientific discourse has drifted to other meanings and there is no consensus on a definition. Object detection is often entwined with temperament, situations, traffic, and disposition. In surveillance, Object detection is frequently defined as a complex state of problem that results in physical changes. These changes influence the thought and behavior of objects. According to other theories, object detection is not a causal force but simply synchronized components.

As human beings' speech is amongst the most natural ways to express ourselves. We depend so much on it that we recognize its importance when resorting to other communication forms like emails and text messages where we often use emojis to express the emotions associated with them. As object detection plays a vital role in communication with the absence of vision impairment restraining the movement of the persons the detection and analysis of the same are of vital importance in today's digital world of remote communication. Object detection is a challenging task because emotions are subjective. There is no common consensus on how to measure or categorize them. We define an SER system as a collection of methodologies that process and classify speech signals to detect emotions embedded in them. Such a system can be used in various application areas like interactive voice-based-assistant or caller-agent conversation analysis. In this study, we attempt to detect underlying emotions in a recorded speech by analyzing the acoustic features of the audio data of recordings.

2 LITERATURE SURVEY

[1] *Real-time implementation of person tracking through webcam***.** Real-time person detection and tracking is an important task in various computer vision applications. For robust persons tracking the factors like person shape variation, partial and full occlusion, and scene illumination variation will create significant problems. We introduce a person detection and tracking approach that combines Prewitt edge d detection and Kalman filter. The target persons' representation and the location prediction are the two major aspects of person tracking this can be achieved by using these algorithms. Here real-time person tracking is developed through a webcam. Experiments show that our tracking algorithm can track moving persons efficiently under person deformation, and occlusion and can track multiple persons.

[2] Person Detection with Deep Learning: A Review. Due to persons detection's close relationship with video analysis and image understanding, it has attracted much research attention in recent years. Traditional person detection methods are built on handcrafted features and shallow trainable architectures. Their performance easily stagnates by constructing complex ensembles which combine multiple low-level image features with high-level context from person detectors and scene classifiers. With the rapid development of deep learning, more powerful tools, which can learn semantic, high-level, and deeper features, are introduced to address the problems existing in traditional architectures. These models behave differently in network architecture, training strategy, optimization function, etc. In this paper, we provide a review of deep learning-based person detection frameworks. Our review begins with a brief introduction to the history of deep learning and its representative tool, namely the Convolutional Neural Network (CNN). Then we focus on typical generic person detection architectures along with some modifications and useful tricks to improve detection performance further. As distinct specific detection tasks exhibit different characteristics, we also briefly survey several specific tasks, including salient person detection, face detection, and pedestrian detection. Experimental analyses are also provided to compare various methods and draw some meaningful conclusions. Finally, several promising directions and tasks are provided to guide future work in both person detection and relevant neural network-based learning systems.

[3] Histograms of oriented gradients for human detection**.** We study the question of feature sets for robust visual person recognition; adopting linear. SVM-based human detection as a test case. After reviewing existing edge and gradient-based descriptors, we show experimentally that grids of histograms of oriented gradient (HOG) descriptors significantly outperform existing feature sets for human detection. We study the influence of each stage of the computation on performance, concluding that fine-scale gradients, fine orientation binning, relatively coarse spatial binning, and high-quality local contrast normalization in overlapping descriptor blocks are all important for good results. The new approach gives near-perfect separation from the original MIT pedestrian database, so we introduce a more challenging dataset containing over 1800 annotated human images with a large range of pose variations and backgrounds.

[4] Region-Based Convolutional Networks for Accurate Persons Detection and Segmentation. Person detection performance, as measured on the canonical PASCAL VOC Challenge datasets, plateaued in the final years of the competition. The best-performing methods were complex ensemble systems that typically combined multiple low-level image features with high-level context. In this paper, we propose a simple and scalable detection algorithm that improves mean average precision (MAP) by more than 50 percent relative to the previous best result on VOC 2012-achieving a MAP of 62.4 percent. Our approach combines two ideas: (1) one can apply high-capacity convolutional networks (CNNs) to bottom-up region proposals to localize and segment persons and (2) when labeled training data are scarce, supervised pre-training for an auxiliary task, followed by domain-specific fine-tuning, boosts performance significantly. Since we combine region proposals with CNNs, we call the resulting model an R-CNN or Region-based Convolutional Network

3 Problem statement

Object recognition is one of the challenging applications of computer vision, which has been widely applied in many areas e. g. autonomous cars, Robotics, Security Tracking, Guiding Visual tracking of paired People, etc. With the rapid development of deep learning, many algorithms were improving the relationship between video analysis and image understanding. All these algorithms work differently with their network architecture but with the same aim of detecting multiple objects within a complex image. The absence of vision in the absence of restraint is the movement of the person in an unfamiliar place and hence it is very essential to take help from our technologies and trained them to guide blind people whenever they need it. Technically, human detection is a key step in the implementation of these applications. With the advantage of high detection rates, deep learning methods have been widely employed on edge devices to detect human objects.

4 Proposed methods

Algorithm:

Step 1: start.

Step 2: Give the image to the model.

Step 3: write the program in python.

Step 4: use the image information dataset.

Step 5: Use the CNN technique for programming.

Step 6: Get the output.

Step 7: Gives the detected image.

Step 8: Get the details about the object in audio form.

We propose a system that will detect every possible day-to-day multiple people on the other hand prompt a voice to alert the person about the near as well as farthest persons around them. To get audio we will use web speech API to produce speech

- **CNN:**

In deep learning, a convolutional neural network (CNN, or Conv-Net) is a class of deep neural networks, most applied to analyzing visual imagery. They are also known as shift invariant or space invariant artificial neural networks (SIANN), based on the shared weight architecture of the convolution kernels that shift over input features and provide translation equivariant responses.

Censure regularized versions of multilayer perceptron. Multilayer perceptions usually mean fully connected networks, that is, each neuron in one layer is connected to all neurons in the next layer. The "full connectivity" of these networks makes them prone to overfitting data. Typical ways of regularization, or preventing overfitting, include penalizing parameters during training (such as weight decay) or trimming connectivity (skipped connections, dropout, etc.) CNN takes a different approach towards regularization: they take advantage of the hierarchical pattern in data and assemble patterns of increasing complexity using smaller and simpler patterns embossed in their filters. Therefore, on a scale of connectivity and complexity, CNNs are on the lower extremity.

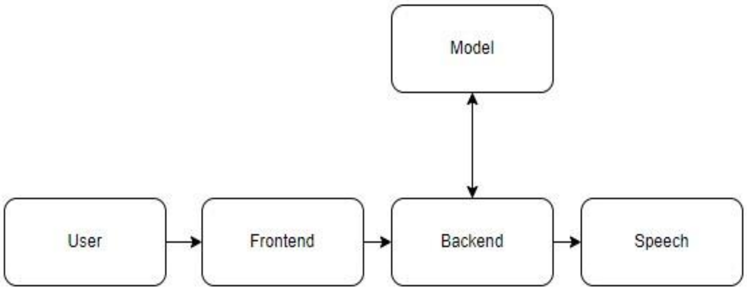


Fig 01: System Architecture

The name “convolutional neural network” indicates that the network employs a mathematical operation called convolution. Convolutional networks are specialized types of neural networks that use convolution in place of general matrix multiplication in at least one of their layers.

Yolo is a part of person detection, Person detection is a computer technology related to computer vision and image processing that deals with detecting instances of semantic persons of a certain class (such as humans, buildings, or cars) in digital images and videos. Well-researched domains of person detection include face detection and pedestrian detection. Person detection has applications in many areas of computer vision, including image retrieval and video surveillance.

Every person’s class has its special features that help in classifying the class – for example, all circles are round. Person class detection uses these special features. For example, when looking for circles, persons that are at a particular distance from a point (i.e., the center) are sought. Similarly, when looking for squares, persons that are perpendicular at corners and have equal side lengths are needed. A similar approach is used for face identification where eyes, nose, and lips can be found and features like skin color and distance between eyes can be found.

MODULES:

- **User**:

Data gathering:

The Dataset used was English- the French dataset which contains small sentences in English and French. We took 20000 English sentences and discarded the French translation and encrypted the English sentences word by word.

Pre-processing:

All encrypted sentences were represented in One Hot representation of the sentences and then turned in word embedding with 250 features

Model Building

We used two 2 LSTM layers one acting as an encoder and another as a decoder. We passed the encrypted data through the encoder and save the weights of the encoder and passed along with target data to the decoder which tries to predict the decrypted version of the encrypted data.

View Results

Users view the generated results from the model.

- **System**

Model Checking

The system checks model accuracy and we got 74.83% after 200 epochs and a 0.00001 learning rate

Generate Results

The system takes the input text from the users, encrypts it, and produces the decrypted text

5 Experimental Setup

The experiments were carried out using Python software.

This Patient Information System is based on Tkinter. The project has a graphical user interface provided by the Python programming language and SQLite. It provides a GUI where the user can detect an object with audio feedback.

Required specifications:

> PyCharm
> Python
> Html
> CSS
> JS
> Django

Testing:

Table no 01: Test Cases

Input	Output	Result
Input features	Tested for different features given by the user on the model.	Success
Images	We were able to detect persons using yolov3 and use web speech API to generate speech.	Success

There are four main testing phases in this project quality management:

- Unit Testing:

  We test individual units of code to make sure they are working as intended and functioning correctly.

- Integration Testing:

  We test how different individual units of code work together.

- System Testing:

  It is the third phase and tests the entire system to make sure it is functioning as intended

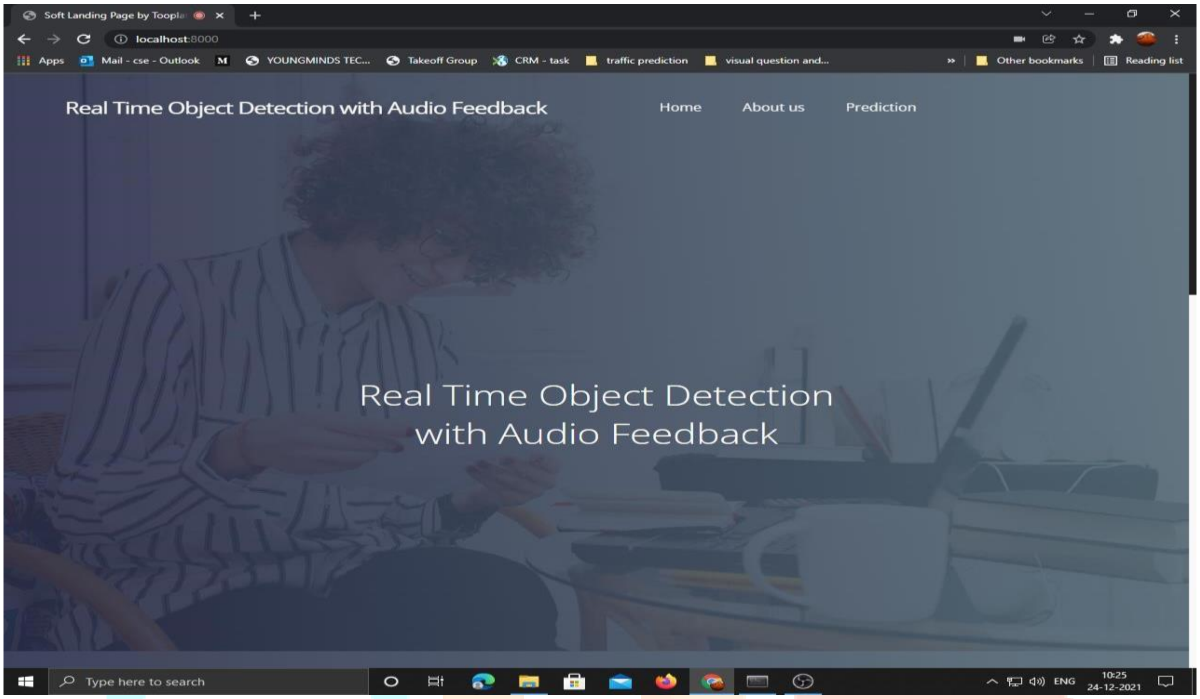
- Acceptance testing:

  Lastly, we have done this to make sure that the system meets all the requirements for the blind person.

6 Results

After we run the code, the output will be displayed in below fig:01-05.

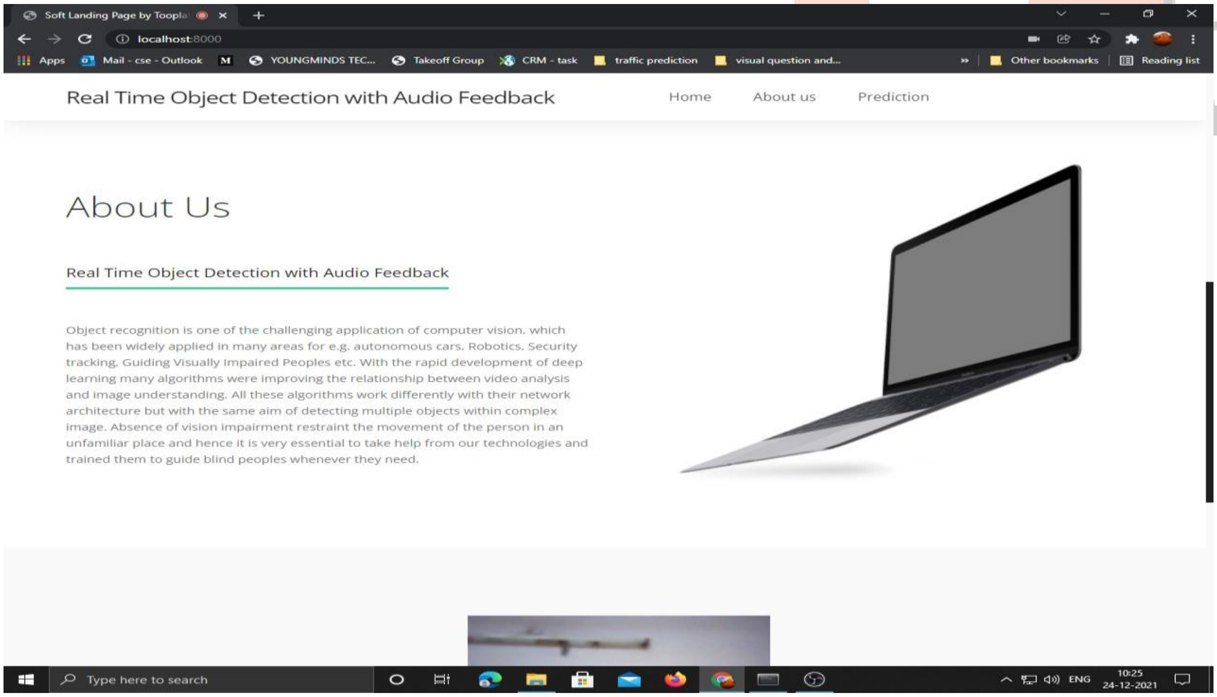
The output will be the detection of objects like persons, cup, bottle etc,. with bounding boxes and object name on top.



Real Time Object Detection with Audio Feedback

Real Time Object Detection with Audio Feedback

Fig 01: home page.



Real Time Object Detection with Audio Feedback

About Us

Real Time Object Detection with Audio Feedback

Object recognition is one of the challenging application of computer vision, which has been widely applied in many areas for e.g. autonomous cars, Robotics, Security tracking, Guiding Visually Impaired Peoples etc. With the rapid development of deep learning many algorithms were improving the relationship between video analysis and image understanding. All these algorithms work differently with their network architecture but with the same aim of detecting multiple objects within complex image. Absence of vision impairment restraint the movement of the person in an unfamiliar place and hence it is very essential to take help from our technologies and trained them to guide blind peoples whenever they need.

Fig 02: About Project

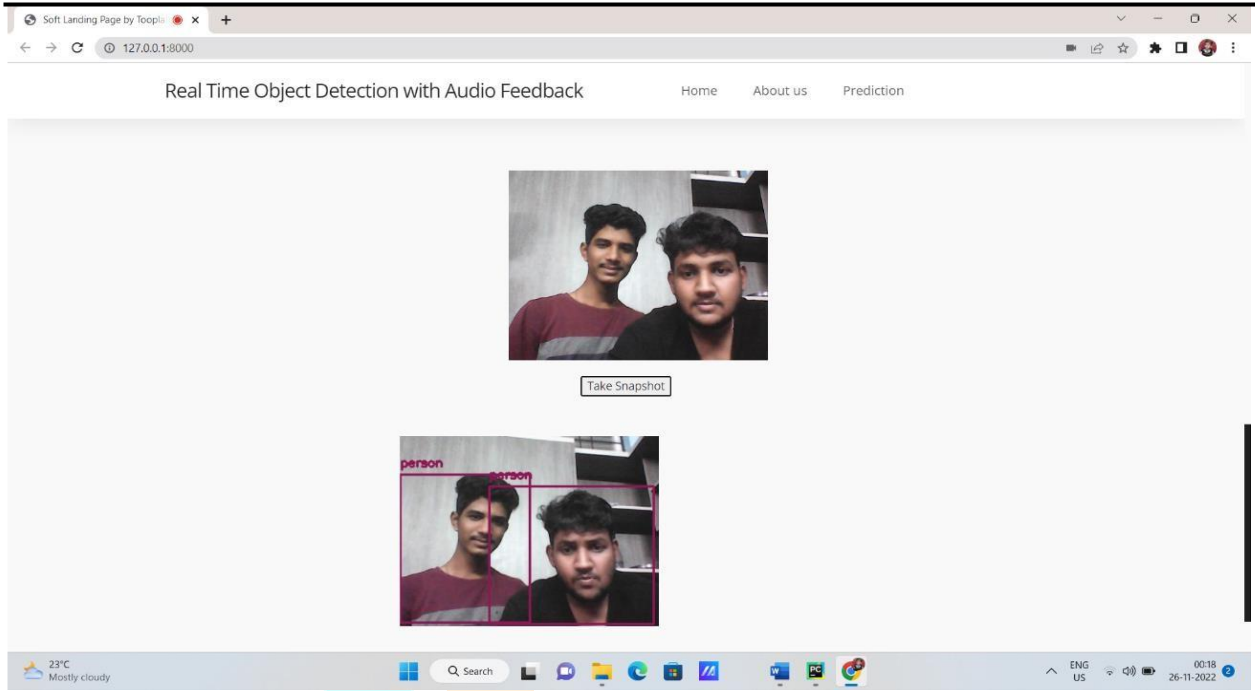


Fig 03: Prediction Output

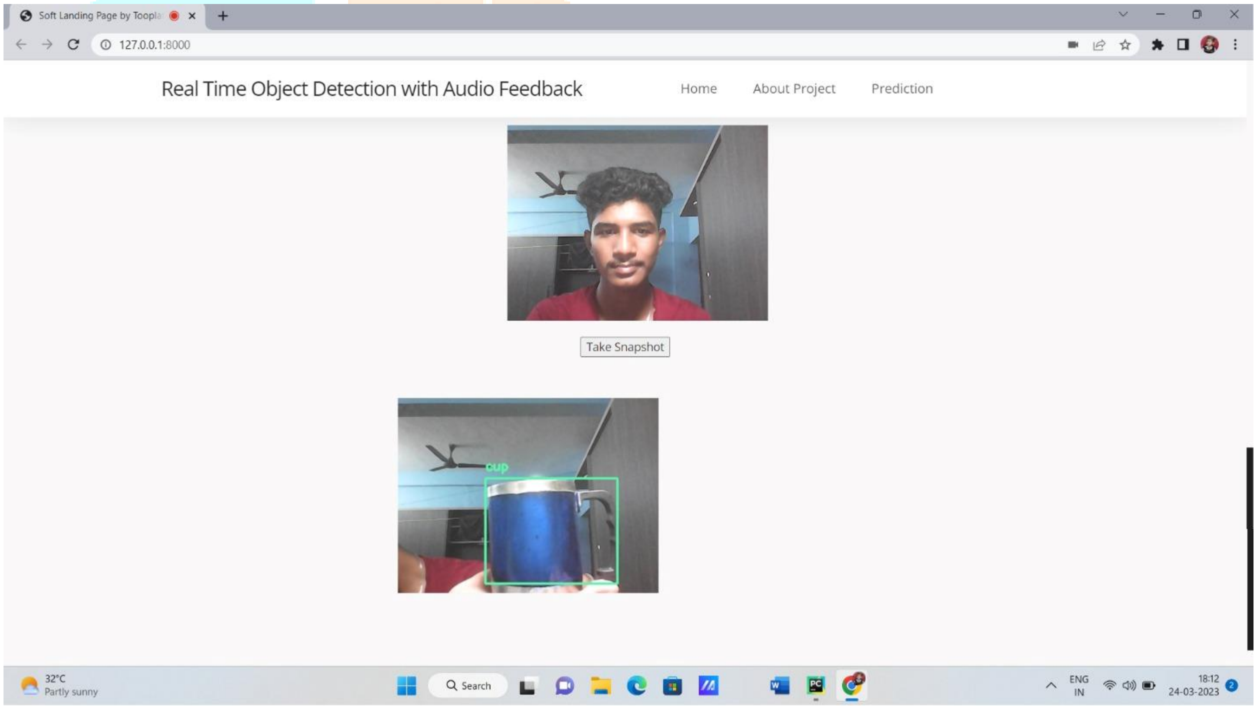


Fig 04: Prediction Output

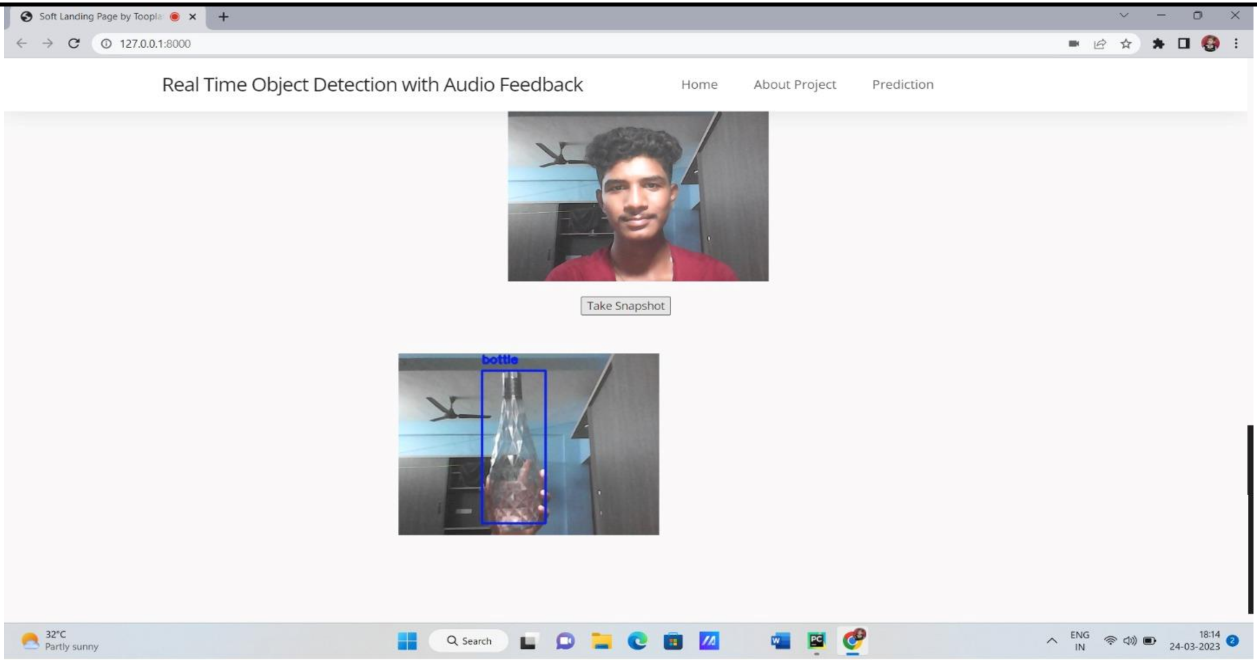


Fig 05: Prediction Output

7 Conclusion

In this chapter, we were able to detect persons using yolov3 and use web speech API to generate speech. Our project on Machine Learning-based on Object detection in surveillance systems provides a better guide Absence of vision impairment restraint the movement of the person in an unfamiliar place and hence it is very essential to take help from our technologies and trained them to guide blind people whenever they need it. Technically, human detection is a critical step in the implementation of these applications. With the advantage of high detection rates, deep learning methods have been widely employed on edge devices to detect human objects. This method has been implemented using TensorFlow, OpenCV, Webcam, and Python. Our project aims to time by reducing manpower. The final output demonstrates the system's ability to overcome the challenges regarding object detection and an automated voice feedback system. This system can be implemented in every platform for blind persons to freely do things by themselves without other person's help. By designing an interface and using Pre-trained Models Like YOLOv3. Then the images Predicted in the Front-end are detected and give audio feedback to users.

8 FUTURE STUDY

In future work, the performance of the generated algorithm could be improved. In the feature extraction part, extracted features from the auditory model may be enhanced. Instead of using mean and standard deviation, more complex methods could be used to extract features from the auditory model output. Besides, modulated signals are not the only output generated by an auditory model. Human auditory system to the brain, phase information of the first three auditory filter bank outputs. Results have shown that when the leave object sample out method is implemented, the highest accuracy rates are obtained for all three databases when compared with speaker-dependent and independent cases. In leaving one speech sample out method, all speakers are included in the training part. This shows that there is a hyperplane that can classify all seven emotions. CNN selects the hyperplane from many choices which maximize the margins. Since in the speaker-independent case, the number of training samples is low, CNN selects the hyperplane accordingly. On the other hand, when one speech sample out method is implemented, the generated hyperplane also segments the training samples in the speaker-independent case. To overcome this issue, and to generalize algorithms into the speaker or language independence, a normalization in features could be searched. Besides, the generated algorithm could be tested with noisy data, which fits the data. In that case, the algorithm could be extended to real-life cases. Since CNN is a binary classifier, binary decision trees are generated. Yet the generated binary tree may not fit all languages. Therefore, instead of a binary classifier, multi-class classifiers may increase the success rate and could work properly for many languages. Besides multi-class classifiers, using ensemble learning many different models could be fused. In the scope of this thesis, features are extracted using only the auditory model. The advantage of ensemble learning is that, developed many different Object detection recognition algorithms could be fused to obtain healthier results.

9 References:

[1] S. Cherian, & C. Singh, "Real Time Implementation of Persons Tracking Through the webcam," International Journal of Research in Engineering and Technology, 128-132, (2014)

[2] Z. Zhao, Q. Zheng, P.Xu, S. T, & X. Wu, "Person detection with deep learning: A review," IEEE transactions on neural networks and learning systems, 30(11), 3212-3232, (2019).

[3] N. Dalal, & B. Triggs, "Histograms of oriented gradients for human detection," In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05) (Vol. 1, pp. 886-893). IEEE, (2005, June).

[4] R. Girshick., J. Donahue, T. Darrell, & J. Malik, "Region-based convolutional networks for accurate persons detection and segmentation," IEEE transactions on pattern analysis and machine intelligence, 38(1), 142-158, (2015).

[5] X. Wang, A. Shrivastava, & A. Gupta, "A-fast-run: Hard positive generation via adversary for person detection," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2606- 2615), (2017).

[6] S. Ren, K. H, R. Girshick, & J. Sun, "Faster r-CNN: Towards real-time person detection with region proposal networks," In Advances in neural information processing systems (pp. 91-99), (2015).

[7] J. Redmon, S. Divvala, R. Girshick, & A. Farhadi, "You only look once: Unified, realtime person detection," In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788), (2016).

[8] J. Redmon, & A. Farhadi, "YOLO9000: better, faster, stronger," In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7263-7271) (2017).

[9] J. Redmon & A. Farhadi, "Yolov3: An incremental improvement," ArXiv preprint arXiv: 1804.02767, (2018).

[10] R. Bharti, K. Bhadane, P. Bhadane, & A. Gadhe, "Persons Detection and Recognition for Blind Assistance," International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 06, (2019).