SFU
Simon Fraser University

RETINA: Real Time Speech Activated Assistant

Anuj Saboo, Rishabh Jain, Ankita Kundra

PROBLEM

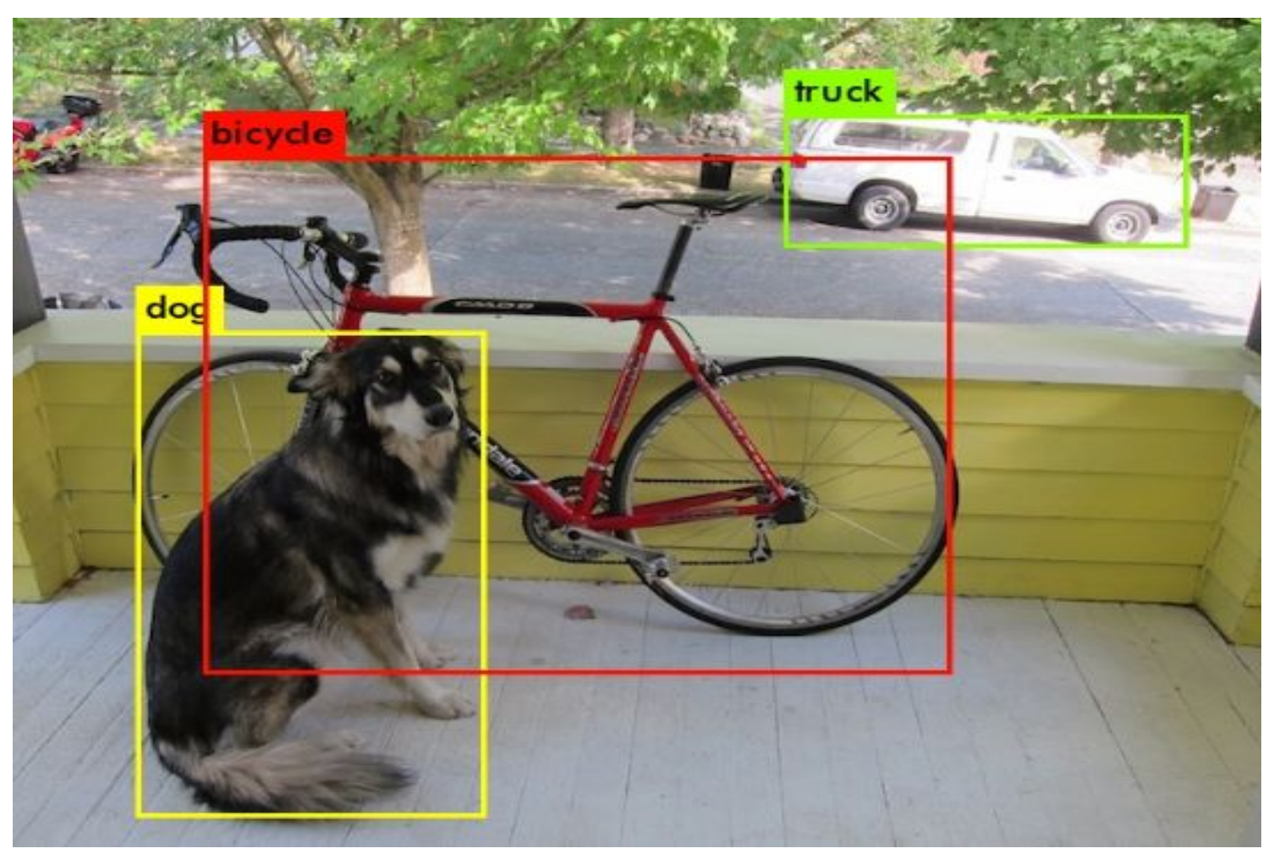
Object Detection is an exciting field in Machine Learning and Computer Vision. Working with robots and drones to assist humans in simple tasks or dangerous locations is essentially a necessity in today's world. Our application is motivated from that very idea to enable a human to talk to a system and detect an object in the live video feed. This involves threaded programming to enable both video and speech inputs working simultaneously with the model to classify the object.

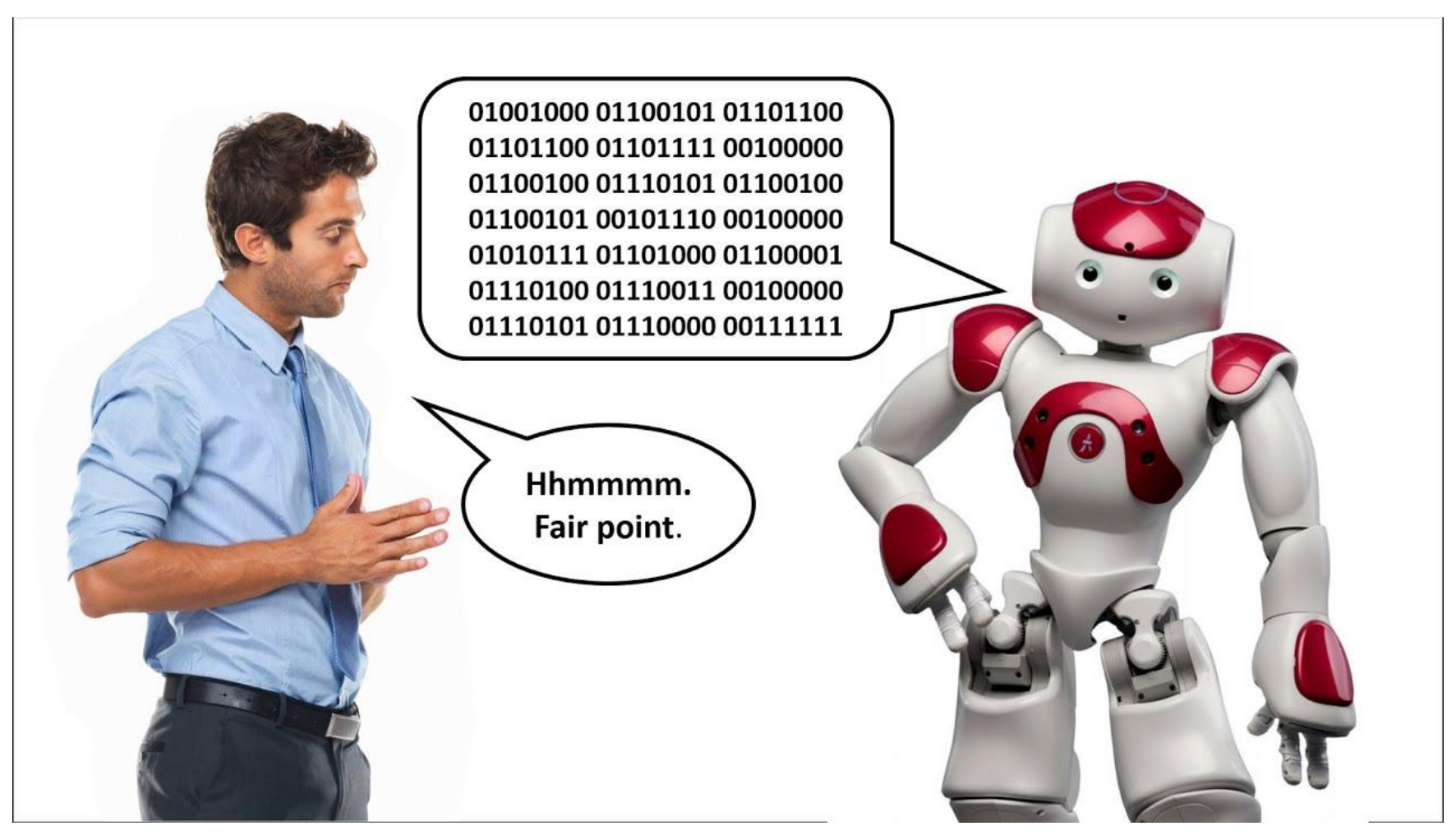
A wide variety of algorithms are available to perform object detection but the results of YOLO v3 are superior to them. We use the model trained on COCO dataset to detect various objects in our surroundings. In addition to this, we train our own model with various objects encountered to achieve the same objective.

The task of speech recognition still remains an issue due to interference of noise. Relevant parameters were tuned but much more can still be done to achieve a better translation.

YOLO v3 over v2

YOLO v3 is an improvement over v2 which used a 30-layer architecture and struggled with small object detections. YOLO v3 incorporates the important elements lacking in v2 and uses a variant of Darknet, which originally has 53 layer network trained on Imagenet. For the task of detection, 53 more layers are stacked onto it, giving us a 106 layer fully convolutional underlying architecture for YOLO v3.





Speech Assistance

The user input is passed as a WAV file to the speech recognizer aided with google recognizer to interpret the speech and pass it to the model for detection in the live video feed.

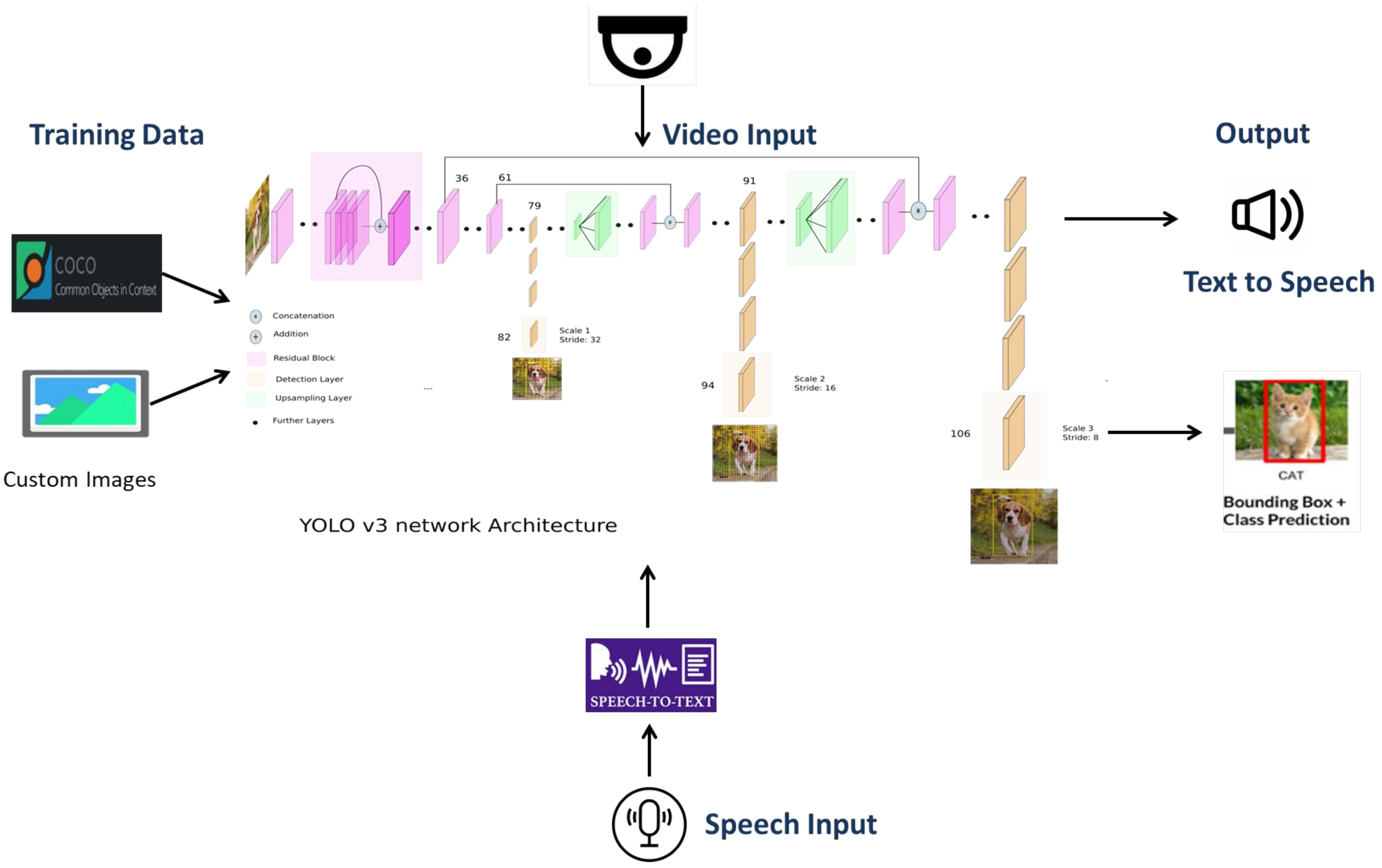
MODEL ARCHITECTURE

We use a pre-trained model for detecting COCO dataset objects and further train a new model to detect objects in a custom data set.

The application continues to look in a live webcam feed generated through OpenCV. In parallel, post a wakeup command, a text input is passed to the system and converted to text using Speech Recognizer.

YOLO v3 is used to detect the existence of the object(name passed through speech) in the live camera feed.

The output result is a speech feedback for the detection result along with a bounding box with category classification prediction in the video feed.



YOLO v3 network Architecture

APPLICATIONS

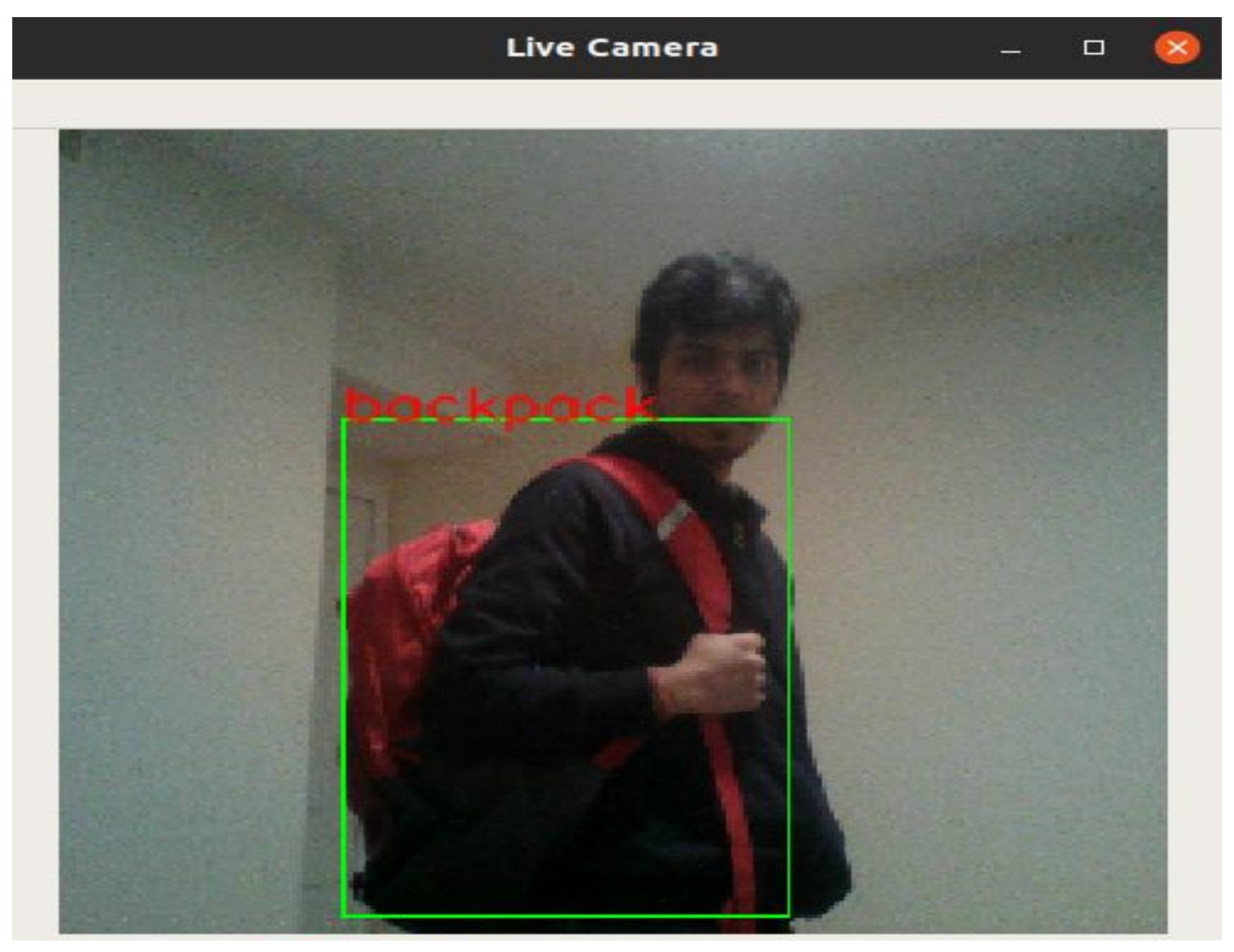
- Service Robots can act as human aid in essential but simple services to manage everyday tasks
- Conversation enabler with machines to act on detection of objects and work in dangerous locations. eg. underground mines, enemy weapon detections etc.

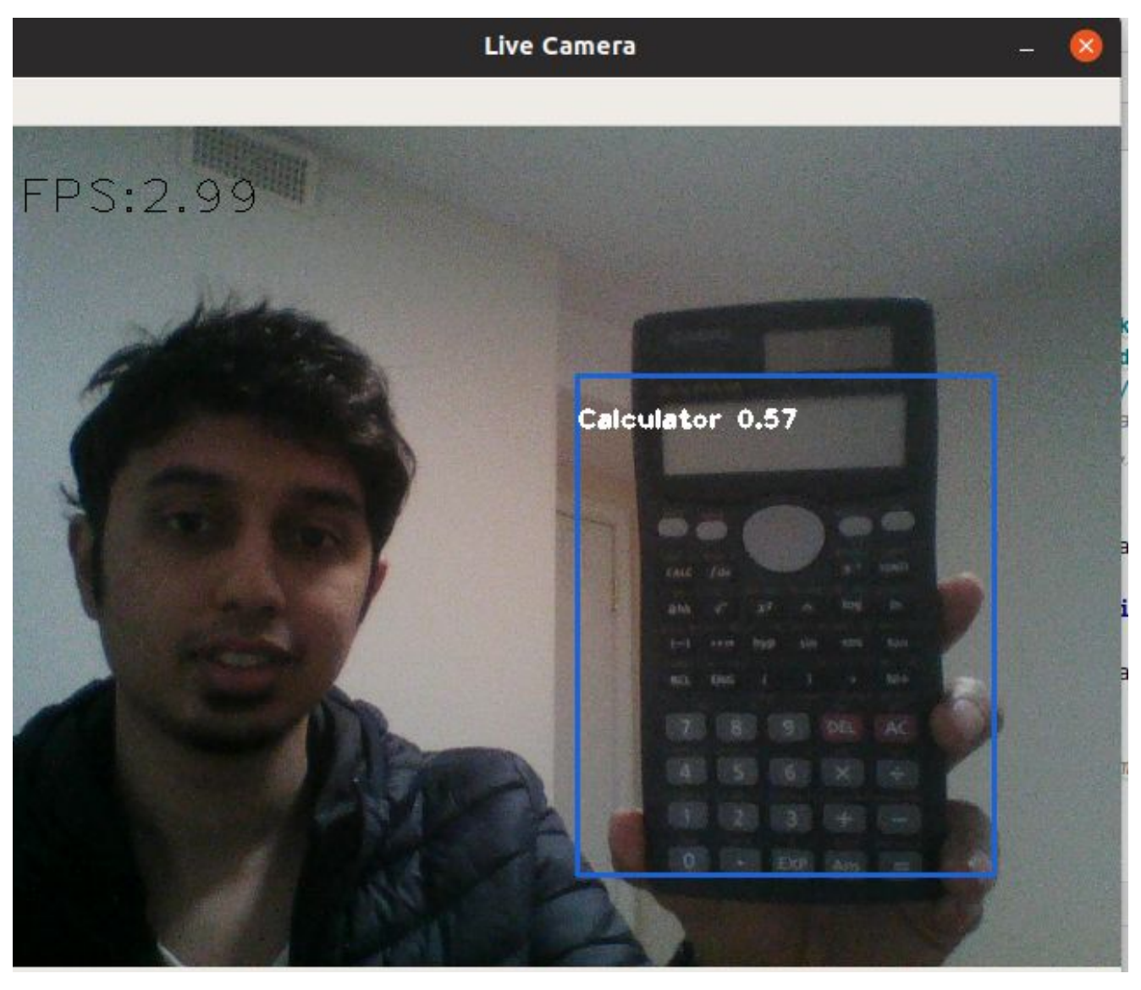
FUTURE WORK

- Larger custom dataset to train our own model to increase classification accuracy
- Designing an android app to use the phone's microphone and camera to run the application
- Tune parameters related to speech processing to improve human to machine conversation in noisy environments

RESULTS

- Input Speech is translated to text through Speech Recognizer
- Object is detected in the video frame through YOLO v3
- Voice feedback from system to confirm detection
- The application can successfully detect COCO dataset objects as well as self trained objects. This can be extended to remotely instruct drones and robots to detect objects through their eyes(camera) and further perform action based on requirement.


Live Camera
backpack


Live Camera
FPS:2.99
Calculator 0.57