HIGHLY OPTIMIZED DETECTION USING ELECTRONIC RETINAS

(H. O. D. E. R.)

Parth Panchal
B.E Student,
Dept. Of Information Technology
Shah & Anchor Kutchhi Engineering College
Mumbai, India

Ravnit Sehgal
B.E Student,
Dept. Of Information Technology
Shah & Anchor Kutchhi Engineering College
Mumbai, India

Kartik Rao
B.E Student,
Dept. Of Information Technology
Shah & Anchor Kutchhi Engineering College
Mumbai, India

Saiyam Khambhaliya
B.E Student,
Dept. Of Information Technology
Shah & Anchor Kutchhi Engineering College
Mumbai, India

Dr. Bhavesh Patel
Ph.D (Computer Science & Engineering)
Principal
Shah & Anchor Kutchhi Engineering College
Mumbai, India

Jalpa Mehta
M. Tech (Computer Science),
Assistant Professor
Shah & Anchor Kutchhi Engineering College
Mumbai, India

Abstract - Nowadays, Artificial intelligence (AI) is becoming part of our lives. It impacts all aspects of our lives: our day to day activities, the way we interact with others, our jobs or occupation, and therefore the choices that we make. Undoubtedly, Artificial Intelligence is the next big thing in Computer Science. The research conducted by technology companies in the field of Artificial Intelligence controls various verticals including Health Care, Automobile, Financial Services, Manufacturing, and Retail. In this paper, we propose a way to help visually challenged people with the help of AI and Machine Learning. We aim to develop an application for visually challenged people which will integrate the module's of Person Recognition, Object Detection, Image Captioning, Text to Voice and Distance Calculation into one unit. It will improve the accuracy and reduce the time taken by each module to provide the results.

I. INTRODUCTION

The aim of this project is pretty simple for us, to make the world around us a better place, even for the differently abled, so that they experience all that life has to offer.

When it comes to H.O.D.E.R, we strive to give the visually impaired a whole new set of eyes, giving them a new perception of what life around them looks like. Thankfully, the times we live in allow us to make this possible and the scope, to be honest, is unsurprising and boundless. H.O.D.E.R is based on YOLO (You Only Look Once), the same technology that Google has been working on to bring about the revolutionary 'self-driving car'. Today, this technology has matured to the point that its demand in every sector is highly prioritized, as every industrialist envisions to keep up with whatever can ease human efforts and intervention.

II. BACKGROUND AND MOTIVATION

We, as the youth of today's progressive society, strive to make the world around us a better place for each and every one of us, impartially and believe that no one deserves any less just because they are challenged in certain aspects of life. We believe that everyone irrespective of their background must be allowed to exercise their right to equal opportunity.

H.O.D.E.R (Highly Optimized Detection using Electronic Retina) is a product to help blind people to detect over 9000 objects and recognize people they see in their everyday life on a daily basis with an option to teach the algorithm about more. The proposed system Highly Optimized Detection using Electronic Retinas (H.O.D.E.R) integrated with Artificial Intelligence will be used to construct a network application that will include training based on You Only Look Once Algorithm having 24 layers running at the same time, which will learn or train objects and persons with the one-shot training feature. It will not only calculate the distance between the object and user but also convert the text to speech which the user will be able to hear through his/her headset. We are also providing multilingual feature and variation of speech for the user's convenience and Artificial Intelligence automated investigation method, based on learning agent.

III. PROPOSED APPROACH

The objective of the proposed system Highly Optimized Detection using Electronic Retinas (H.O.D.E.R) integrated with Artificial Intelligence will be used to:

A) Object Detection and Person Identification

Detect objects with the objects with high accuracy and train the new objects which are not included.

B) Distance Calculation

Calculate the distance between the person and object and convert it from text to speech which the user will be able to hear through his/her headset.

C) Image Captioning and Text to Speech

Generate report on the basis of the investigation that will include artificial intelligence automated investigation method, based on the learning agent in the form of text and then converting it into audio format for the user.

All of these modules will then be integrated. As the proposed product is a mobile application, for the Android platform. We have observed that the mobile devices have limited computing power as well as storage and so generates a huge delay between the image capturing and speech output. As it will not be feasible to let the mobile phones do the computation and hold the training data, we have to resort to other technologies. We propose the application be deployed on the cloud. This will help speed up the process and have a low delay.

A. Object Detection and Person Identification

Object detection is the most basic step in computer vision tasks. As the product is supposed to act as the eyes of the user, its most important task would be to first detect the object fast but also with high accuracy. This can be done via Machine Learning techniques which need features and a Support Vector Machine for the classification or deep learning which in turn doesn't need the features, but use Convoluted Neural Networks for an end to end object detection. We have decided to use the later technique's approach using YOLO i.e You Only Look Once. [6]

YOLO9000 is a state-of-the-art, real-time object detection system that can detect over 9000 object categories[1]. YOLOv3 is extremely fast and accurate. In mAP measured at .5, IOU YOLOv3 is on par with Focal Loss but about 4x faster. We can easily change the tradeoff between speed and accuracy simply by changing the size of the model without retraining the entire model again. This helps improve ease of change according to our needs.

Previously detection used to be done using classifiers or localizers, but YOLOv3, the full image goes through a single neural network. This neural net splits the image into regions and estimates bounding boxes and probabilities for each area. These bounding boxes are weighted by the predicted probabilities. This makes it extremely fast, more than 1000x faster than R-CNN and 100x faster than Fast R-CNN.[4]

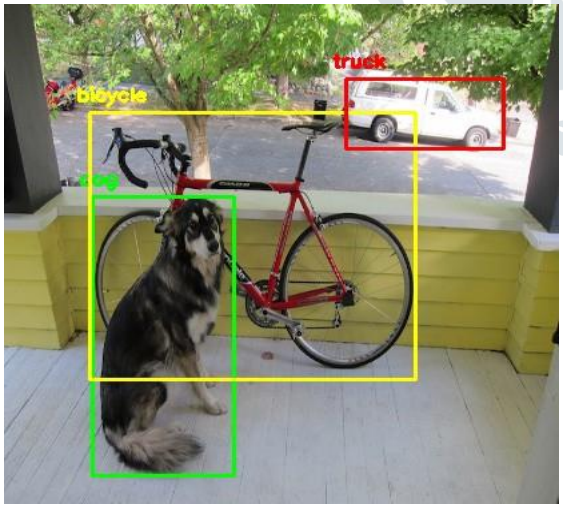

Figure 1: Object Detection using YOLO

Similarly, the same approach is used for Person Identification where the trained model can differentiate it from other objects. If we train the model, we can add the feature of person recognition to it but the drawback to this approach is that our all 9000 pre-trained objects are lost and we have to train it from the scratch.

B. Distance Calculation

The YOLO Model upon recognizing the objects/entity that is present in the video frame that is being projected to the model at real time, bounds the objects within rectangular boxes, that is, enclosing it to ensure that the object is detected with the most accurate conclusion.

Bounding boxes formed around the entity, or in other words, the contours around the bounded object have to separate from other objects in the frame if any, for which the model first converts the frame from the camera input into grayscale, for finding out the noise in the frame which helps to make sure that the object detected is distinct in nature and also, gives the model an approximation of where the object begins and ends. Post detection, the real-time frame is sent back to the model with the boundary around it.

Figure 2. Distance Calculation on implemented model

These Bounded Boxes with respect to the frame that is being projected to the model are traced with the help of pixel coordinates on the axes. Upon calculating the width of the box bounded around the object/entity which is done in pixels, it needs to be converted into inches which are done with the help of the triangle similarity[2]

$$F = (P \times D) / W$$

where W is the Width of the object that is detected, D is the distance of the object from the camera which is projected to the model at real time. Our calculated value is P which denotes the apparent width in pixels, calculated with the help of the pixel coordinates received from the frame. The Focal Length F of the object will vary for each object in the frame and has been set assuming the average size of the object in the frame.

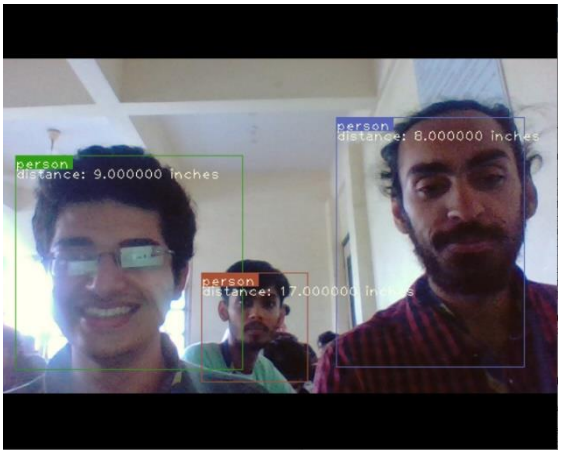

Figure 3: Person recognition and distance calculation using H.O.D.E.R.

The distance of the object in the frame from the camera is then computed in real time and returned to the model, which is converted to speech and spoken out by the tts module.

C. Image Captioning and Text to Speech

Suppose you see this picture –

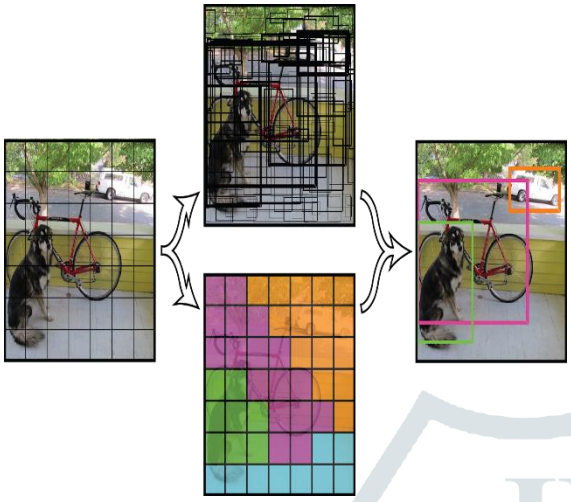

Figure 4: YOLO Convolutional Model Breakdown

- The brain, depending on its maturity and vocabulary would have a different definition for each image. However, the AI has to describe the image in the simplest of its meaning which would tell exactly what it sees in its simplest jargon. The system would have to be trained accordingly to frame sentences based on when and what it sees, making it easy for the person to know what's happening around them. The captioning of the image will have the output in text format.

- But this format is not useful for the user as the users are primarily those with sight impairment, and so it must be seen to it that it is in a format which they can easily understand. The output in a pad with braille which changes according to the output was thought about, but not everyone can read braille nor is it feasible for production. Instead, a simpler method was thought in which the text would be converted to speech. This speech can be then modulated as per the user's requirements.

Variation in volume, speed, and even the language can be varied as per the user's needs. We have used Pyttsx 3 2.5 to convert text to speech. It is cross-platform and can run on both Python 2 and 3.Pyttsx is completely offline and works seamlessly and has multiple texts to speech-engine support too. It contains sapi5, nsss and espeak TTS Engines.

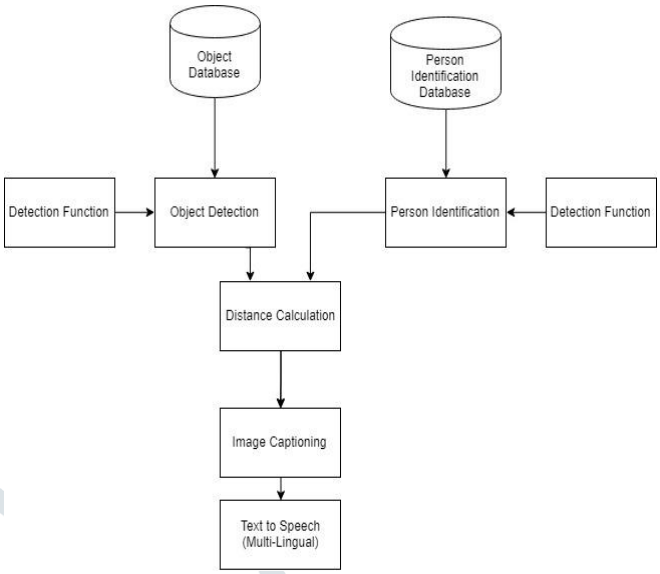

Figure 5. HODER System Flow with respect to YOLO Model

IV. Accuracy analysis with other existing object detection models

When it comes to object or person recognition, almost all existing object detection models, including R-CNN, ResNet, or for that matter, even YOLO follow the same methodology of processing the different objects that are present in the frame.The comparison, thus, between the models has to be superficially done on the basis of the balance between accuracy, and the speed of detecting the objects. The selection of the model is thus weighed upon the availability of better existing technology, which outweighs the drawbacks of existing models.

The mean average precision (mAP), which maps the average precision of the objects detected in the images frame gives us an idea about the balance between the accuracy and speed of object detection for the respective model. The PASCAL VOC, (Visual Object Classes) which is a unification of various datasets has been used as reference dataset for testing the accuracy of the various models in comparison. The following table compares the detected objects with their precisions.

Table 1: PASCAL VOC2012 test detection results. YOLOv2 performs on par with state-of-the-art detectors like Faster R-CNN with ResNet and SSD512 and is 2 − 10× faster. [7]

Method	data	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
Fast R-CNN	07++12	68.4	82.3	78.4	70.8	52.3	38.7	77.8	71.6	89.3	44.2	73	55	87.5	80.5	80.8	72	35.1	68.3	65.7	80.4	64.2
Faster R-CNN	07++12	70.4	84.9	79.8	74.3	53.9	49.8	77.5	75.9	88.5	45.6	77.1	55.3	86.9	81.7	80.9	79.6	40.1	72.6	60.9	81.2	61.5
YOLO	07++12	57.9	77	67.2	57.7	38.3	22.7	68.3	55.9	81.4	36.2	60.8	48.5	77.2	72.3	71.3	63.5	28.9	52.2	54.8	73.9	50.8
SSD300	07++12	72.4	85.6	80.1	70.5	57.6	46.2	79.4	76.1	89.2	53	77	60.8	87	83.1	82.3	79.4	45.9	75.9	69.5	81.9	67.5
SSD512	07++12	74.9	87.4	82.3	75.8	59	52.6	81.7	81.5	90	55.4	79	59.8	84.3	84.7	83.3	50.2	50.2	78	66.3	86.3	72
ResNet	07++12	73.8	86.5	81.6	77.2	58	51	78.6	76.6	93.2	48.6	80.4	59	92.1	85.3	84.8	80.7	48.1	77.3	66.5	84.7	65.6
YOLOv2 544	07++12	73.4	86.3	82	74.8	59.2	51.8	79.8	76.5	90.6	52.1	78.2	58.5	89.3	82.5	83.4	81.3	49.1	77.2	62.4	83.8	68.7

The YOLO Model trades off speed and accuracy with ease in comparison to all other models as it is quite flexible with in terms of processing the image from the frame after capturing it. The image resolution size can be manually adjusted if necessary to meet the hardware requirements.

Table 2: Results on COCO test-dev 2015. Table adapted from [8][9][10][11]

Detection Frameworks	Train	mAP	FPS
Fast R-CNN	2007+2012	70	0.5
Faster R-CNN VGG-16	2007+2012	73.2	7
Faster R-CNN ResNet	2007+2012	76.4	5
YOLO	2007+2012	63.4	45
SSD300	2007+2012	74.3	46
SSD500	2007+2012	76.8	19
YOLOv2 288 × 288	2007+2012	69	91
YOLOv2 352 × 352	2007+2012	73.7	81
YOLOv2 416 × 416	2007+2012	76.8	67
YOLOv2 480 × 480	2007+2012	77.8	59
YOLOv2 544 × 544	2007+2012	78.6	40

V. Future Scope

Like every project, even our project is not perfect and has a lot of scope for improvement. We plan to improve our project by reducing the dependency on the mobile phone. We plan to integrate our project with custom hardware which resembles a walking stick. The stick can have a camera and can be connected to the Android phone. It will be easier for the user to use it than carry the android phone .Another alternative is to integrate it with a device which resembles smart spectacles so that the hands will not be needed.

This will be challenging, but overcoming challenges is the key to technological advancement. We hope our application inspires others to help people and encourage others to help us improve.

VI. Conclusion

The application developed aims to achieve an efficient way to detect, identify objects and person. This system provides information via voice which can be heard loud or with the help of earphones. This encouraged many people to use this application as the models which are used to make this application has the highest percentage of accuracy. Therefore, a combination of various models used in making this application would be perfect for blind people to do their daily activities.

References

[1] Redmon, Joseph, and Farhadi, Ali, YOLO9000: Better, Faster, Stronger

[2]Adrian Rosebrock, Image Processing, https://www.pyimagesearch.com/2015/01/19/find-distance-camera-objectmarker-using-python-opencv/

[3] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. E. Reed. SSD: single shot multibox detector. CoRR, abs/1512.02325, 2015.

[4] J. Redmon. Darknet: Open source neural networks in c. http://pjreddie.com/darknet/

[5] Natesh M Bhat, Text to Speech https://pypi.org/project/pyttsx3/

[6] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. International journal of computer vision, 88(2):303– 338, 2010

[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. E. Reed. SSD: single shot multibox detector. CoRR, abs/1512.02325, 2015.

[8] R. B. Girshick. Fast R-CNN. CoRR, abs/1504.08083, 2015

[9] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. arXiv preprint arXiv:1506.01497, 2015

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. arXiv preprint arXiv:1506.02640, 2015.

[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. E. Reed. SSD: single shot multibox detector. CoRR, abs/1512.02325, 2015