|  |
| --- |
|  |
| Real Time Object Detection through Webcam |
| Project Report  Submitted to:  Prof. Dr. Sher Muhammad Doudpota |

|  |
| --- |
| Submitted By:  Aqsa Majeed (021-19-0014)  Muhammad Talha Munir (021-19-0034)  Muhammad Yamin (021-19-0029) |

Contents

[Introduction 2](#_Toc107523227)

[Problem statement 2](#_Toc107523228)

[Dataset Discussion 3](#_Toc107523229)

[Methodology 3](#_Toc107523230)

[1. Hand detection model 4](#_Toc107523231)

[2. Object Detection Model 5](#_Toc107523232)

[Major Outcomes 6](#_Toc107523233)

[Timeline 6](#_Toc107523234)

[Conclusion 7](#_Toc107523235)

[Project Implementation 7](#_Toc107523236)

[References 13](#_Toc107523237)

# Introduction

Computer Vision is a subfield of Artificial Intelligence that concentrates on how a computer can understand and learn by detecting and recognizing objects from digital images and videos. In the 1950s, early computer vision research used some of the first Neural Networks to recognize object edges and classify them as circles and squares. However, because neural networks require a lot of training data and computing power, it wasn't very successful at the time. With the passage of time, when the internet got matured computer vision also accelerated and was used in the 1970s for the interpretation of hand-written text for visually impaired people. Later, in the late 1990s, when the internet became so popular and everybody joined, it provided a large amount of data for the training of neural networks and hence computer vision also improved significantly. Consequently, popular applications like face recognition, live object detection, and pattern matching came into being.

Computer vision is very well needed which is used in image processing for detecting the malicious cancerous cells, classification of cars from CCTV videos, the live process of football matches, self-driving cars, healthcare, and augmented and virtual reality. Previously, both object detection and object segmentation are implemented in machine learning models but to the best of our knowledge, no model provide the output in voice command.

In this project, our objective is to implement object detection as well as object segmentation to detect the real-time object through webcam. First of all, the hand will be classified whether it is empty or holding any object. If the hand is empty then the computer should give a voice message stating “Hand is empty” otherwise tell the object name which user is holding inside the hand. To achieve our aim, we will be using OpenCV library. OpenCV is a large open-source library for computer vision, machine learning, and image processing, and it currently plays an essential part in real-time operations, which are critical in today's systems. It may be used to detect items, faces, and even human handwriting in photos and videos.

# Problem statement

Our objective is to detect the real-time objects through a webcam. For this purpose, both object detection and object segmentation techniques will be used simultaneously to complete the objective. Initially, to the best of our knowledge, there are very few models which work on both the techniques simultaneously and there is no model which responds to the user in a voice that either the hand is empty or which object is in the hand of the person.

# Dataset Discussion

The dataset that we will be using is COCO and a pre-trained model that is trained on COCO dataset. We have switched from Image-Net dataset coco because it is lightweight and works well on common object detection problems. The MS COCO dataset is a large-scale object detection, segmentation, and captioning dataset published by Microsoft. [Machine Learning](https://viso.ai/deep-learning/deep-learning-vs-machine-learning/) and Computer Vision engineers popularly use the COCO dataset for various computer vision projects.  The COCO dataset contains challenging, high-quality visual datasets for computer vision, mostly state-of-the-art neural networks. 80 object categories, the “COCO classes”, which include “things” for which individual instances may be easily labeled (person, car, chair, etc.) 91 stuff categories, where “COCO stuff” includes materials and objects with no clear boundaries (sky, street, grass, etc.) that provide significant contextual information. With COCO, Microsoft introduced a visual dataset that contains a massive number of photos depicting common objects in complex everyday scenes.

Understanding visual scenes is a primary goal of computer vision; it involves recognizing what objects are present, localizing the objects in 2D and 3D, determining the object’s attributes, and characterizing the relationship between objects. Therefore, algorithms for object detection and object classification can be trained using the dataset. COCO stands for Common Objects in Context, as the image dataset was created with the goal of advancing image recognition. The COCO dataset contains challenging, high-quality visual datasets for computer vision, mostly state-of-the-art neural networks.

The COCO dataset classes for object detection and tracking include the following pre-trained 80 objects:

'person', 'bicycle', 'car', 'motorcycle', 'airplane', 'bus', 'train', 'truck', 'boat', 'traffic light', 'fire hydrant', 'stop sign', 'parking meter', 'bench', 'bird', 'cat', 'dog', 'horse', 'sheep', 'cow', 'elephant', 'bear', 'zebra', 'giraffe', 'backpack', 'umbrella', 'handbag', 'tie', 'suitcase', 'Frisbee', 'skis’, ‘snowboard', 'sports ball', 'kite', 'baseball bat', 'baseball glove', 'skateboard', 'surfboard', 'tennis racket', 'bottle', 'wine glass', 'cup', 'fork', 'knife', 'spoon', 'bowl', 'banana', 'apple', 'sandwich', 'orange', 'broccoli', 'carrot', 'hot dog', 'pizza', 'donut', 'cake', 'chair', 'couch', 'potted plant', 'bed', 'dining table', 'toilet', 'tv', 'laptop', 'mouse', 'remote', 'keyboard', 'cell phone', 'microwave', 'oven', 'toaster', 'sink', 'refrigerator', 'book', 'clock', 'vase', 'scissors', 'teddy bear', 'hair drier', 'toothbrush'

# Methodology

In this project, we aimed to design a machine learning model that can predict the object held by a human hand. Object detection is the combination of some subtasks. Many of us get confused between object classification and object detection. The simple expiation is object classification is just assigning an image a class by detecting the features, object localization is drawing bounding boxes around the image, and object detection is the combination of both object classification and object localization. [Figure 2](#figure2) shows how object classification is different from object detection. Object detection is a computer vision task that involves the detection of objects in images or videos. The computer vision again is the subfield of artificial intelligence that train and understand the visual world. Using video and images as input computer vision models could accurately identify the objects. Computer vision is a machine learning system that uses Convolutional Neural Network (CNN) and Deep Learning to perform fast unsupervised learning on large amounts of visual information and interpret data in a manner like the movement of the human eye. The computer works the same as our brain works. The human eye captures the information and sends it to the brain and the brain interprets it and recognizes the objects.

Our task revolves around two subtasks that are recognition of human hands and object recognition. So, we will be detecting the human hands and the machine learning model will generate an audio message that” Hand is empty” and if there is an object in the human hand it will announce the object's name. The methodology revolves around the following steps.

## Hand detection model

A picture containing text

Description automatically generatedHand detection is a basic kind of computer vision application. There are dozens of algorithms and models available that can not only detects the human hand but also examine human finger positioning and movement. However, in this task, we only have to detect and localize the human hand position. We will use a machine learning model to detect the bounding box of the hand. Traditional hand detection methods primarily use low-level imaging features such as skin color and shape to detect the area of ​​the hand. Today, CNN based detection approaches have proven to be more robust and accurate due to the deep, discernible features learned. OpenCV is a powerful python library that provides a solution to computer vision-based detection problems. OpenCV provides CVzone (computer vision zone) a package that gives us a built-in in tool to detect the hand. Just calling a function by a function we can got the bounding boxes of the hand. Long with other information left of right hand and center coordinates and etc.

Figure 1: Hand detection model

## Object Detection Model

Object detection as discussed earlier, evolved around two sub-tasks that is image classification and image localization. There are various datasets and machine learning models available that provide good accuracy. For example, CIFAR-10, Open Images, BDD100K, and ImageNet datasets are popular for object detection. Deep learning provides powerful tools for object detection. Deep learning has gained popularity in the recent machine learning world due to the result obtained from image classification and detection. The object detection model first tries to find all real-world entities in the frame and then tries to find the position and label of every instance. The reason behind deep learning's popularity is the large dataset and powerful Graphics Processing Units (GPUs). So, we will be using a deep learning model for object detection that will be trained on the ImageNet dataset.

We expect to get the bounding box of the detected object and use this bounding box to compare it with the bounding box of the detected hand. On the overlapping result, we will decide whether the hand is empty or not, and if the hand is not empty then the model will tell the name of the object in the hand.

A picture containing graphical user interface

Description automatically generated

Figure 2: Difference between object classification and object detection

# Major Outcomes

After completion of this project, we will be having a web interactive dashboard that will ask the user to upload a video (.mp4) or allow the webcam permissions. Once the permission is granted it will open the webcam of the system. On the backend of this system, a machine learning model will be deployed which will be trained on the ImageNet dataset of the objects. The webcam video will be the input of the model and the output will be the video with the segmentation of the object and the detection with a clear visual (Bounding Box) around that object and the name of that object will be displayed. Moreover, the system will also be able to respond in the voice that which object is currently in the hand of the person and also detect whether the hand is empty or not.

# Timeline

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Activity** | **2022** | | | | | | | | | | | |
| **May** | | | | | | **June** | | | | | **July** |
|  | Week1 | Week2 | Week3 | | | Week4 | Week1 | Week2 | Week3 | | Week4 | Week1 |
| Requirements  Gathering |  |  |  | | |  |  |  |  | |  |  |
| Project  Proposal |  |  |  | |  |  |  |  |  | |  |  |
| Design &  Implementations of ML model |  |  |  |  | |  |  |  |  | |  |  |
| Frontend Development |  |  |  |  | |  |  |  |  | |  |  |
| Final  Integration |  |  |  |  | |  |  |  |  |  |  |  |
| Final Documentation |  |  |  |  | |  |  |  |  | |  |  |

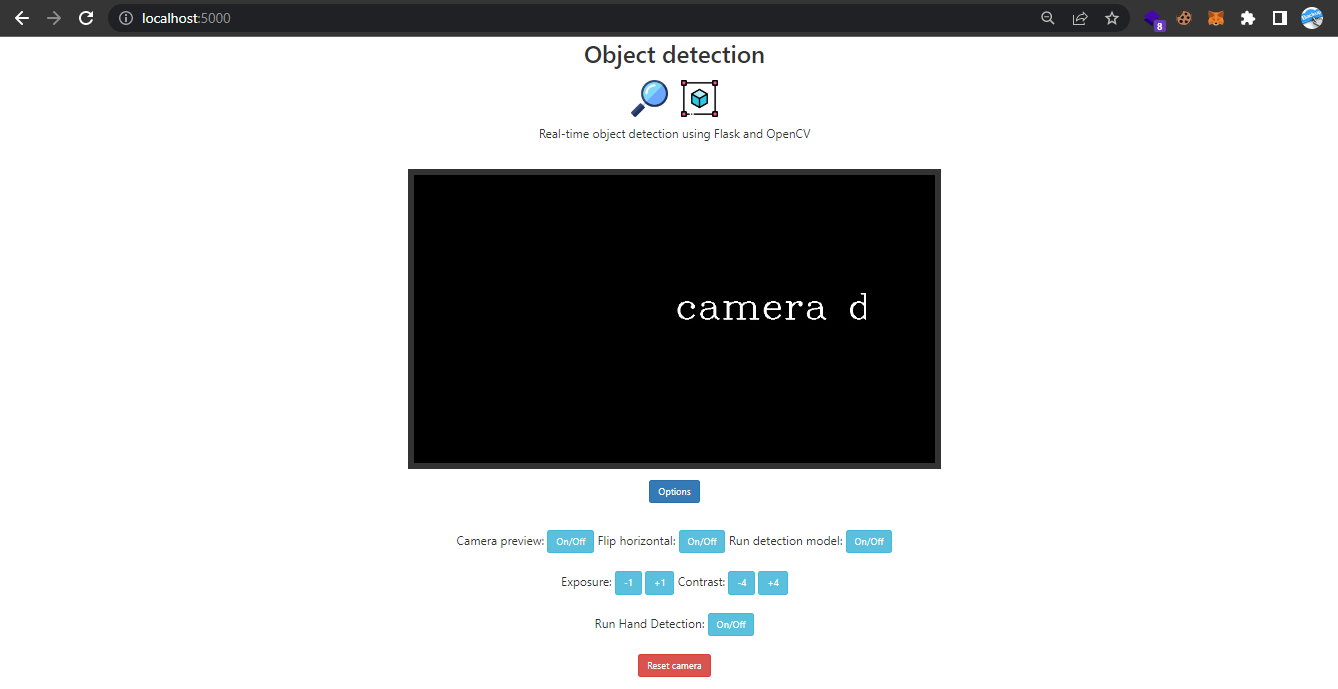
# Conclusion

Machine learning is reshaping the world through transforming a variety of industries, including healthcare, education, transportation, food, entertainment, and various manufacturing lines, among others. Similarly, object detection become a major concern in many fields such as traffic surveillance, detection of stolen things, the medical field, etc. a lot of work is done previously in this domain. COCO dataset will be used to train the model and deep learning and open CV techniques will be used to first detect the object through object segmentation and to recognize the object through object detection technique. The model will respond in voice and tells which object is in the hand of the person. The resultant system will have a momentous significance in the society and it will surely help multiple fields i.e., health care, smart irrigation, traffic surveillance and detection of stolen things even through CCTV cameras.

# Project Implementation

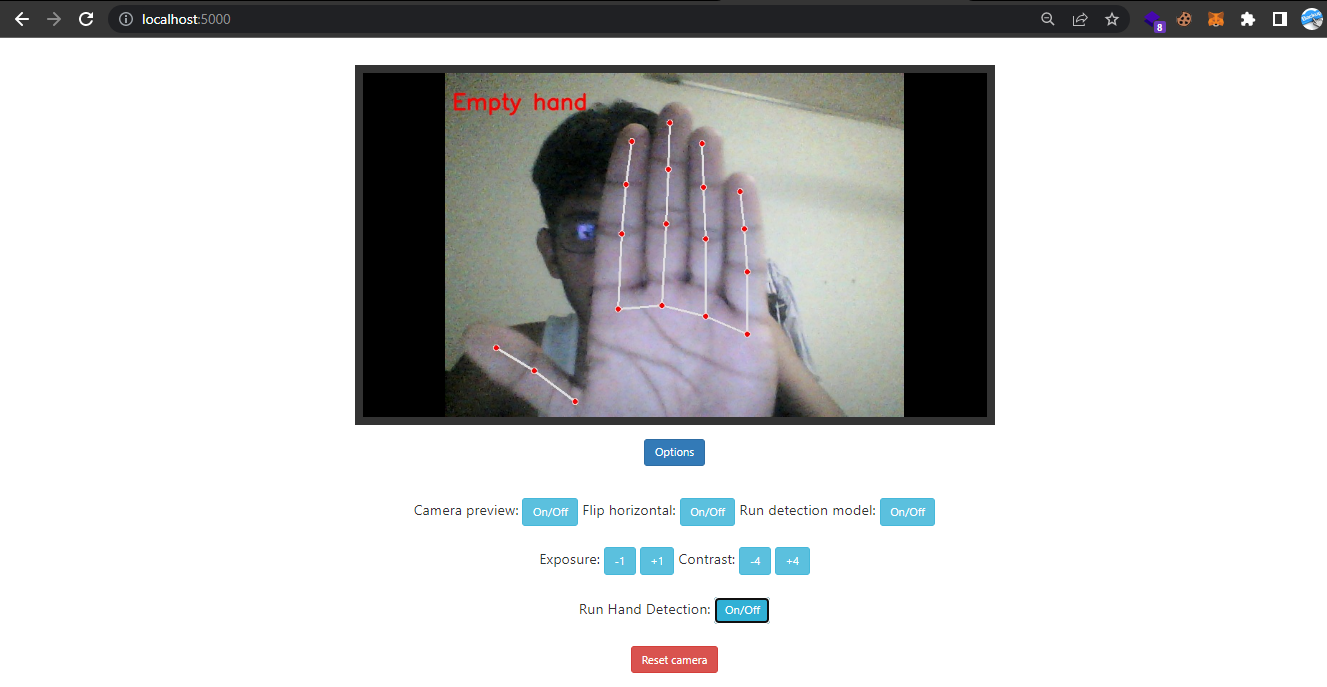
The resultant project is being deployed in the local host by designing the dashboard in the Flask framework for the direct interaction if the users. Our deployed model was trained on the 81 different classes which includes wine glass, person, train, cow, bus, spoon, glass and 74 other classes. The result was displayed in the frames ion real time and the system also responds in through voice that which object is in the hand of the person.

Some screen shots of the deployed system are as follow.

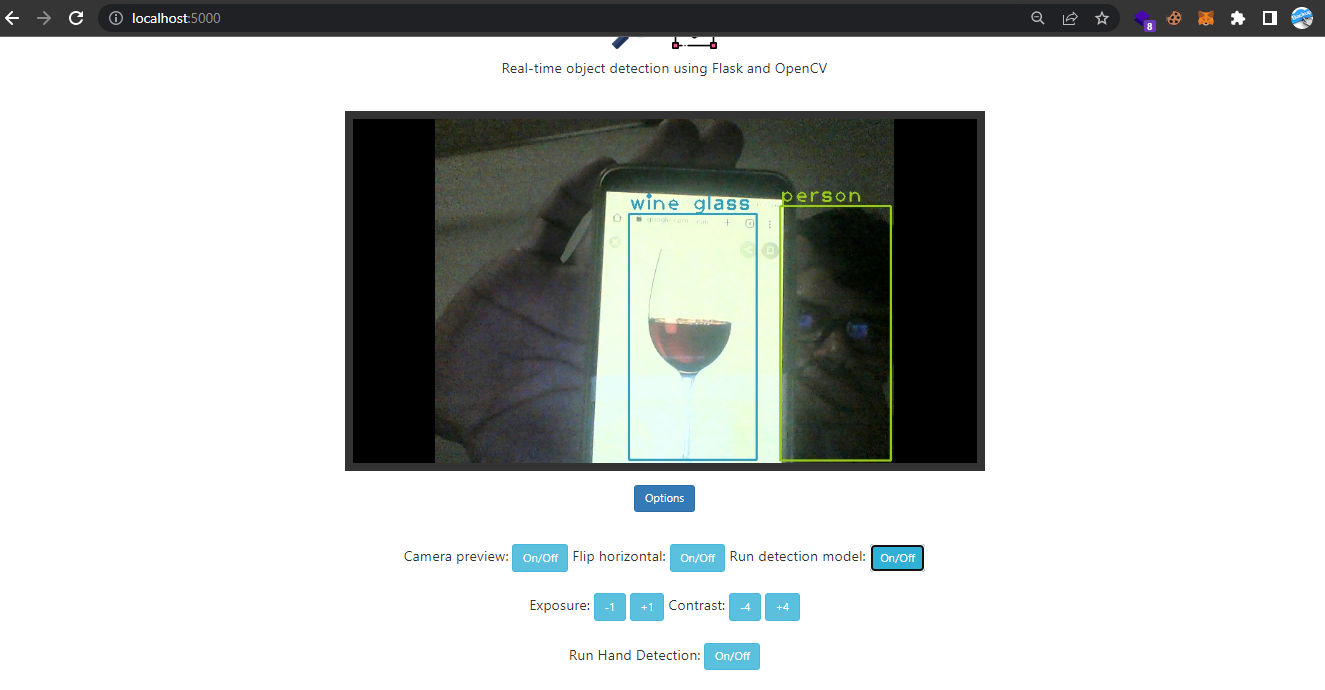
Figure 2: home page of the interface

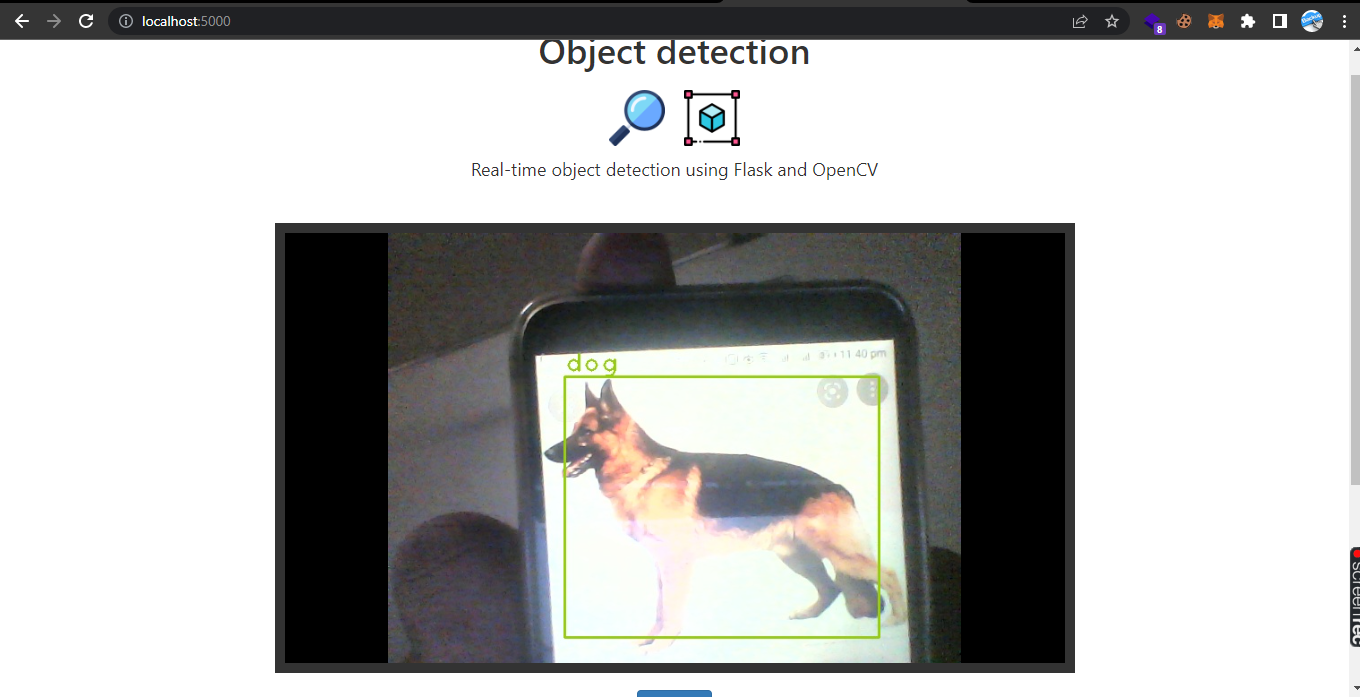
First objective of our project was to detect the hand of the person either the hand of the person in empty or not?

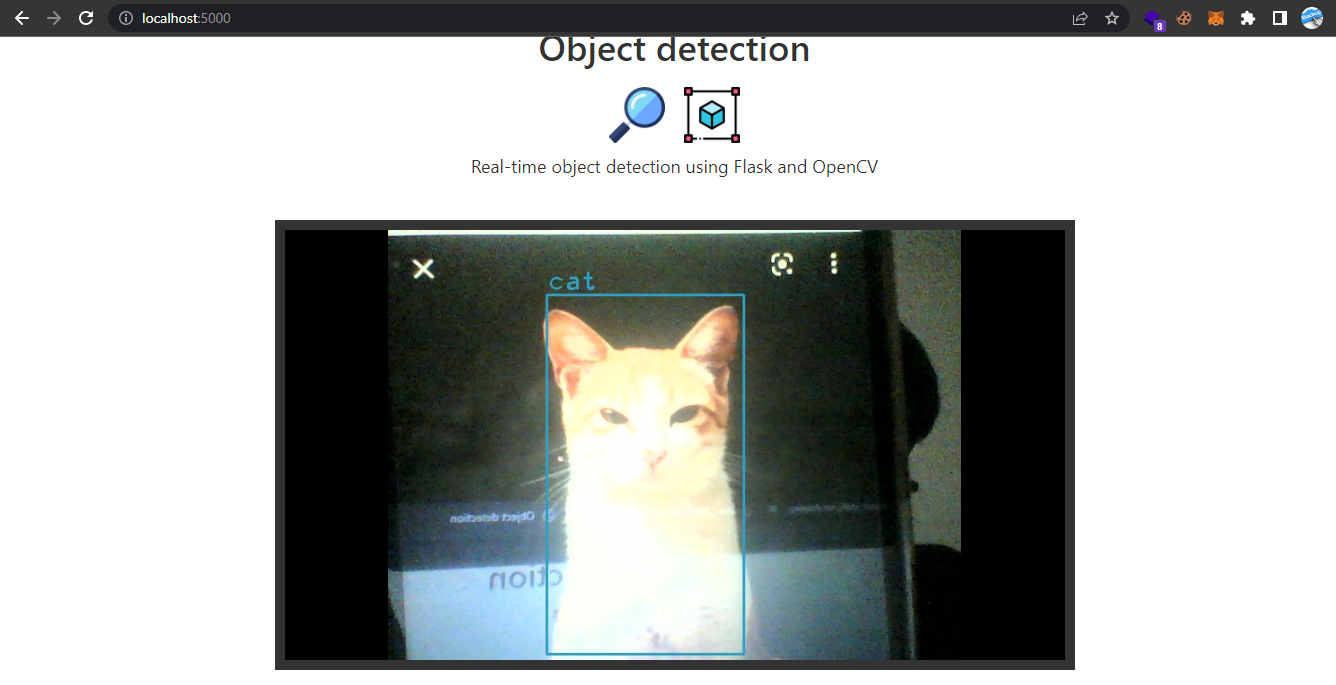
So the implementation of this module is as follow

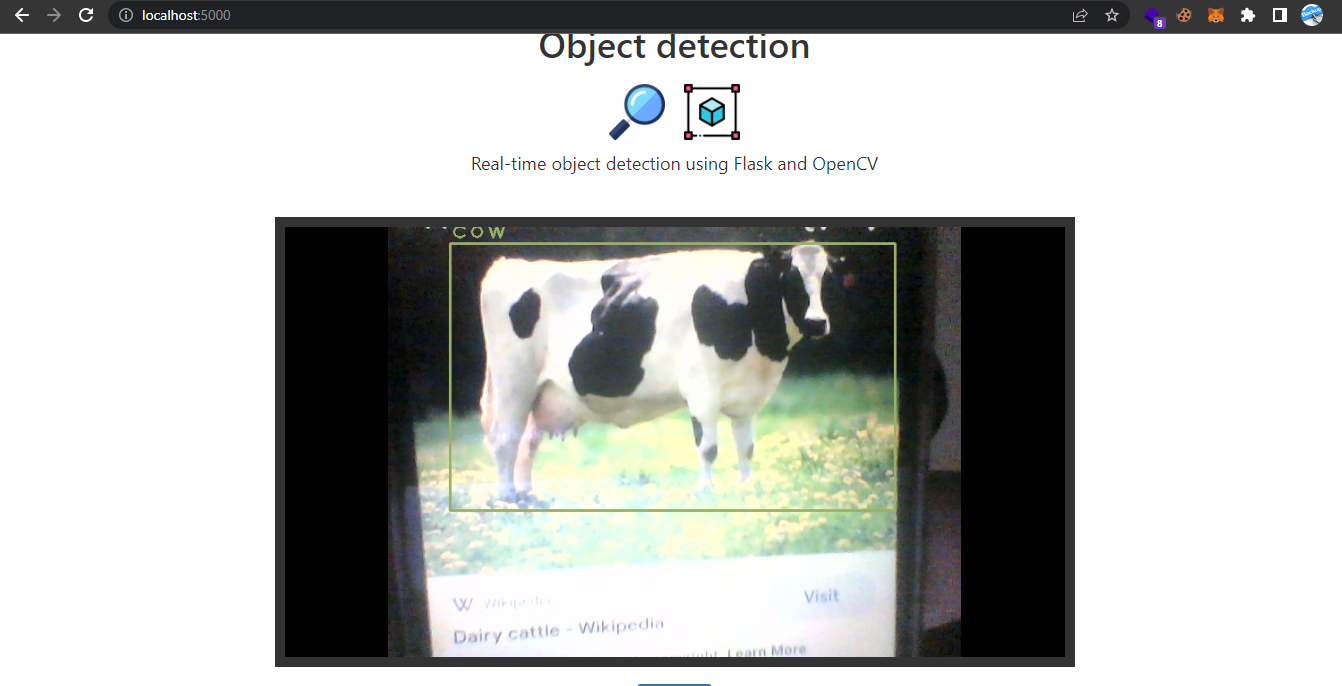
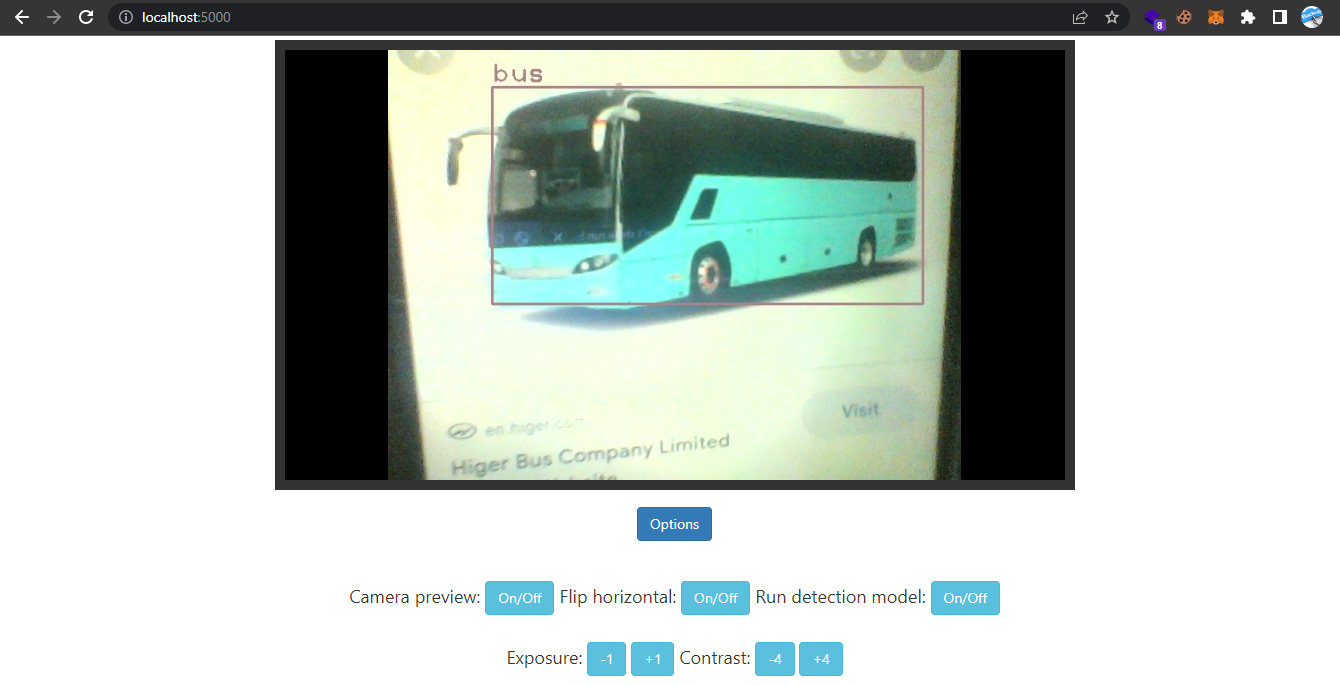
Figure 3: empty hand detection

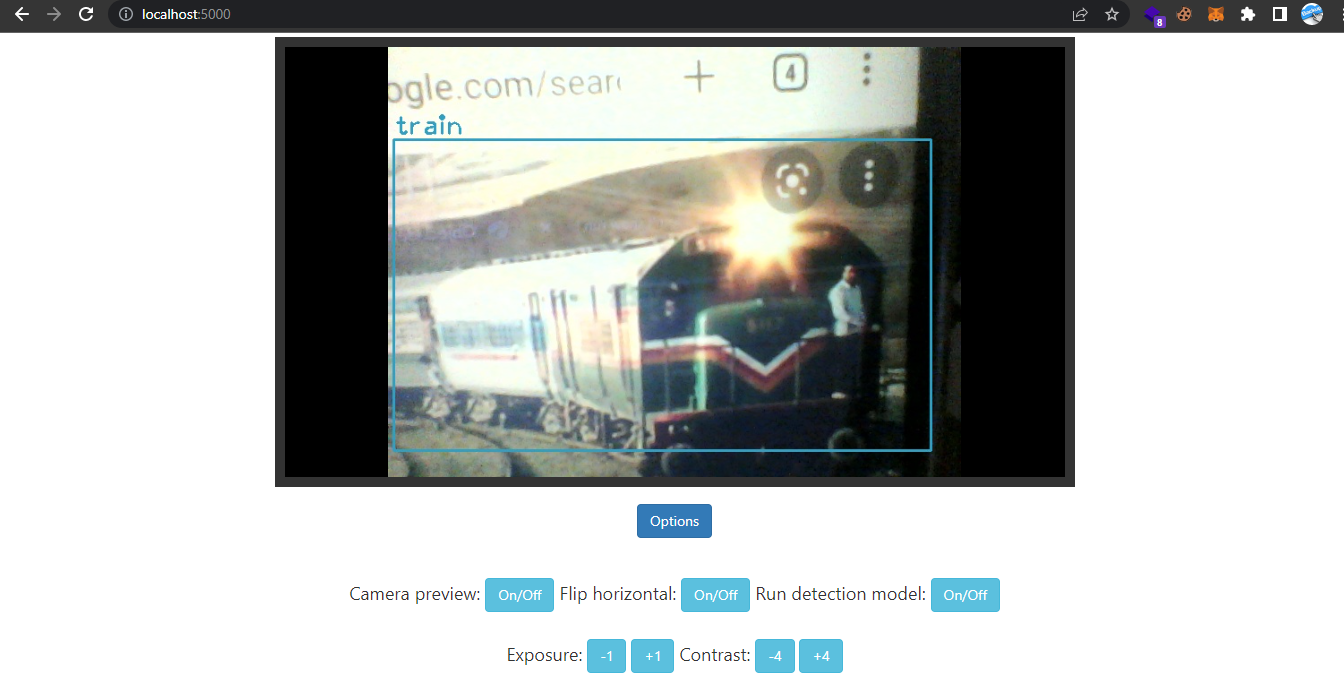
Finally, the system also detects the different objects in the hand and on some of the objects the system is being tested and the results are as follow

Figure 4: system tested on wine glass

Figure 5: system is tested in the Dog

Figure 6: System is tested on the Cat

Figure 7: System is tested on the CowFigure 8: System is tested on the Train

Figure 9: System is tested on the Train

# References

1. Kumar, A., Zhang, Z.J. and Lyu, H., 2020. Object detection in real time based on improved single shot multi-box detector algorithm. EURASIP Journal on Wireless Communications and Networking, 2020(1), pp.1-18.
2. Lu, S., Wang, B., Wang, H., Chen, L., Linjian, M. and Zhang, X., 2019. A real-time object detection algorithm for video. Computers & Electrical Engineering, 77, pp.398-408.
3. Ibm.com. 2022. *What is Computer Vision? | IBM*. [online] Available at: <https://www.ibm.com/topics/computer-vision> [Accessed 14 May 2022].
4. Sas.com. 2022. *Computer Vision: What it is and why it matters*. [online] Available at: <https://www.sas.com/en\_us/insights/analytics/computer-vision.html#:~:text=Computer%20vision%20is%20a%20field,to%20what%20they%20%E2%80%9Csee.%E2%80%9D> [Accessed 14 May 2022].
5. Verdict. 2022. *History of computer vision: Timeline*. [online] Available at: <https://www.verdict.co.uk/computer-vision-timeline/> [Accessed 14 May 2022].
6. *ImageNet Information Available* at:< <https://image-net.org/challenges/LSVRC/2017/#vid>>
7. *Detection models:*

<<https://cv.gluon.ai/build/examples_detection/demo_ssd.html#sphx-glr-build-examples-detection-demo-ssd-py>>

1. *Object localization and object detection:*

<<https://machinelearningmastery.com/object-recognition-with-deep-learning/>>

1. *CVone Madiapipe*:

<<https://www.analyticsvidhya.com/blog/2021/07/building-a-hand-tracking-system-using-opencv/>>