JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

EYEMATE – Object Detection with Voice Commands for Visually Impaired

[1]Sagar Sharma, [2]Saransh Aggarwal, [3]Sandeep Kumar
[1]B. Tech Student, [2]B. Tech Student, [3]Associate Professor
[1]Computer Science and Engineering,
[1]Sharda University, Greater Noida (U.P), India

Abstract: Object detection is used in almost all real-world applications, such as autonomous navigation, visual systems, face recognition, and many more. Instead of using object detection techniques, it aids in identifying an image's contrasting features and generates a knowledgeable and competent comprehension, much like how human vision functions. This study begins with a succinct introduction to deep learning, which has produced noteworthy performance as a result of these visual representations. This clarifies the function of deep learning on convolutional neural networks (RCNNs) for object detection algorithms that assist the blind by identifying objects by name. EYEMATE's primary role is to record live visual input utilizing a smartphone or portable camera. A complex object detection algorithm then processes this input. This algorithm is capable of identifying a broad variety of items that are frequently seen in daily life, including obstacles, traffic signs, products on store shelves, and more. Modern machine learning techniques are used by the system to increase its accuracy and extend its object detection capabilities. EYEMATE incorporates a natural language processing component that allows users to give voice commands in order to promote smooth interaction. Users can ask the system to describe their surroundings, identify specific things, or offer assistance with activities like navigating unfamiliar places by speaking commands to it. Users are guaranteed timely and pertinent information by means of synthesized speech or haptic feedback as the system's reaction.

Index Terms **: Object Detection, RCNN, Neural Network, Deep Learning**

I. INTRODUCTION

Visual impairment can pose significant challenges to individuals, affecting their ability to navigate and interact with the world around them. The simplest tasks that many take for granted, such as recognizing objects, reading signs, or avoiding obstacles, can be immensely challenging for those with visual impairments [4]. Assistive technologies have played a pivotal role in mitigating these challenges, striving to empower visually impaired individuals and enhance their quality of life.[8] One such groundbreaking technology is EYEMATE, a comprehensive system that combines the capabilities of object detection and voice commands to create a powerful tool for those who are visually impaired. EYEMATE represents a significant step forward in the quest for accessibility and independence for the visually impaired. By fusing cutting-edge computer vision techniques with natural language processing, this system aims to provide real-time visual assistance and information to users, helping them better understand and interact with their surroundings.[10] In this introduction, we will explore the critical need for such a technology, the core features of EYEMATE, and its potential to revolutionize the way visually impaired individuals experience the world. Advancements advances in technology have led to the creation of a wide range of assistive tools and software intended to meet the particular requirements of people who are blind or visually impaired. The ability of text-to-speech software, braille displays, and navigation apps to assist users in understanding and interacting with the visual world has previously been proven [26] However, by emphasizing real-time object detection and natural language interaction, EYEMATE advances this assistive technology. Using voice recognition, machine learning, and artificial intelligence, EYEMATE offers a comprehensive solution.. At its core, the system uses a camera or smartphone to capture the user's surroundings.[26] These visuals are then processed by a state-of-the-art object detection algorithm that can identify a wide range of objects, such as obstacles, street signs, products on store shelves, and more. What sets EYEMATE apart is its integration of natural language processing, allowing users to issue voice commands, inquire about their environment, and request object identification. The system reacts by giving the user immediate access to contextually relevant information through synthesized speech or haptic feedback. The fundamental purpose of object detection in computer vision is to determine the location and existence of specific items in a scene. It can be applied to a wide range of real-world scenarios, such as autonomous vehicles crossing difficult terrain and intelligent surveillance systems that guarantee public safety. Adding voice input greatly advances this technique. Imagine an interface that is easy to use for users of all skill levels, one that not only recognize items but also reports its findings verbally. As we go into the mechanics of the process, we will examine the algorithms, neural networks, and data-driven techniques that enable object detection with voice feedback. Also, we'll examine the mutually advantageous relationship between natural language processing and computer vision, demonstrating how these fields work together to generate comprehensive and important results. We'll also see how the promise of this technology extends well beyond comfort, transforming entire sectors, enhancing accessibility, and altering the nature of human-machine interaction. The possibilities are as groundbreaking as they are

diverse, ranging from using augmented reality to change retail experiences to helping the blind navigate their environment. So settle down for a thrilling ride as we explore the fascinating field of voice feedback object detection. The hypothesis states that this study will uncover a hitherto undiscovered aspect of human-machine interaction—one that combines sound and vision to create a truly immersive and transformative experience—which will lay the groundwork for real-world implementations.Finding the presence and placement of particular items in a scene is the basic function of object detection in computer vision. It has a wide range of practical uses, from intelligent surveillance systems that ensure public safety to autonomous cars navigating challenging terrain. This technique advances significantly when voice feedback is added. Imagine a system that not only recognises items but also verbally reports its findings, resulting in a simple-to-use interface for users of all expertise levels. We will explore the algorithms, neural networks, and data-driven methods that enable object detection with voice feedback as we delve into the mechanics of the process. We'll also look at the mutually beneficial interaction between computer vision and natural language processing, revealing how these disciplines cooperate to produce thorough and significant findings. Additionally we'll see how this technology's promise goes far beyond the realm of convenience, revolutionising entire industries, improving accessibility, and changing how people and machines interact. The possibilities are as varied as they are ground-breaking, from assisting the blind in navigating their surroundings to transforming retail experiences through augmented reality. So grab a seat as we set out on an exciting voyage into the world of voice feedback object detection. According to the theory This investigation is expected to reveal a new dimension in human-machine interaction—one that blends sight and sound to produce a genuinely immersive and transforming experience—that serves as the foundation for practical applications.It is a holistic solution that harnesses the power of artificial intelligence, machine learning, and voice recognition. At its core, the system uses a camera or smartphone to capture the user's surroundings. These visuals are then processed by a state-of-the-art object detection algorithm that can identify a wide range of objects, such as obstacles, street signs, products on store shelves, and more. What sets it apart is its integration of natural language processing, allowing users to issue voice commands, inquire about their environment, and request object identification. The system responds with synthesized speech or haptic feedback, providing instant and contextually relevant information to the user.

II. Literature Review

LoCAR: Low-Cost Autonomous Robot for Mobile Nets-Based Object Detection and Voice Command The Low-Cost Autonomous Robot (LoCAR) for object identification with voice command is designed and implemented in this work (Figure 1). Upon receiving an instruction, the gadget searches its surroundings for the specified location and proceeds in that direction. AI Vision [1]: Smart speakers are designed with advanced voice interaction and customized object detection capabilities. The cutting-edge "AI Vision Smart Speaker" combines sophisticated voice interaction features with specialized object identification abilities to produce a singular and engrossing user experience. This intelligent speaker uses AI-powered visual technologies to communicate with the real world, going beyond conventional voice assistants. [2] Blind Voice Guidance Object Detection System People with visual impairments can feel more secure and independent thanks to the "Detection System for the Blind with Voice Guidance."

This technology helps users successfully navigate their surroundings by using sophisticated computer vision algorithms to identify things in the surrounding area and deliver useful audio cues.[/3] Microsoft Holo Lens: 3D Audio Localization and Object Detection The "Detection of Objects" "Microsoft HoloLens Featuring 3D Audio Localization" is an innovative innovation that improves the augmented a reality experience through the use of real-time 3D audio and object recognition localization with the HoloLens from Microsoft voice command labelling in the platform. [4] Convolutional Neural Nets for Combinations of Realistic Image Identification The framework integrates CNN features with friendly voice labelling to boost user interaction and identification of the object precision. It can be utilised for several purposes, and developments in the future may make even more adaptable and efficient system. [5] Object Detection and Voice Guidance The Smart App for the Visually Impaired Approach This project radically alters The manner in which blind people engage with their environment. Using a combination of live object monitoring using more technologies, the Technology provides a welcoming and empowering remedy that improves movement and autonomy as well as establishes new chances for consumers in a variety of scenarios.[6] R-CNN Acceleration: Real-Time Object Identification In in order to offer a cohesive architecture For object detection, Quick R-CNN integrates a Proposal Network for Regions (RPN) with Quick CNN R. Region is created by the RPN. suggestions, which Fast R-CNN later Sorts and polishes.[7] A Single Exam: Instantaneous, Consolidated YOLO's cohesive architecture makes advantage of the entire picture to calculate the item and class probability. bounding boxes in just one assessment. That allows low-resource gadgets to recognize items in actual time.[9] One-shot Detector With Multi Box SSD, one-shot detection is possible. method for forecasting item boundary boxes and class rankings for various object sizes and ratios of aspects. This facilitates the ability to large range of sizes swiftly and accurately.[9] "Permit the Ill See" offers a workable solution to bridge the divide. between visual impairment and object recognition. The device has the potential to significantly improve the lives of people who are blind by the use of AI, instant response as well as speech translation. They will be able to communicate with their environment more effectively. Ten [] Thai Blind Using CLIP to Select an Object Speech-based detection and positioning Demand It processes real-time pictures. captured using a camera and the CLIP model for identifying objects. It is able to identify objects in the scene and offer a written account of the subject.11] Mask is based on Cover R-CNN Quicker R-CNN. R-CNN, which improves on example division by incorporating a second mask forecasting division. With concurrent forecasting bounding boxes, class labels, With object masks, producing precise segmenting an object instance. When [12] Dense Object Focal Loss in Retina Net A fresh loss function is proposed by Retina Net. concentrated loss, to get over the course imbalance in the way objects are detected. Dedicated loss enhances the performance of detecting, particularly for uncommon classes, by raising the weights allocated to difficult and incorrectly classified data.13] A YOLOv3 YOLOv3's Gradual Advancement gets better. on YOLO by making use of a larger both in terms of design and multiple detection balances. This improves the detection of the model. and precision talents. In Scalable [14] as well as Effective Object: Effective Det suggests a mixture. scaling strategy that achieves a middle ground the relationship between the depth, width, and resolution to ensure maximum effectiveness and precision. To produce a cutting-edge item detecting outcomes, the method utilizes significantly fewer parameters.[15] In Triplets of Crucial Points for Center Net Identifying Items with Keypoint triads to represent objects, Center Net suggests a unique method

that integrates keypoint estimation and object identification simultaneously .At [16] Convolutional Nets that Can Be Deformed for Identification of Deformable Objects DCNs, or convolutional networks, are an superiority over conventional convolutions, since they include teachable countermeasures to alter the locations of the sample filters with convolutions. Now, the model can depict item alterations and deformations more precise.[17] CornerNet-Lite for Efficient Keypoint -Based Object Detection Maintaining its keypoint-based detection system approach and improving the architecture in order to identify objects in real time, CornerNet is enhanced by CornerNet-Lite. The. High efficiency is the outcome with little to no reduction in accuracy[18].

III. Methodology

The process of creating EYEMATE, an object detection system with voice instructions designed for people with vision impairments, is a multi-phase procedure that includes several critical steps. Project planning and definition, which clearly define the project's goals, scope, and the particular needs of the target user group, are crucial first steps. After that, it's time to set up the hardware and software, choose the right hardware, and install the required software tools. The crucial next stage is data collecting, which entails putting together a varied dataset of photos depicting objects that blind people frequently come across and painstakingly labeling and bounding boxing each image for object detection. The creation of the object detection model starts after the data collection stage is finished. This could entail building a bespoke model or modifying already-trained models, such as Faster R-CNN or YOLO[6], that have been refined with the use of the gathered dataset[19]. Metrics like Mean Average Precision are utilized to thoroughly evaluate the performance of the model (mAP). Concurrently, Automatic Speech Recognition (ASR) models or APIs such as Google's Speech Recognition are used to create the voice command recognition system. These models are trained on a specific dataset of voice commands.How YOLO work :-

1.**Residual Blocks**

Figure 1: Residual Blocks

The image is divided into various grids. Each grid has a dimension of S x S. The following image shows how an input image is divided into grids.

2. Bounding Box Regression- The following characteristics are present in each bounding box in the image: Length (bw) Altitude (bh)Class: The letter c stands for class (e.g., person, car, traffic signal, etc.). Box center bounding (bx, by)A bounding box example can be seen in the image below. A yellow outline has been used to depict the bounding box. YOLO uses a single bounding box regression to predict the height, width, center, and class of objects. In the image above, represents the probability of an object appearing in the bounding.

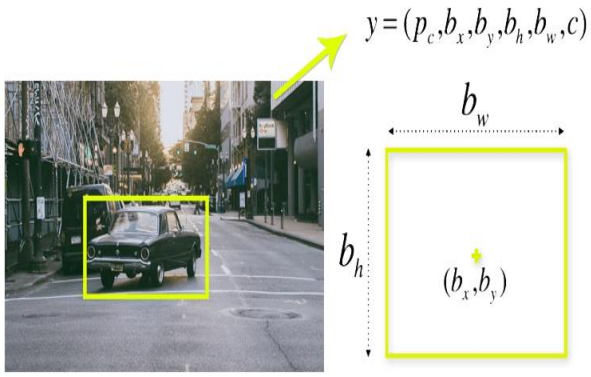

Figure 2: Bounding Box Regression

3.**Interaction over union** - In object detection, the phenomenon known as intersection over union (IOU) characterises how boxes overlap. YOLO creates an output box that precisely encircles the items by using IOU.The task of anticipating the bounding boxes and their confidence scores falls on each grid cell. If the expected and actual boundary boxes match, the IOU is equal to 1. Bounding boxes that are not equivalent to the actual box are removed using this procedure. The IOU is illustrated in the following illustration in a straightforward manner.

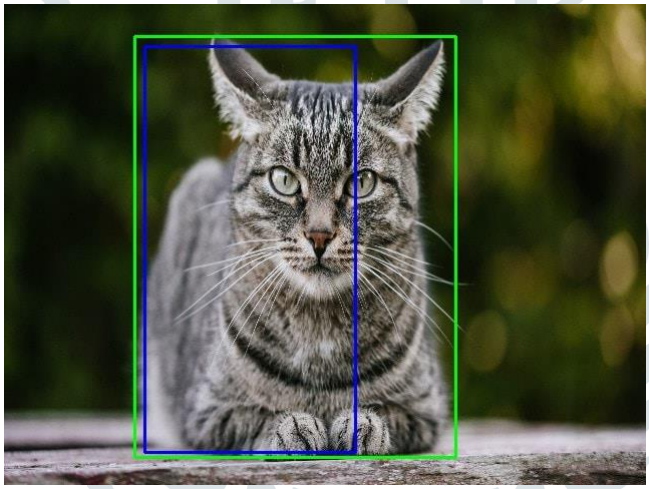

Figure: 3 Intersection of Union

Combination of all three steps - The image is first split up into grid cells. B bounding boxes are predicted by each grid cell, along with their confidence values. To determine each object's class, the cells make predictions about the class probabilities.A car, a dog, and a bicycle are just a few examples of the at least three classes of items that are visible. A single convolutional neural network is used to make all of the predictions at the same time.The predicted bounding boxes and the actual boxes of the items are guaranteed to be equal by intersection over union. This occurrence removes superfluous bounding boxes that don't match the object's dimensions (height and breadth). The final detection will be made up of distinct bounding boxes that precisely match the objects. For instance, the pink bounding box surrounds the vehicle. while the yellow border box encloses the bicycle. The blue bounding box has been used to highlight the dog.

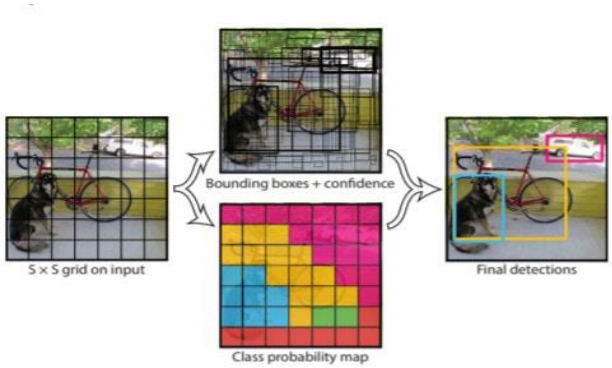

Figure 4: Architecture of YOLO Algorithm [26]

The voice command recognition and object detection systems are flawlessly synchronized throughout the crucial integration step. Interaction is made easier using an interface that is easy to use. The object detection model continuously examines real-time video stream while it is in use. When an object is detected, the user receives vocal feedback from the system that includes specifics about the object's type, position, and other pertinent information like distance.[17] Users can also give voice commands to the system, giving it instructions to carry out particular tasks like finding specific goods, identifying objects, or describing their surroundings. Subsequently, the system undergoes extensive user testing and feedback collecting with visually challenged individuals to evaluate its overall performance, accuracy, and usability. The system is then improved and refined using the input received. Features for customization and accessibility are combined to guarantee that the system accommodates the preferences of each user.[20] Users receive thorough documentation, and lines of help are set up to answer questions and fix problems.

Dataset: The COCO dataset comprises 330kimages and 80 labels. Among the dataset's labels are: -Animals-Flowers Traffic signals - Individuals-buildings -things like computers, smartphones and laptops.-Food items And much more in a similar manner...A. Figuring out where the image is: every identified item requires that we create a bounding box to enable determine the position of the image. Then, with reference to the photo frame, using the height and width of that particular enclosed space. The item position inside the bounding box is calculated with five values. The position of the object is shown by the initial four figures are bx, by, bw, and bh. The sixth figure, BC, shows how much of a box something is housed in. BC is calculated by multiplying the items discovered within the box and IOU Crossing Over Union. Given that tagged data should be utilized in training, the label "y" is recognized and is characterized as Triangle Similarity [11] The thing's both the width and the angle at the camera expand as it gets closer to the camera. The. Triangle similarity formula is (PxD)/W = F. When the system locates the object a user wants to find, voice guidance is a feature that readily gives information to certain users, such as those who are visually impaired[22].For voice feedback, the estimated portion of the object and the identified class label of the detected object are concatenated as text. Alpha Text-to-One can utilize speech API. Cloud TTS API for Google:- Users can combine hundreds of real voices with it.-It is translated into several languages. -The developer can communicate with users through many languages, applications, and devices, making it an easy-to-use API.

IV. Detailed Methodology

A low-level design would go into the specifics of how each element of the highlevel design is implemented in a real-time object detection system with voice feedback. I'll briefly describe the main elements and how they interact down below: connect to the camera or the source of the video feed.at a predetermined frame rate, record video frames. Pre-trained object identification models should be loaded, like YOLO or Faster RCNN.Set the model and related weights to their initial values.Set up the model for instantaneous inference. the input module should send you the video frames.Frames should be preprocessed (e.g., resized, normalised) to meet the model's input specifications.Send the object detection model preprocessed frames for inference. Utilise the loaded model to infer on the preprocessed frames. For detected objects, extract bounding boxes, class labels, and confidence ratings. Non-maximum suppression should be used to get rid of duplicate detections and boost accuracy. Use an object tracking method (such as the Kalman filter or SORT) to link items found in successive frames. Keep track of the movement of the objects and update their locations. Get data from the object detection and tracking modules regarding discovered objects, their locations, and any pertinent circumstances. Transform this data into written feedback. Input text into a Text-to-Speech (TTS) engine to provide speech feedback. Specify the reasoning behind when and what should be given as voice feedback. This reasoning can involve circumstances like object thresholds, object categories, or user preferences. Real-time object detection with voice feedback is an exciting application that combines computer vision and natural language processing. To develop such a system, you will need various tools, libraries, and formulae. Here's an outline of what you'll require: 1. Hardware: Computer: A powerful computer or server capable of handling real-time object detection tasks. A GPU (Graphics Processing Unit) is highly recommended for faster inference. Camera: A camera or webcam for capturing the real-time video feed. 2. Software and Libraries: Deep Learning Framework: Choose a deep learning framework like TensorFlow, PyTorch, or Keras for building and training your object detection model. Object Detection Model: You can use pretrained models like YOLO (You Only Look Once), Faster R-CNN, or SSD (Single Shot MultiBox Detector) to perform object detection. Speech Recognition Library: Use a speech recognition library like Google's SpeechRecognition or Mozilla's DeepSpeech for converting spoken words to text. Text-to-Speech (TTS) Library: Libraries like gTTS (Google Text-to-Speech) or pyttsx3 can be used for generating voice feedback. OpenCV: This library is useful for image and video manipulation, which you'll need for capturing frames from the camera and drawing bounding boxes around detected objects. Python: You'll likely use Python as the primary programming language for integrating all these components. 3. Data: Training Data: For training your object detection model, you'll need a labeled dataset with images containing the objects you want to detect, along with bounding box annotations. Speech Data: For voice feedback, you may need a dataset of audio clips containing various spoken words and phrases. Object Detection Algorithm: You don't necessarily need to derive new formulas, as object detection models come with their own architecture and formulas. However, you should understand how these models work, including anchor boxes, non-maximum suppression, and confidence scores. Speech Recognition: Some key techniques involve using Hidden Markov Models (HMMs) or deep learning models like recurrent neural networks (RNNs) or transformers for speech recognition. Text-to-Speech (TTS): TTS models are usually based on deep learning, such as WaveNet or Tacotron, which convert text into speech.

Formulae:

Precision = True Positives / (True Positives + False Positives)
Recall = True Positives / (True Positives + False Negatives)
F-Measure = (2 * Precision * Recall) / (Precision + Recall)
Confusion Matrix:

	Positive Prediction	Negative Prediction
Positive Class	True Positive (TP)	False Negative (FN)
Negative Class	False Positive (FP)	True Negative (TN)

VI. Result and Discussion

We have created an Android application for object detection as part of this investigation. The output is shown by the system as the object's name and its percentage probability. Therefore, only items with a probability higher than the specified threshold probability will be detected by the system. The mean accuracy chart prior to and following enhancement. Additionally, the system speaks text into audio using the Google API and uses the device's speakers to indicate an object's position in inches .Because little YOLO is employed to implement the system on the Android platform, object recognition accuracy is decreased.

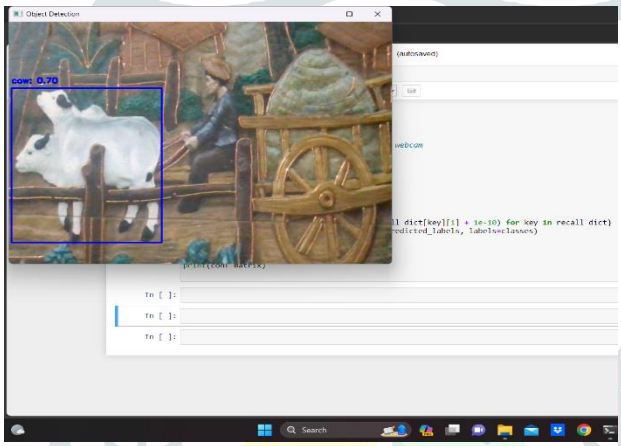

Figure 5: Detect the cow and give the name of that object

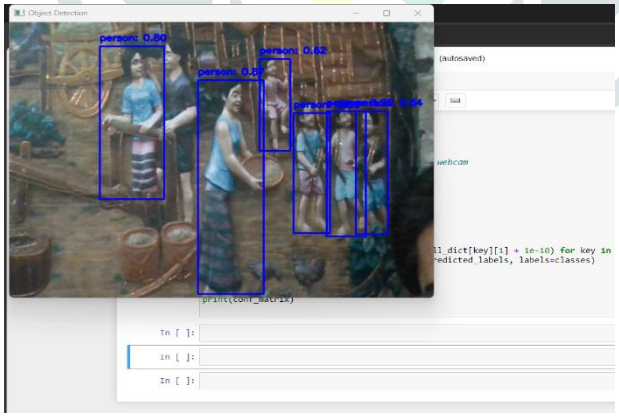

Figure 6: Detect The Persond and gives the name of the person.

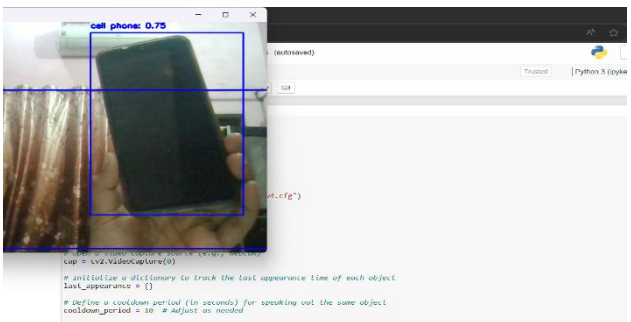

Figure 7 : Detect the cell phone and Speak the name of that object

VII. Comparision Table

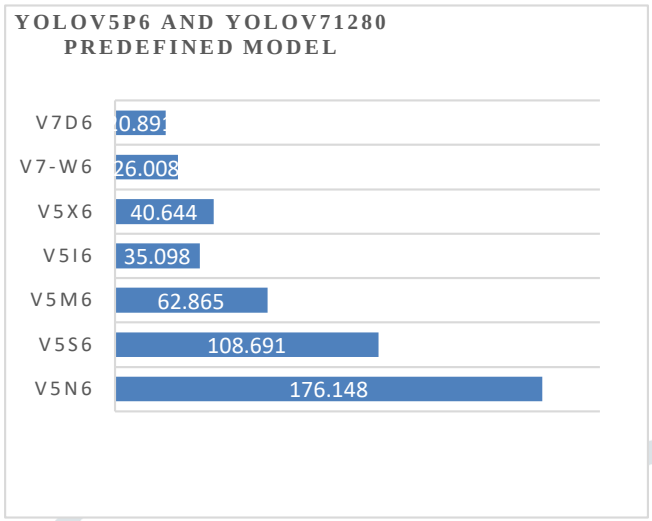

Figure 9: FPS performance comparison of YOLOv5 P6 and YOLOv7 models pretrained with 1280 image resolution.

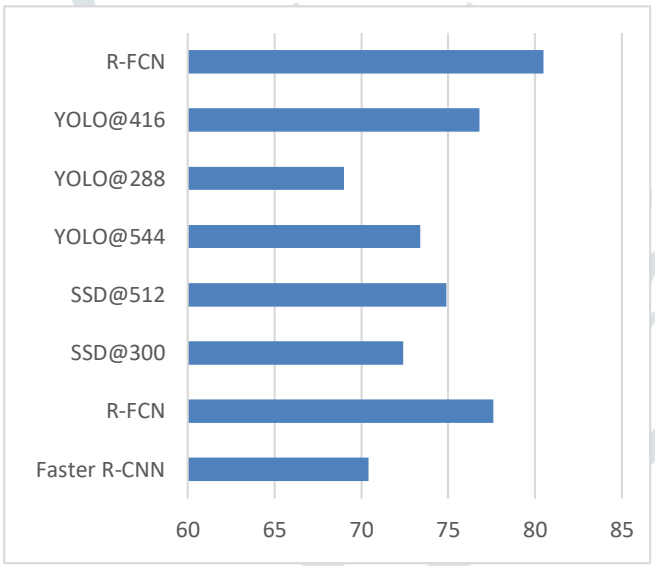

Figure10: Comparison of Algorithms

VIII. Conclusion and Future Work

The current system helps those who are blind. This system can be integrated into smartphones or other comparable handy gadgets. People with visual impairments will benefit from this system by being able to access item information such as name and position. Create an object detection system using deep learning algorithms, like YOLO, then use those same algorithms to estimate the position of objects. For individuals who are blind or visually challenged, this technology offers voice instruction. This system is specifically made to assist blind individuals in an efficient manner. On the other hand, accuracy can be raised. Additionally, the current system operates on the Android operating system, which may be improved to make it compatible with all convenient gadgets.

References

1. Abadi, M., et al.: TensorFlow: large-scale machine learning on heterogeneous systems Software.
2. Fu, K.S., Young, T.Y.: Handbook of Pattern Recognition and Image Processing. Academic Press(2020).
3. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2414–2423 (2017).
4. Hawkins, D.M.: The problem of overfitting. J. Chem. Inf. Comput. Sci. 44(1),.
5. Izadi, S., et al.: Kinectfusion: real-time 3D reconstruction and interaction using a moving depth camera. In: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, pp. 559–568. ACM.

6. Kingma, D.P., Ba, J.L.: Adam: a method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations.
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp.
8. Lawrence, S., Giles, C.L., Tsoi, A.C., Back, A.D.: Face recognition: a convolutional neural-network approach. IEEE Trans. Neural Network.
9. LeCun, Y., Bengio, Y., et al.: Convolutional networks for images, speech, and time series. handb. Brain Theory Neural Netw. 3361(10),
10. Maturana, D., Scherer, S.: Voxnet: a 3D convolutional neural network for real-time object recognition. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 922–928. IEEE .
11. Milgram, P., Takemura, H., Utsumi, A., Kishino, F.: Augmented reality: a class of displays on the reality-virtuality continuum. In: Telemanipulator and Telepresence Technologies, vol. 2351, pp. 282–293. International Society for Optics and Photonics(2017).
12. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1717–1724.
13. Rauschnabel, P.A., Ro, Y.K.: Augmented reality smart glasses: an investigation of technology acceptance drivers. Int. J. Technol. Mark. 11(2), 123–148(2019).
14. Rekimoto, J.: Matrix: a realtime object identification and registration method for augmented reality. In: Proceedings of the 3rd Asia Pacific Computer Human Interaction, pp. 63–68. IEEE (2018).
15. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI, vol. 4, p. 12 (2017).
16. Szegedy, C., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2019).
17. D. Tran, L. D. Bourdeva, R. Fergus, L. Torresani, and M. Paluri. “Learning spatiotemporal features with 3d con-volutional networks”. In ICCV, 2015
18. K. S. Varunn, I. Puneeth and T. P. Jacob, ”Hand Kinesis Recognition and Implementation for Disables using CNN’S,” 2019 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2019, pp. 0592- 0595, doi: 10.1109/ICCSP.2019.8697980.
19. Polat, Husiyin, and Homay Danaei Mehr.Object Detection Featuring 3D Audio Localization for Microsoft Holo Lens (2018).
20. Gibran Benitez-Garcia, Muhammad Haris, Yoshiyuki Tsuda, Norimichie Ukita, Convolutional Neural Networks for Image Recognition in Mixed Reality Using Voice Command Labeling (2019).
21. Mukul Nair, Sherly Noel, "Hand Kinesis: Object Detection and Voice Guidance for the Visually Impaired Using a Smart App9(2020).
22. Dimtris Nikos Fakotakis, Stavros Nousiass, Gerasimos Arvanitis, Evangelia I: Faster R-CNN: Towards Real-Time Object Detection In this study, Faster R-CNN, a deep learning framework for object detection. (2018)
23. LeCun, Y., Bengio, Y., et al.: You Only Look Once: Unified, Real-Time A real-time object detecting system is called YOLO with deep CNN (2017).
24. Fu, K.S., Young, T.Y.:Single Shot MultiBox Detector SSD is a one-time object detection method based on deep learniing.(2020)
25. Izadi, S., et al.:Object Detection and Position using CLIP with Thai Voice Command for Thai Visually Impaired (2022) IEEE.
26. Ying Ma, Tianpei Xu, Kanigchul Kim: RetinaNet: Focal Loss for Dense Object Focal loss is a feature of RetinaNet's dense object detecting system(2021) .
27. F. K. H. Quik,YOLOv3: An Incremental Improvement An updated version of the original YOLO architecture is presented in YOLOv3(2019).
28. Shreyashii Narayan Sawant and M. S. Kumbhar :CenterNet: Keypoint Triplets for Object detection A keypoint-based object detection framework is presented by CenterNet(2018).
29. K. S. Varunn, I. Puneeth and T. P. Jacob, ”Hand Kinesis:Deformable Convolutional Networks for Object Detection Deformable convolutional operation is introduced by DCN(2020).
30. Ansari, Aqueib, and Dushyant Kumar Singh: CornerNet-Lite: Efficient Keypoint Based Object Detectio CornerNet-Lite is a productive variant of objecct detection system(2018).