International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 04 | April - 2023 Impact Factor: 8.176 ISSN: 2582-3930

OBJECT DETECTION AND TEXT TO SPEECHCONVERSION BASED ON YOLOV7 USING DEEP LEARNING

1st KOTA TEJASWINI
B.Tech Final Year
Department of Electronics & Communication Engineering
R.V.R. & J.C. COLLEGE OF ENGINEERING
GUNTUR, Andhra Pradesh-522019
tejaswini.kota68@gmail.com

2nd KONDA SAI CHAITHANYA
B.Tech Final Year
Department of Electronics & Communication Engineering
R.V.R. & J.C. COLLEGE OF ENGINEERING
GUNTUR, Andhra Pradesh-522019
saichaithanyakonda@gmail.com

3rd KETHAVATHU LAKSHMI BAI
B.Tech Final Year
Department of Electronics & Communication Engineering
R.V.R. & J.C. COLLEGE OF ENGINEERING
GUNTUR, Andhra Pradesh-522019lakshmi9908565513@gmail.com

***Abstract*—Object detection is a computer vision technique that locates objects in images or videos by creating bounding boxes around them. In this paper, we propose a model based on object detection using deep learning technologies along with text to speech conversion.An object detection system uses a deep learning model to detect objects using YOLO (You Only Look Once) and text-to-speech (TTS) to synthesize a voice announcement about each object. The system we used is built using python OpenCV tool and Google text to speech (gTTS) is used to convert text into audio segment. First variations of YOLO algorithm are compared and then the best one is used according to result we get it by training it on COCO dataset. After the object is detected, the name of the detected object is displayed then the voice output is generated by using Google Text To Speech(gTTS) module. The contribution we make is to present a visual substitution system that uses features extraction and matching to recognize objects with a voice feedback.**

***Index Terms*—Object Detection, YOLO, Open CV, python, Google Text To Speech**

I. INTRODUCTION

Computer vision research has expanded rapidly and successfully in recent years. An object detection system is one of the first tasks in a computer vision system, since it allows further information about the detected object to be obtained. In this project, we explored the possibility of using the hearing sense to understand visual objects. The sense of sight and hearing sense share a striking similarity. Using a voice feedback system, we built a real-time object detection system. We used YOLO algorithm trained on the COCO dataset to identify the object present in the image. Then the label of the object is identified and then converted into audio by using Text to Speech conversion which will be the antici- pated output. A computer vision task called object detection involves identifying and locating objects in images or videos. There are many applications for it, such as surveillance, self- driving cars, or robotics. As a result of object detection, other important AI vision techniques such as image classification, image retrieval, and object co-segmentation can be employed to extract meaningful information from real-life objects. The detection of objects can be roughly divided into two categories based on how many times the same image is passed through the network (single-shot detectors versus two-stage detectors). In single-shot object detection, predictions about the presence and location of objects are made based on a single pass through the input image. By processing an entire image in one pass, it is computationally efficient. YOLO is a single-shot detector that processes an image using a fully convolutional neural network (CNN). The two-shot object detection method makes predictions about the presence and location of objects using two passes of the input image. In the first pass, proposals or potential locations for objects are generated, and in the second pass, the proposals are refined and final predictions are made.

II. RELATED WORK

In computer vision tasks like face detection and face recognition, object detection is widely used. It can also be used for tracking objects, such as tracking a ball during a football match, tracking the movement of a cricket bat, or tracking a person in a video. Colour characteristics have been used in many previous works to find items. In order to find flowers

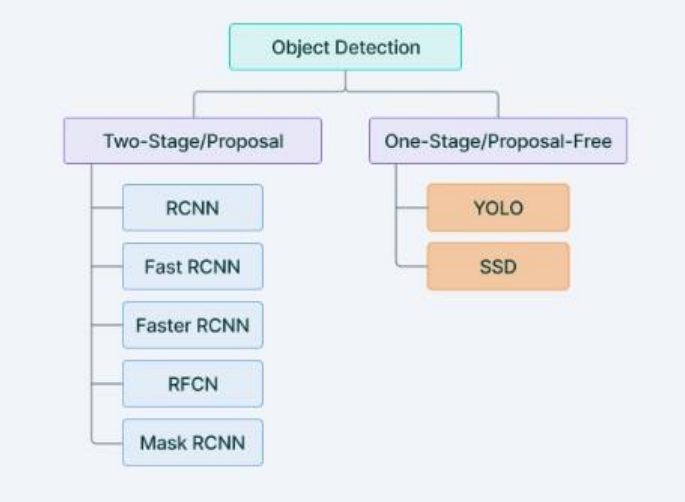

Fig. 1. Single-shot and two-shot object detection

at the scene, the authors used the colour separation method using thresholding. The colour of the flower was considered to be the basic commodity of an agricultural product. The colour of the flower is determined by histogram analysis in the HSV colour field. Afterwards, the flowers were separated from the background and obtained from photographs. Researchers are finding things based on physical characteristics. A two- dimensional shape can be detected in an image as well as a type of the shape can be detected. The known shapes are identified by dividing the images into corresponding regions, determining the shape element, and using it to identify the shape type. Objects at a particular distance from a point (i.e. the center) are sought when looking for circles. Similarly to find squares, it is important to find objects with equal side lengths and perpendicular corners. Face identification uses a similar approach, where eyes, noses, and lips can be identified, along with features like skin color and distance between eyes.

III. DATASET

The dataset we used was COCO. The COCO acronym stands for Common Objects in Context. In computer vision, the COCO dataset is widely used. There are more than 330,000 images, each annotated with 80 categories of objects. It is possible to train object detection models using the COCO dataset.This dataset contains annotations that can be used to train machine learning models that recognize, label, and de- scribe objects. Datasets provide bounding box coordinates for 80 different types of objects, which can be used to train models that detect bounding boxes and classify objects. The COCO dataset, a large data for object detection, segmentation, and captioning, can be used to train deep neural networks.Some features you can anticipate from MS COCO:

- Object segmentation
- Recognition in environment
- Superpixel stuff segmentation
- Pretrained images of 330k
- Object instances of 1.5 million
- 80 classes of different objects
- 91 stuff categories
- There are 5 captions per image
- 250,000 people with key points

IV. METHODOLOGY

The project aims to include state of the art technique for object detection with the goal of achieving high accuracy with a real-time performance. For many of the object detection systems, the deep learning-based approach relies on other computer vision techniques for assistance, leading to slow and non-optimal performance. This project uses an end-to- end deep learning approach to solve the problem of object detection. In this paper, we introduce YOLO, a new method for detecting objects. You Only Look Once(YOLO) is used to detect the objects present in the image. The bounding boxes and class probabilities are directly predicted by a single neural network from full images. As the whole detection pipeline is a single network, it can be optimized end-to-end directly on detection performance.

A. Programming Language and libraries

We implemented our proposed models using Python pro-gramming language. We used the following libraries:

- NumPy for numerical computations
- Pandas for data manipulation
- LabelImg for labelling the object names
- gtts pyttsx3 playsound is used for text to speech conver-sion

B. Problem Statement

Many existing models do real-time object detection with voice feedback, but the major disadvantage is that they use older algorithms like Effective Det, R-CNN, ATSS, ASFF or YOLOv3, YOLOv4. With these algorithms, accuracy and real time speed don't always go hand in hand, if accuracy is good, real time speed is slow, and vice versa. Our proposed system uses YOLOv7 as the main difference from the existing one. A new loss function called "focal loss" is used in Yolo v7. In previous versions of YOLO, a standard cross-entropy loss function was used, which is less effective for detecting small objects. It was previously possible to detect small objects using YOLO's cross-entropy loss function, but this is less effective at detecting small objects in the current version. As well as having a higher resolution than previous versions, YOLO v7 also has an improved user interface. 608 by 608 pixels are processed in this version, compared with 416 by 416 pixels in version 3. Because of the higher resolution, YOLO v7 is capable of detecting smaller objects with better accuracy. Aside from that, it can process images at speeds of 155 frames per second, which is much faster than other advanced algorithms for detecting objects. In addition to that, this project uses Python3. A 30 frame per second frame rate is obtained by initializing the camera using the OpenCV library. The algorithm takes them into account. As soon as the object has

been identified, the system uses an algorithm called YOLOv7 that has been trained on the COCO dataset. Through text-to-speech conversion, the object identification is converted to an audio segment.

C. YOLO

YOLO, or You Only Look Once, is an algorithm that detects and recognizes different objects in a picture. In YOLO, object detection is performed as a regression problem and class probabilities are provided. The YOLO algorithm uses convolutional neural networks (CNN) to detect objects in real-time. There is only one propagation that occurs throughout the neural network when making predictions, so the algorithm only looks at the image once. As compared to other methods of object identification, the YOLO model is the fastest and most accurate. YOLO's main advantage is its quickness. A frame rate of 45 frames per second is used here. To provide an abstract description of things to its network, the model is constructed in a concise manner. When applied to the COCO dataset, the You Only Look Once (YOLO) architectural algorithm results in a quick and effective deep learning tech- nique for recognizing objects. The framework of Yolo mainly consists of three components:

- Backbone
- Head
- Neck

Through the neck, the Backbone extracts essential features of an image and feeds them to the Head. A feature pyramid is created by the Neck based on the feature maps collected by the Backbone. Final detections are made on the output layers of the head.

YOLO algorithm works based on three techniques:

- Residual blocks
- Bounding box regression
- Intersection Over Union (IOU)

1) Residual blocks: First, the image is divided into various grids, each with a dimension of S x S. As you can see in the image below, there are many grid cells with the same dimensions. Grid cells will detect objects that appear within them. A grid cell responsible for detecting an object center, for instance, will be used if the object center appears within that cell.

2) Bounding box regrssion: Bounding box highlights the objects in an image by drawing an outline around them. Each bounding box can be described using four parameters. They are:

- Width(Bw)
- Height(Bh)
- Class(C),
- Centre of the bounding box(Bx,By)

There is one more predicted value Pc which is the probability that there is an object in the bounding box. If there is an object present in the bounding box, then Pc value corresponds to the number one (Pc=1). If there is no object present in the bounding box, then the Pc value corresponds to the number zero (Pc=0). In relation to the enveloping grid cell, Bx, By are the coordinates of the center of the bounding box. Bw, Bh corresponds to the width and the height of the bounding box. We can have many classes as per our requirement. The final vector representation for each bounding box is represented by Y = [Pc, Bx, By, Bw, Bh, C1, C2]

Fig. 2. Residual blocks

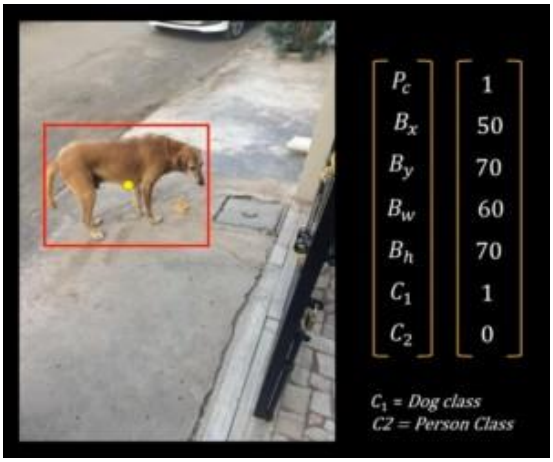

Fig. 3. Parameters defining the bounding box

From the fig-3, we can say that Pc is equal to one indicating the presence of the object. The co-ordinates Bx and By represent the centre of the bounding box. Bh represents the height of the bounding box. Here we have taken two classes i.e C1 and C2. The class C1 indicates a dog and the class C2 indicates a person. Since the object present in the bounding box is dog, the value of the class C1 becomes one (C1=1). Therefore, the value of the class C2 is zero (C2=0). Hence, the vector presentation of the bounding box is represented as Y = [1, 50, 70, 60, 70, 1, 0]

Anything can be included in the class, such as a person, a car, a traffic light, etc. An object's height, width, center, and class can be predicted using YOLO's single bounding box regression.

3) Intersection Over Union (IOU): When detecting objects, Intersection Over Union is used to describe how boxes cross

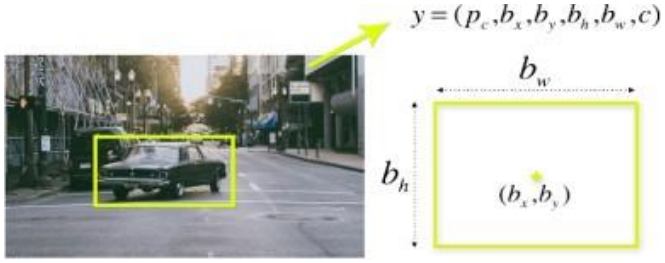

Fig. 4. Bounding box around the object i.e car

over each other. IOU is used by YOLO to provide an output box for the objects that perfectly surrounds them. Based on the confidence scores for the bounding boxes, each grid cell determines their bounding boxes. In the case of a real bounding box which is the same as the predicted bounding box, then the IOU is equal to 1. The process of eliminating bounding boxes that are not equal to their real counterparts is accomplished by this mechanism.

$$\text{IOU} = \frac{\textit{Intersection area}}{\textit{Union area}}$$

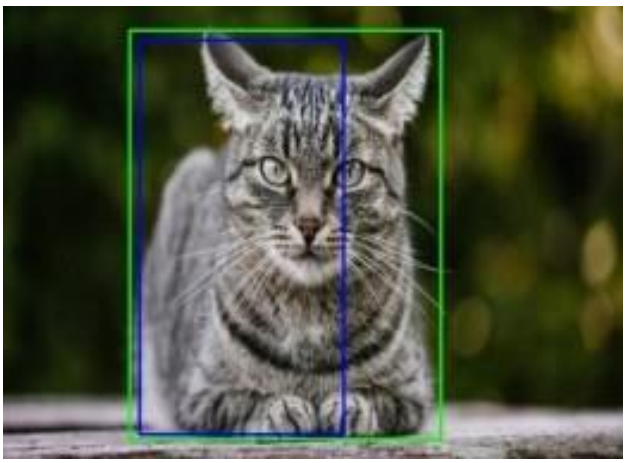

Fig. 5. Intersection over union

The fig-5 showing two bounding boxes, one in blue and one in green. As you can see, the blue box represents the predicted outcome and the green box represents the actual outcome. As long as both bounding boxes are equal, YOLO will ensure a fair result.

D. YOLO V7

By introducing several architectural reforms, YOLOv7 increases speed and accuracy. The different versions of YOLO differ from YOLOV7 by its architecture. The architectural reforms present in YOLOV7 are:

- E-ELAN
- Model Scaling for Concatenation-based Models

In YOLOv7's backbone, the computational block is called E-ELAN, which stands for Extended Efficient Layer Aggre- gation Network. In YOLOv7, the E-ELAN architecture allows the model to learn more effectively by using "expand, shuffle, merge cardinality" to continuously improve the network's learning ability without destroying its gradient path.

Scaled models are designed to meet the needs of different applications by adjusting key attributes of the model. Scaling a model can optimize its width (number of channels), depth (number of stages), and resolution (size of the input image).

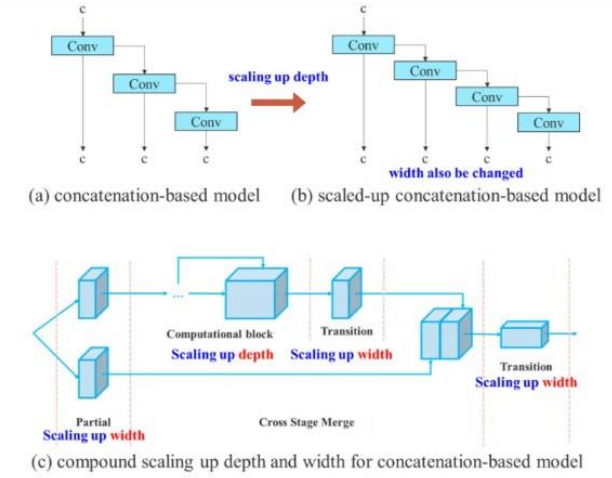

Fig. 6. Model Scaling

E. Open CV

OpenCV which is a Python library, allows to perform image processing and computer vision tasks. Currently, OpenCV plays a major role in real-time operation, which is crucial in today's systems because it is a huge open-source library for computer vision, machine learning, and image processing. The OpenCV library captures images and videos in the BGR format for 8-bit unsigned integers.The captured images can be divided into 3 matrices, BLUE, GREEN, and RED (hence the name BGR) with integer values ranging from 0 to 255. These pixels are so small in genuine pictures that the natural eye cannot distinguish them. It can identify objects, faces, and even human handwriting from images and videos.

F. Voice Generation

When the system detects the desired object, a voice is generated to mention the detected object. An essential component of voice generation is PYTTSX3. Pyttsx3 is a Python library for converting text to speech. For voice alerts, we also used Google Text to Speech (GTTS). There are a wide variety of English accents that Google Text to Speech contains for users from all over the world. In addition to being very easy to use, it converts the text into audio files that can be saved as mp3 files. It also supports many regional languages, which is helpful to those who are not fluent in English.

V. Block Diagram

The working of the system is represented in the below block diagram. The input image is taken from the user's camera. The system checks if any objects are detected in the image using YOLO V7 algorithm. If an object is detected, the system identifies the object by classifying the object categories. Then bounding boxes are created around the objects along with the object name. Later the object name is converted to speech. Then the system generates an audio output of the identified object using gTTS.

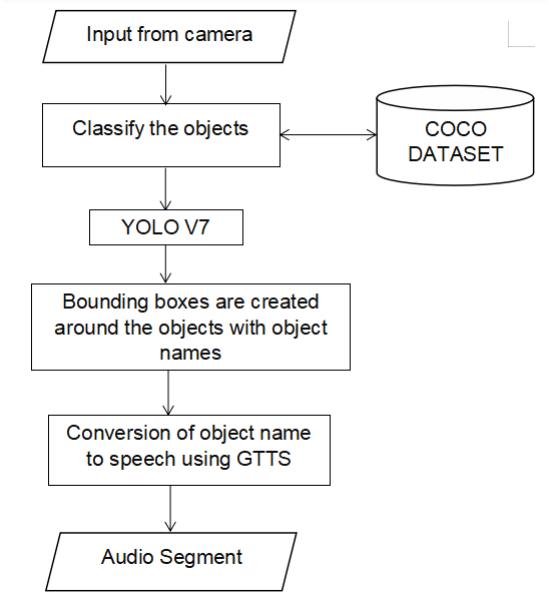

Fig. 7. Block Diagram

VI. Experimental results

A object detection system based on YOLOV7 using deep learning is designed for object recognition in a image or a video and then converting the object names to speech. Here, are few reults of object detection.

Fig. 8. A Remote and mouse are detected

Fig. 9. A plotted plant is detected

Fig. 10. A person, backpack and bicycle are detected

It can not only detect single object but also multiple objects present in the image. After detection of objects in the image, the name of the particular object along with the probablity score will be displayed. Hence, we get detected objects along with its object name. Therefore, the object names are converted to speech. In the audio, we can hear the object names present in the image.

VII. Other Applications of Object Detection

- **Face and person detection:-** The majority of face recognition systems are powered by object detection. To identify an individual from a group of people, it can be used to detect faces, classify emotions, and feed the results to an image-retrieval system. The most common use case for object detection is face detection, and you probably already use it to unlock your phone. A person

detection system can also be used in retail stores to count people or ensure social distancing metrics.

- **Intelligent video analytics:-** In intelligent video analytics (IVA), object detection is used wherever CCTV cameras are present in retail venues to understand customer in- teractions. An anonymization pipeline blurs out people's faces and de- identifies individuals in these video streams. When using IVA, some use cases preserve privacy by only looking at people's shoes, placing cameras below knee level, and ensuring the system captures a person's presence without looking at their identifiable features directly. It is commonly used in factories, airports, and transportation hubs to track queue lengths and access to restricted areas.
- **Performing a defect inspection:-** Object detection can be used by manufacturers to spot production line defects. A neural network can be trained to detect minute defects in fabrics, injection molded plastics, or even folds in fab- ric. The deep learning approach in object detection is able to detect defects in objects with heavy variations, such as food, and can do so more accurately than traditional machine learning methods.
- **Autonomous Driving:-** A self-driving car relies on object detection to recognize pedestrians, traffic signs, and other vehicles. Now a days, AI utilizes object detection to check the environmental and surrounding threats, such as oncoming vehicles or obstacles.
- **Medical Diagnosis:-** In the medical field, object detec- tion has led to many breakthroughs. Object detection using CT and MRI scans has become extremely useful for diagnosing diseases due to the heavy reliance on images, scans, and photographs in medical diagnostics. It is also used to detect the X-Ray reports and brain tumours.

VIII. Conclusion and Future work

In this proposed model we used image, voice generation modules for the development of the model. Our model in- tended to achieve real-time Object Detection using YOLO algorithm with Voice Feedback. As of now accuracy is good but in case if we want to increase the accuracy we have to train the model with more object/images in the dataset. We can further enhance this model by adding a Facial Recognition system to it and also to locate the exact position of the object/person which will help identify people and the exact location of them in an image and relay it to a visually challenged person.

References

[1] Prachi Tijare, Pranali Warkhede, Lina Godbole, Sayali Thakre, Rohini Dhakate and Ankit Mahule. " Object Detection, Convert Object name to text and text to speech", International Journal of Innovative Research in Engineering, Vol 3, PP: 45-51, April 2022.
[2] Rajeshwar Kumar Dewangan, Dr. Siddharth Chaubey. "Object Detection System with Voice Output using Python", IJRTI, Vol. 6, Issue 3, 2456- 3315, 2021.
[3] Punyaslok Sarkar, Anjali Gupta. "Object Recognition with Text and Vo- cal Representation", by ResearchGate, DOI: 10.9790/9622-1005046377,May 2022.
[4] Zhong-Qiu Zhao, Shou-tao Xu, and Xindong Wu. "Object Detection with Deep Learning: A Review", by IEEE transactions on neural networks and learning systems for publicaton, 1807.05511v2, Apr 2019.
[5] Jun Deng, Xiaojing Xuan, Weifeng Wang, Zhao Li, Hanwen Yao and Zhiqiang Wang. "A review of research on object detection based on deep learning", by Journal of Physics: Conference Series, DOI: 10.1088/1742- 6596/1684/1/012028, 2020.
[6] Meian Li, Haojie Zhu, Hao Chen, Lixia Xue and Tian Gao. "Research on Object Detection Algorithm Based on Deep Learning", by Journal of Physics: Conference Series, DOI:10.1088/1742-6596/1995/1/012046,2021
[7] Sunil and Gagandeep. "Study of Object Detection Methods and Appli- cations on Digital Images", IJSDR, Vol. 4, Issue 5, May 2019.
[8] Ajeet Ram Pathak, Manjusha Pandey and Siddharth Rautaray. "Appli- cation of Deep Learning for Object Detection", by ScienceDirect, Vol. 132, 1706-1717, 2018.
[9] Pavuluri Jithendra, Tummala Vinay Sai, Raj Kumar Mannam, Ramini Manideep and Shahana Bano. "Cognitive Model for Ob- ject Detection based on Speech-to-Text Conversion", by IEEE, DOI: 10.1109/ICISS49785.2020.9315985, Jan 2018.
[10] Okeke Stephen, Deepanjali Mishra and Mangal Sain. "Real Time object detection and multilingual speech synthesis", by IEEE, DOI: 10.1109/ICCCNT45670.2019.8944591, December 2019.
[11] A Mahesh Babu, CH Rithwik and R Sumukh. "Object Detection and converting text to speech", International Research Journal of Modern- ization in Engineering Technology and Science, Vol. 04, Issue. 06, June-2022.
[12] Prinsi Patel and Prof. Barkha Bhavsar. "A Survey on Object Detection with Voice", Research Journal of Engineering and Technology (IRJET), Volume: 08, Issue: 03, Mar 2021.