EYE FOR THE VISUALLY IMPAIRED – YOLOV3 OBJECT DETECTION WITH A VOICE

CS 534 COMPUTER VISION PROJECT

BY
KOUSHIK YELLISETTY (KY278)
KAUTILYA ATUL JOSHI (KJ474)
RALLA JASHWANTH YADAV (JR1756)

PROBLEM STATEMENT

- Dementia and vision problems are significant challenges faced by many individuals, making daily activities such as recognizing objects and navigating their surroundings more difficult.

Vision
Confused
Problems
Dementia
Where is
Dementia
my ?
Where is
Vision
Confused
Problems

OUR APPROACH

- Our project uses YOLOv3 object detection to help people with dementia and vision impairments. recognize and interact with their surroundings. We've added a voice recognition interface that allows users to give voice commands for detecting specific objects. Once detected, the system provides a voice output that precisely locates the object, making it easier for users to interact with their environment.
- By using this model, we hope to provide a reliable and accurate tool for enhancing their quality of life and increasing their independence

FLOW OF MODEL

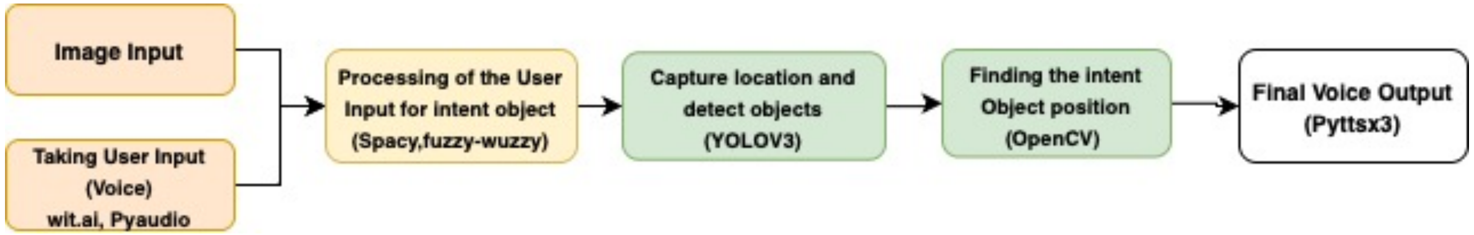


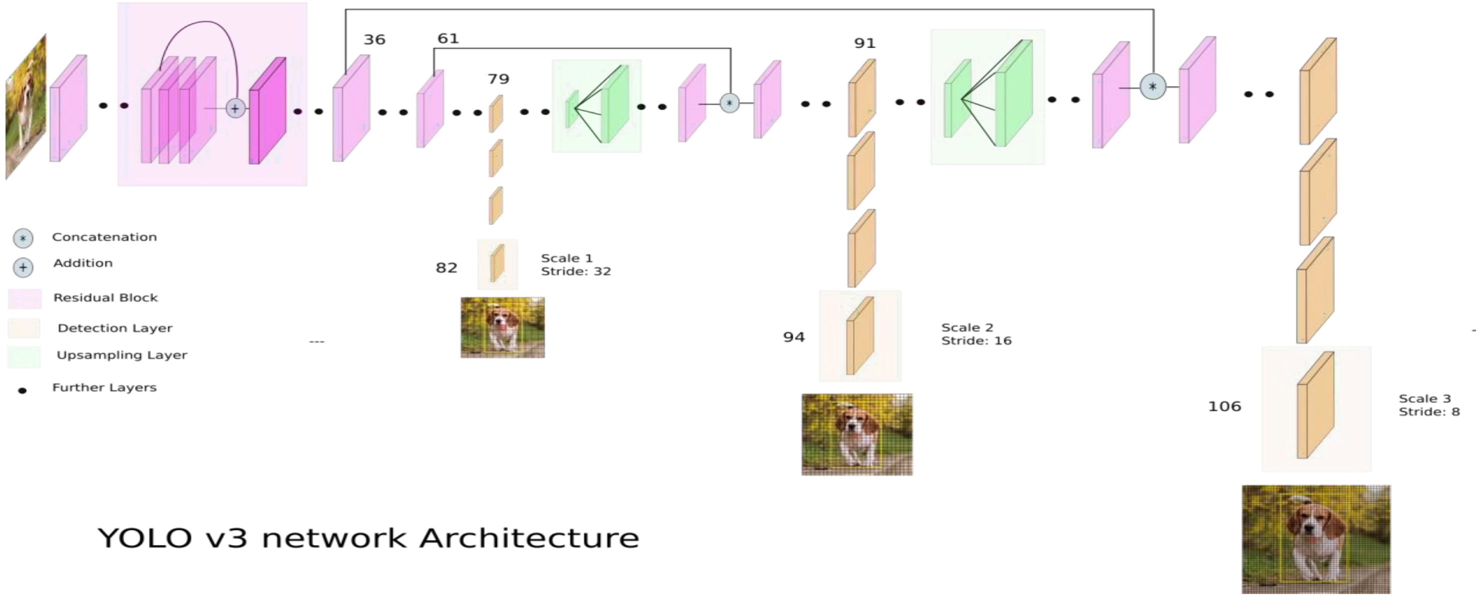
Image Input
Taking User Input (Voice) wit.ai, Pyaudio
Processing of the User Input for intent object (Spacy,fuzzy-wuzzy)
Capture location and detect objects (YOLOV3)
Finding the intent Object position (OpenCV)
Final Voice Output (Pyttsx3)

- Firstly, system will take input image and voice input from the user and process the input for the intent object.
- After that system receives input from the user about which object, they are looking for, and then passes this information to YOLOv3.
- YOLOv3 accurately detects the intended object and nearby objects, and provides the precise location of the intended object, which is then relayed back to the user through voice output.
- By using YOLOv3 for object detection, we aim to provide a reliable and efficient tool for enhancing the independence and quality of life of individuals with dementia and vision problems

YOLOV3

- YOLOv3 is a state-of-the-art object detection algorithm that uses a deep neural network and anchor boxes to achieve high accuracy.
- The YOLOv3 architecture consists of a feature extractor, convolutional layers, and a final detection layer that outputs bounding boxes and class probabilities.
- YOLOv3 uses non-max suppression to filter out overlapping bounding boxes and improve detection accuracy.
- Hyperparameters such as the learning rate and anchor box scales can be tuned to improve YOLOv3's performance.
- The YOLOv3 loss function: L = λ{coord} * L{coord} + λ{noobj} * L{noobj} + L_{class} + λ{obj} * L{obj}
- The YOLOv3 anchor box equation: t_x = \sigma (w_x) + c_x
- The YOLOv3 confidence score equation: Pr(obj) * IOU^{truth}{pred} = Pr(obj) * Pr(class_i) * IOU^{truth}{pred,class_i}

YOLOV3 ARCHEITECTURE



- The YOLOv3 architecture consists of a feature extractor, convolutional layers, and a final detection layer that outputs bounding boxes and class probabilities.

MODIFIED YOLOV3 FOR INTENT OBJECT DETECTION

- In our modification, we obtained the central coordinates of each detected object by calculating the midpoint of the bounding box using the top left corner coordinates and width and height.
- By comparing the distance between the intent object and detected objects, we were able to determine the position of the object of interest relative to the nearest detected object
- And we also even added the commands related to object position. Whether the object is right, left or near to the object so that it will be helpful to user.
- . Our model is capable of detecting any object present in the COCO dataset and Custom dataset(book, bottle and watch)

VOICE ASSISTANT

- (INPUT)speech-to-text: The user's speech is recorded using Python's Pyaudio module and converted to text format using the Wit.ai interface.
- Processing of Text: The text is tokenized, stop words are removed using spacy framework, and the fuzzy-wuzzy module is used to find the most accurate object name. And passes the object name to YOLOV3.
- (OUTPUT)Text-to-Speech: The final step of the voice assistant is performed using Pyttsx3 module, which creates a voice engine object, sets the output voice, and passes the text argument to the say() function for voice output

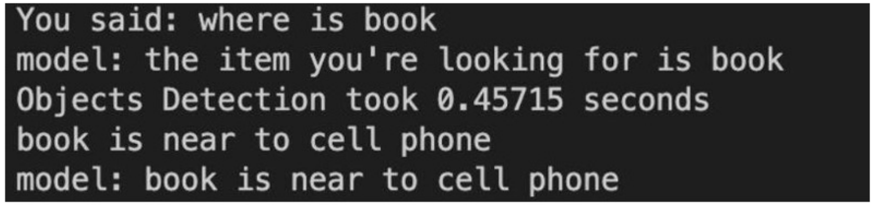
CUSTOM DATASET

- Created a custom dataset of three classes, bottle, book and watch, each consisting of 100 images with the help of OIDv4_ToolKit.
- As we have converted a live video which consists of the aforementioned classes and converted into images and labelled them using OIDv4_ToolKit and created the data set.
- Partial labelling of the generated images was done manually, then referring the manual labelling, auto-labelling feature of the OIDv4_ToolKit was used to generate labels for the rest of the images.
- After that we have trained this dataset in darknet neural network framework with yolov3 training configuration in order to generate and acquire weights for our YOLOv3 model on our custom dataset.

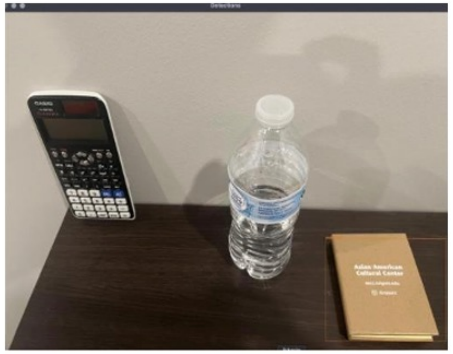
RESULTS

User Input: Where is Book

Output: Model provides the output through voice. Here is the snapshot of the output.

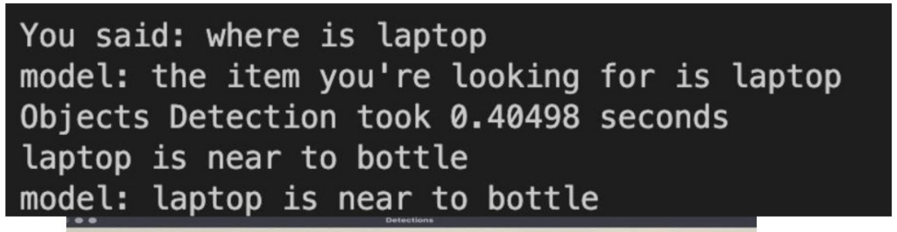


```
You said: where is book
model: the item you're looking for is book
Objects Detection took 0.45715 seconds
book is near to cell phone
model: book is near to cell phone
```

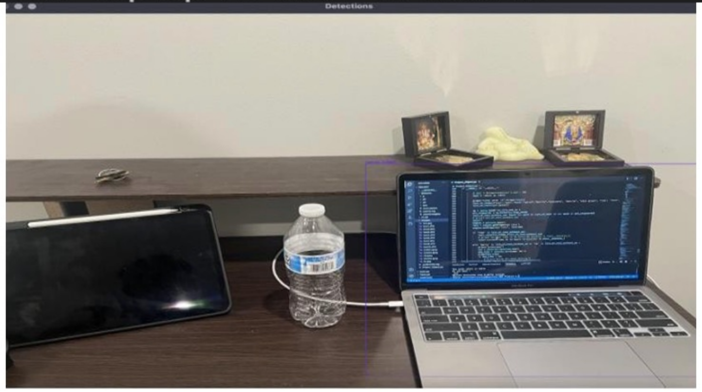


User Input: Where is Laptop

Output: Model provides the output through voice. Here is the snapshot of the output.



```
You said: where is laptop
model: the item you're looking for is laptop
Objects Detection took 0.40498 seconds
laptop is near to bottle
model: laptop is near to bottle
```



SCOPE OF IMPROVEMENT

- Instead of a computer application, a dedicated SoC (System on Chip) device can be made to listen to user's input commands 24X7
- Providing more precise output commands with additional details about the estimated location of intended object to other objects in the vicinity
- Fine-tuning our model using transfer learning techniques can improve its accuracy in detecting a wider range of objects and make it more customizable to any dataset.
- Attention mechanisms can be incorporated into our model to improve its ability to focus on relevant information and reduce false detections.
- Model running on video rather than images

THANK YOU