REAL-TIME OBJECT DETECTION FOR VISUALLY IMPAIRED PEOPLE

[1]Devayani T, [2]Yasmine S.K.A

[1]Student, [2]Assistant Professor
[1]Department of Computer Science
Auxilium College (Autonomous), Vellore – 632006

Abstract: **Those who are visually impaired (VIPs) make up a sizeable segment of the population and can be found everywhere in the world. Technology has recently demonstrated its presence in every field, and cutting-edge gadgets help people in their daily lives. Our work creates a clever, intelligent system for VIPs to aid movement and guarantee their safety. The suggested method uses an automated voice to deliver real-time navigation. VIPs are able to sense and comprehend their surroundings even when they cannot see the objects around them. Also, a web-based application is created to guarantee their security. In order to develop a clever and effective object detection system, the work proposed an Android application and Convolutional Neural Network (CNN).**

Index Terms**: Object detection, Recognition, Convolutional Neural Network, YOLO algorithm, Android application.**

1. INTRODUCTION

The World Health Organization (WHO) has reported that 285 million of the world's population is eyeless or visually bloodied. Out of these, 39 million are eyeless. The major conditions that beget visual impairments include refractive error, glaucoma, trachoma, corneal, darkness, cataracts, diabetic retinopathy, and unaddressed diplopia. Visually disabled Persons(superstars) face difficulties in performing conditioning of diurnal living (ADLs)e.g. occasion of work and training, moving in their surroundings, capability to interact with the terrain, and searching for common objects (inner/out-of-door) at their own or indeed with some backing. The main challenges for superstars are/ object discovery and recognition, currency identification, textual information (sign, symbol) and restatement, mobility/ navigation and safety.

Vision impaired persons typically rely on various forms of assistance, such as a white cane, other people's knowledge, trained dogs, etc. Many VIPs rely on canines or walking sticks to get around. A guide dog is taught to help its owners avoid mishaps caused by objects and barriers while travelling along a fixed path or in a fixed location. When using a walking stick, a person who is blind waves the staff in front of the obstacles in his path to locate them.

VIPs may now undertake tasks that they were previously unable to do thanks to a variety of assistive technology-related strategies, methods, equipment, and programmes that have been developed in the past. These solutions often consist of electronic gadgets with cameras, sensors, and microprocessors that can decide and give the user input via touch or sound. Many of the current item identification and recognition technologies are unable to provide high accuracy and the details required for their safe movement.

In order to give visually impaired persons with a smart and secure way of life, a smart object detection system built on Android applications and Convolutional Neural Networks (CNN) is developed. The major goal is to create a system for VIPs that uses a deep learning architecture for real-time object identification and recognition. It says the names of distant things that are present in the current frame and that can be viewed through the camera's lens. Moreover, it offers voice access to locations. It located an object that was nearby and 5 metres away. Moreover, the object will be kept in a database for future usage.

2. LITERATURE SURVEY

The approach described in this work creates sparse judgement DAGs (directed acyclic graphs) using a set of basic classifiers supplied by an external learning technique like AdaBoost. Casting the DAG design task as a Markov decision process is the essential idea. Based on the status of the classifier being produced at the time, each instance can choose whether to use or skip each base classifier. As a result, the base classifiers are chosen in a data-dependent manner, resulting in a sparse decision DAG.

One hyperparameter that clearly controls the accuracy/speed trade-off is used in the method. On three object-detection benchmarks, the technique is competitive with cutting-edge cascade detectors, and it outperforms them when there are few base classifiers. [1]

Modern object detection performance is achieved by deformable part-based models [1, 2], however because to the non-convex cost function optimisation during training, these models rely on heuristic initialization. This study explores the initialization's limits and improves on past approaches by adding more supervision. In terms of annotated object parts, it explores strong supervision and makes use of it to I enhance model initialization, (ii) enhance model structure, and

(iii)handle partial occlusions. This method has been proven to benefit from semi-supervised learning setups where part-level annotation is only provided for a portion of positive examples, and it can deal with imperfect and incomplete annotations of object parts. Results from experiments to identify six animal classes in the PASCAL VOC 2007 and 2010 datasets are provided. It shows

important advancements. [2]

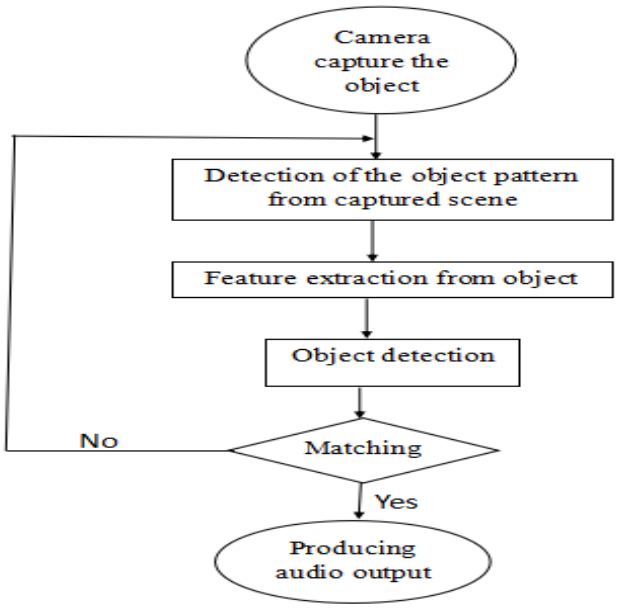
Grayscale photographs still exist, and they can be used to examine the difficulty of finding things. The creation of a learning-based solution to the issue that makes use of a sparse, part-based representation is the main focus. Images of the object class of interest are used to automatically create a vocabulary of distinctive object parts. Images are then represented using parts from this vocabulary together with spatial interactions observed among the parts. A learning technique is used to automatically learn to find instances of the object class in new photos based on this representation. [3]

2.1 EXSISTING METHODS

A raspberry pi digital signal processing (DSP) board is used in the current system. The object detection and recognition module receives a live feed from the video camera and transmits it to the DSP. The output of the object detection will be text. The system is lightweight, tiny in size, and detects in real time, however it cannot classify most everyday things.

The system uses unsupervised deep neural networks to extract global picture features, while supervised DNNs are not employed to extract local attributes. One of the most difficult aspects of object detection is that an object might appear radically different from different angles.

ARCHITECTURE DIAGRAM



3 METHODOLOGY

3.1 PREPROCESSING

The first step is to take a picture. The image is captured with a camera. Preprocessing enhances image intensity by suppressing undesired characteristics or increasing them for later processing. It resizes the image to 448*448 while also normalising the contrast and brightness.

3.2 DETECTION OF OBJECTS

Object detection is the act of locating and recognising real-world object instances in photos or videos, such as a car, bike, TV, flowers, and persons. It is commonly used in image retrieval, security, surveillance, and advanced driver assistance systems. (ADAS). Pre-processing, segmentation, foreground and background extraction, and feature extraction can all be used to find objects in a video stream.

Convolutional Neural Networks (CNN) are capable of detecting and classifying objects in images. Convolutional Neural Network (CNN) is a deep learning algorithm that was created specifically for working with images and videos. It takes photographs as input, extracts and learns the image's attributes, and then classifies the images based on the learnt features.

This algorithm was inspired by the operation of the visual cortex in the human brain. The visual Cortex is a portion of the human brain that processes visual information from the outside environment. It has several layers, each with its own function, such as extracting some information from an image or any visual and then combining all of the information collected from each layer.Similarly, CNN has various filters, and each filter extracts some information from the image such as edges, different kinds of shapes (vertical, horizontal, round), and then all of these are combined to identify the image.

The basic architecture of CNN comprises of different layers as,

- Input layer
- Convolutional layer
- Pooling layer
- Output layer

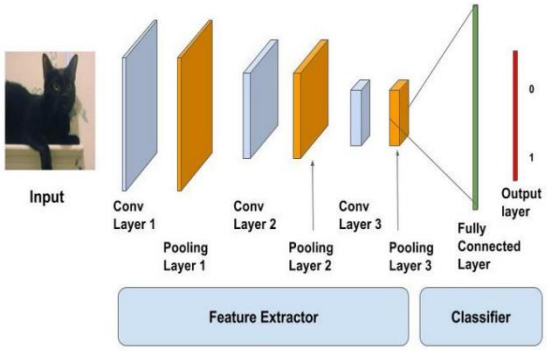


Fig. 1 Convolutional Layer

3.3 INPUT LAYER

As the name says, it is input image and it can be grayscale or RGB. Every image is made up of pixels that range from 0 to 255. It need to normalize them i.e convert the range between 0 to 1 before passing it to the model.

Below is the example of an input image of size 4*4 and has 3 channels i.e RGB and pixel values.

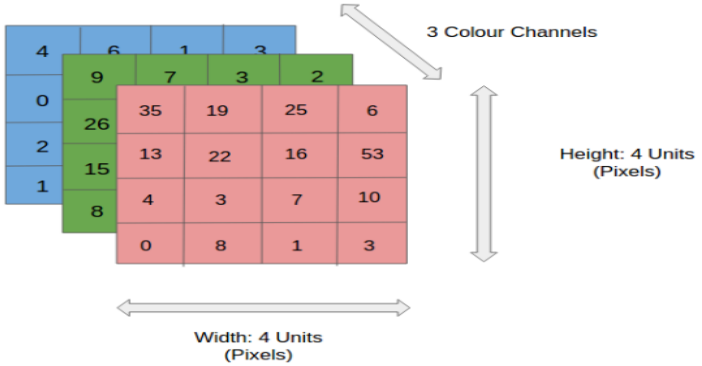


Fig. 2 Input Layer

3.4 CONVOLUTIONAL LAYER

The convolution layer is the layer where the filter is applied to the input image to extract or detect its features. A filter is applied to the image multiple times and creates a feature map which helps in classifying the input image. Let's understand this with the help of an example.

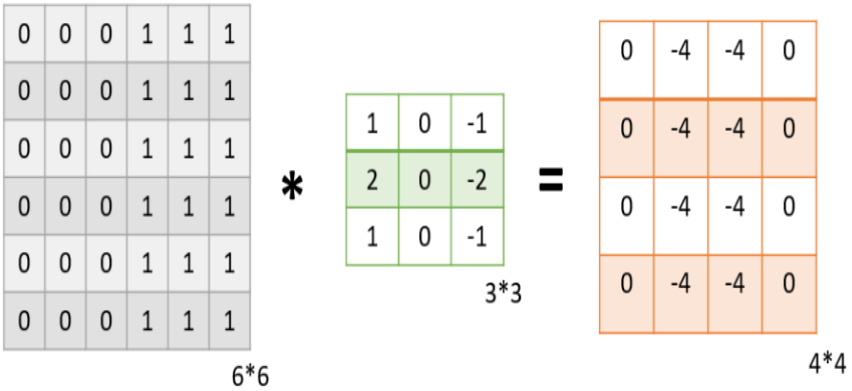


Fig. 3 Convolutional Layer

In the above figure, there is an input image of size 6*6 and applied a filter of 3*3 on it to detect some features. In this example, it have only one filter but in practice, many such filters are applied to extract information from the image.
The result of applying the filter to the image is a feature map of 4*4 which has some information about the input image. Many such feature maps are generated in practical applications. Let's get into some maths behind getting the feature map in the above image.

3 .5 POOLING LAYER

The pooling layer is applied after the Convolutional layer and is used to reduce the dimensions of the feature map which helps in preserving the important information or features of the input image and reduces the computation time.

Using pooling, a lower resolution version of input is created that still contains the large or important elements of the input image.

The most common types of Pooling are Max Pooling and Average Pooling. The below figure shows how Max Pooling works. Using the Feature map which is got from the above example to apply Pooling. Here they are using a Pooling layer of size 2*2 with a stride of 2.

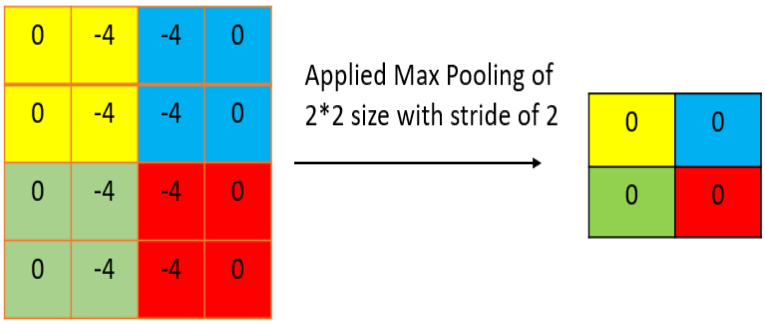


Fig. 4 Pooling Layer

3.6 OUTPUT LAYER

The Fully connected layer is used for classifying the input image into a label. This layer connects the information extracted from the previous steps (i.e Convolution layer and Pooling layers) to the output layer and eventually classifies the input into the desired label.

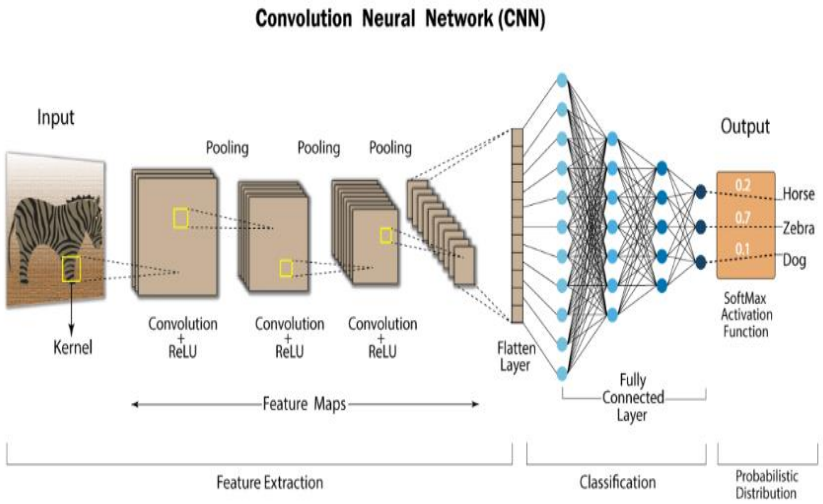


Fig. 5 Complete Form of CNN

3.7 OBJECT RECOGNITION

The YOLO algorithm is used for object recognition. The YOLO algorithm takes an image as input and then detects objects in the image using a simple deep convolutional neural network. The CNN model's architecture serves as the foundation of YOLO. Classification is one of the most dynamic research and application areas, according to YOLO V7. YOLO V7 is an Artificial Intelligence branch. (AI). The YOLO V7 technique was used to train the neural network. The influence of various function combinations while utilising YOLO V7 as a classifier is explored, and the correctness of these functions is analysed for various types of datasets. With the right combination of training, learning, and transfer, the YOLO V7 can be a highly effective tool for dataset classification.

YOLO algorithm has the following steps:

- They split the image into an S*S grid
- Each cell predicts B boxes(x,y,w,h) and confidences of each box: P(Object)
- Each cell predicts boxes and confidences: P(Object)
- Each cell also predicts a class probability.
- Conditioned on object: P(Car | Object)
- Then, combine the box and class predictions.
- Finally, they do threshold detections and NMS

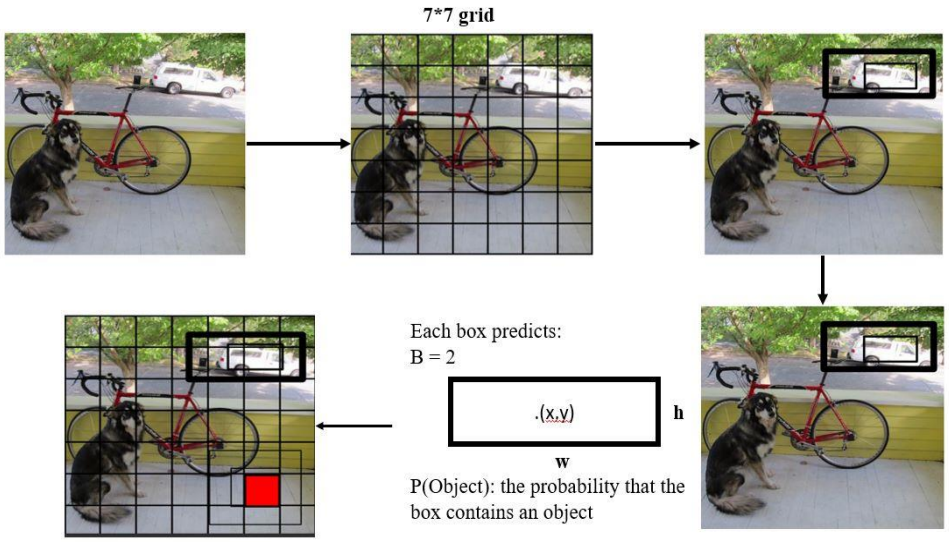


Fig. 6 Object Detection Through YOLO Algorithm

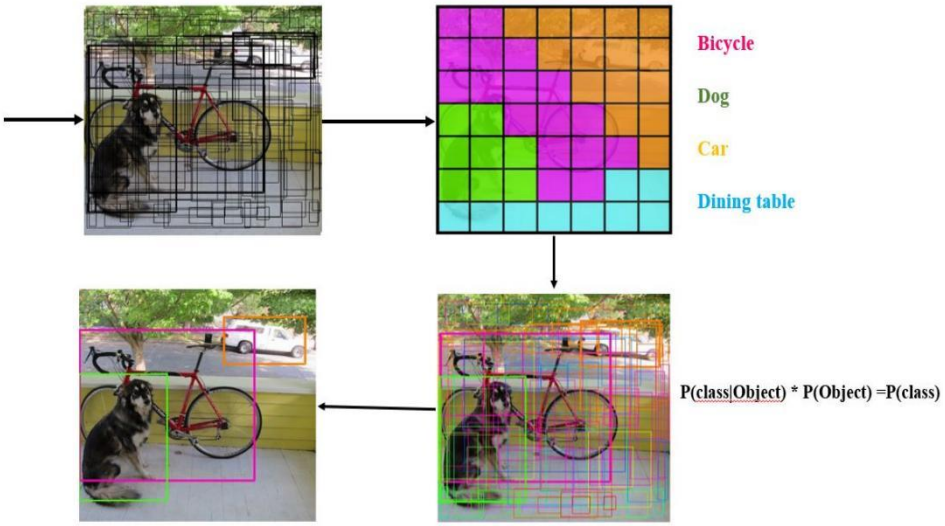


Fig. 7 Object Recognition Through YOLO Algorithm

3.8 TEXT CONVERSION

Optical Character Recognition (OCR) is the process that converts an image of text into a machine-readable text format.

3.9 TEXT-TO-SPEECH CONVERSION

The text processing component's aim is to process the provided input text and generate a suitable phonemic unit sequence. First, the incoming text is parsed. The text-to-speech system (TTS) uses a speech synthesiser to turn text into voice. Java includes a voice API for incorporating speech technology into user interfaces. It defines a platform-independent API for command and control recognizers, dictation systems, and speech synthesisers. It is not included in JDK. It is a third-party speech API designed to promote the availability of diverse implementations. It artificially creates a human voice. A speech synthesiser is a computer system that is used for this purpose. It operates without a hiccup.

4. CONCLUSION

This project briefly describes the Android application and current object detection techniques, such as the CNN family and YOLO. YOLO is a model for unified object recognition. It is straightforward to build and can be trained immediately on full-frame photos. In contrast to classifier-based techniques, YOLO is trained on a loss function that directly corresponds to detection performance, and hence the entire model is learned concurrently. The quickest general-purpose object detector is Fast YOLO. And, in comparison to other detection algorithms, YOLOV2 offers the best balance between real-time speed and good accuracy for object detection across a wide range of detection datasets. Furthermore, because YOLO has a better generalising representation of objects than other

models, it is suited for applications that require fast, robust object recognition. These outstanding and valuable benefits make it worthy of being strongly promoted and popularised. The scope of the dataset is the most pressing next obstacle for machine learning, aside from the structure of each algorithm. The availability of appropriate training data may be a critical component in the learning process in order to achieve desired results.
In the future, this project can be enhanced with other features such as email with voice capability, recognising location, and so on for VIP.A large amount of dataset may be incorporated, and the object detection distance will be extended as well.

REFERENCES:

1. W. Elmannai and K. Elleithy, ‘‘*Sensor-based assistive devices for visually impaired people: Current status, challenges, and future directions*,’’ Sensors, vol. 17, no. 3, p. 565, 2017.
2. ToufiqP. Ahmed Egammal and Anurag mittal (2006),*”A Framework for feature selection for Background Subtraction”,* in Proceedings of IEEE computer Society Conference on Computer Vision and Pattern Recognition.
3. R. Velázquez, *‘‘Wearable assistive devices for the blind,”* in Wearable and Autonomous Biomedical Devices and Systems for Smart Environment. Berlin, Germany: Springer, 2010.
4. L. B. Neto, F. Grijalva, V. R. M. L. Maike, L. C. Martini, D. Florencio, M. C. C. Baranauskas, A. Rocha, and S. Goldenstein, *‘‘A Kinect-based wearable face recognition system to aid visually impaired users,*’’ IEEE Trans. Human-Mach. Syst., vol. 47, no. 1, pp. 52–64, Feb. 2017.
5. C. Shah, M. Bouzit, M. Youssef, and L. Vasquez, *‘‘Evaluation of RU-netratactile feedback navigation system for the visually impaired*,’’ in Proc. Int. Workshop Virtual Rehabil., 2006, pp. 72–77.
6. M. R. U. Saputra, Widyawan, and P. I. Santosa, *‘‘Obstacle avoidance for visually impaired using auto-adaptive thresholding on Kinect’s depth image*,’’ in Proc. IEEE 11th Int. Conf. Ubiquitous Intell. Comput., IEEE 11th Int. Conf. Auton. Trusted Comput., IEEE 14th Int. Conf. Scalable Comput. Commun. Associated Workshops, Dec. 2014, pp. 337–342.
7. Y. Yi and L. Dong, ‘‘*A design of blind-guide crutch based on multisensors*,’’ in Proc. 12th Int. Conf. Fuzzy Syst. Knowl. Discovery (FSKD), Aug. 2015, pp. 2288–2292.