ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 7, Issue 1, January 2020

Virtual Assistant for the Visually Impaired using Deep Learning

[1] Preethi Harris, [2] S.Aparnadevi and [3] N. Mahaswetha

[1][2][3] Department of Information Technology, Sri Ramkrishna Engineering College, India
[1] preethi.harris@srec.ac.in, [2] aparnadevi.1505013@srec.ac.in, [3]mahaswetha.1505050@srec.ac.in

Abstract: **- Vision impairment is one of the top disabilities among mankind. Indians with this disability roughly account to one-third of the world's blind population. Hence accessible visual information improves independence and safety of visually impaired people. With the explosion of data and multimedia, researchers are exploring new avenues to train devices that detect and classify multiple objects within an image. At the outset, the advancements in the field of deep learning can be extended to enhance the life of the visually challenged through smart devices also. Inspired by these findings in literature, a virtual assistant for the visually challenged has been proposed. The assistant detects and classifies objects with an image to provide a voice output for the detected objects. This system has been designed using Mobilenet and Single Shot Detector (SSD) algorithm pertaining to Deep Learning, to incorporate deep learning network for PC and mobile devices and also finds its application for illiterates.**

Keywords: **- Visually challenged, object, Virtual Assistant, Deep Learning, Audio Video, Mobilenet, SSD.**

I. INTRODUCTION

With the rise of autonomous vehicles, smart video surveillance, facial detection and various people counting applications, fast and accurate object detection systems are rising in demand [1,2]. These systems involve not only recognizing and classifying every object in an image, but localizing each one by drawing the appropriate bounding box around it [3,4].

An image classification or image recognition model simply detects the probability of an object in an image. At the outset object localization refers to identifying the location of an object in the image. An object localization algorithm outputs the coordinates of the location of an object with respect to the image. This makes object detection a significantly harder than its traditional computer vision predecessor which is image classification. With the advent of 4th industrial revolution Artificial Intelligence (AI) has become the buzz word [5]. Deep learning which is an essential part of AI has fueled tremendous progress in the field of Computer Vision in recent years, with Neural Networks repeatedly pushing the frontier of Visual Recognition technology[6]. While many of those technologies such as object, landmark, logo and text recognition are provided for Internet-connected devices through the Cloud Vision API[7], the ever-increasing computational power of mobile devices can enable the delivery of these technologies into the hands of users, anytime, anywhere, regardless of Internet connection. However, visual recognition for a device and embedded applications poses many challenges such as models must run quickly with high accuracy in a resource-constrained environment making use of limited computation, power and space. There has been a significant research in the field of deep learning and multi object detection and some of the notable contribution in literature is summarized as follows:

The work of Esraa Elhaririaet al [8], an assistive object recognition system for enhancing seniors quality of life, presents an indoor object recognition system based on the histogram of oriented gradient and Machine Learning (ML) algorithms such as Support Vector Machines (SVMs), Random Forests (RF) and Linear Discriminate Analysis (LDA) algorithms. This system exploits HOG descriptor to identify different indoor objects in a multi-class scenario. The results showed that RF classifier is better than SVM and LDA and achieved good accuracy 80.12 %.

A research on Multi object detection in video surveillance application by Neha Sharma et al[9] has been carried out using 3 techniques namely Gaussian Mixture Model (GMM) technique, Approximate Median filter(AMF),Optical flow method resulted in calculation of false positive, false negative, true positive and true negative by comparing detected output image with the corresponding ground truth image.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 7, Issue 1, January 2020

A study on application of Deep Learning for object detection by Ajeet Ram Pathaka et al [10] compares the features of object detection frameworks, namely, Caffe, Microsoft Cognitive Toolkit ,IBM Vision recognition Service, Google Cloud Vision. The results show infeasibility to process large surveillance data and a need to bring data closer to the sensor where data are generated. This would result into real time detection of objects. The current object detection systems are small in size having 1-20 nodes of clusters having GPUs. These systems were extended to cope with real time full motion video generating frames at 30 to 60 per second. Such object detection analytics could be integrated with other tools using data fusion.

The application of Mobilenets, efficient Convolutional Neural Networks for Mobile Vision Applications by Andrew G. Howard et al [11]proposed a new model architecture called MobileNets based on depthwise separable convolutions. In the work Mobilenet functionality was compared to popular models such as AlexNet and Resnet demonstrating its superior size, speed and accuracy characteristics.

Thus, literature survey done revealed the various techniques and frameworks used for object detection. However a system that can accurately, efficiently and quickly detect and classify objects is in need.

With accessibility to visual information which is of paramount importance to improve independence and safety of blind and visually impaired people, there is a pressing need to develop smart automated systems to assist their navigation, specifically in unfamiliar environments. With the availability of large amounts of data, faster GPUs and better algorithms, it is now easy to train computers to detect and classify multiple objects within an image with high accuracy which will help the visually challenged. The next section describes the proposed work done to assist the visually challenged.

II. PROJECT DESCRIPTION

The current Deep Learning methods [12] exhibit low transparency and interpretability. The models require incredible amounts of data to train. Subsequently the models created are very specialized and narrow and furthermore lack the ability to perform transfer learning between tasks.These functionalities have been extended with the proposed work. In the proposed system,Mobilenet which is based on streamlined architecture that uses Network (LRCN), depthwise separable convolutions to build light weight deep neural networks has been used along with Single Shot Detector(SSD) algorithm to detect objects in real time.The object's classes are detected using the algorithm and then converted as a voice output to assist the visually impaired.This is done using Google Text to Speech (gTTS) and the entire process is depicted in Fig 1.

The basic blocks in the proposed system are:

Video Capture: Real time video is captured through mobile camera.IP Webcam is an android app that is used to stream video captured from the mobile phone to laptop.

Preprocessing: Preprocessing involves resizing the frame to 300 by 300 pixel to match the scale of the feature map converted to gray scale to serve as input to the Mobilenet.

Caffe Mobilenet SSD: Caffeis a deep learning framework characterized by its speed, scalability, and modularity. Caffe works with CPUs and GPUs and is scalable across multiple processors. This framework is suitable for various architectures such as CNN (Convolutional Neural Network), Long-Term Recurrent Convolutional Long Short-Term Memory (LSTM) or fully connected neural networks.MobileNets are a family of mobile-first computer vision models, designed to effectively maximize accuracy while being mindful of the restricted resources for an on-device or embedded application. MobileNets are small, low-latency, low-power models parameterized to meet the PC resource constraints.

Single Shot Detector: achieves a good balance between speed and accuracy [13]. The architecture of SSD runs a convolutional network on input image only once and calculates a feature map. In this work, we run a small 3×3 sized convolutional kernel on this feature map to predict the bounding boxes and classification probability for the

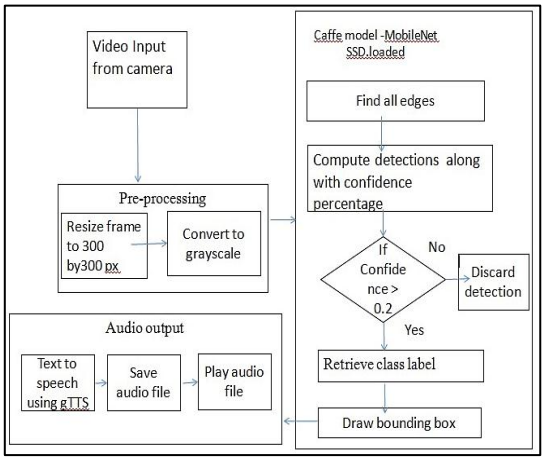

Fig.1. Basic activities depicting proposed system

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 7, Issue 1, January 2020

bounding box to represent object localization. SSD also uses anchor boxes at various aspect ratio. Then to handle the scale, SSD is used to predict bounding boxes generated after multiple convolutional layers. As each convolutional layer operates at a different scale, it is able to detect objects of various scales given via video capture.

The steps of Caffe model-Mobilenet SSD is summarized as

1. The SSD algorithm scans the captured frame for edges
2. The algorithm detects the objects based on the edges and calculates a confidence percentage for each detected object
3. If the confidence of the detected object exceeds the minimum confidence threshold, then the class label of the object is retrieved.
4. As multiple objects can be detected in a single frame, a bounding box is drawn around each detected object.

Audio Output: Once the object is detected, the class label is fed into the audio module and the steps to convert it into speech is as follows:

1. Google Text To Speech (gTTS) API is used to convert the retrieved class label to a voice output
2. The voice output is saved as an mp3 audio file
3. The audio filed is loaded and played every time an object is detected

Algorithm: The entire process of detecting the object from the training dataset, converting it to the relevant class label using MobileNet and SSD and finally into the appropriate voice output is represented using the algorithm: ObjectLabel_to_Audio in Fig 2.

First the real time video is captured through a mobile phone camera. The captured video is processed frame by frame and the frame is subjected to preprocessing. The preprocessed frame is passed to the network. The network detects the objects by identifying the edges. The detected objects along with the confidence percentage are stored in a Numpy array. If the confidence of the detected object is greater than 0.2,the detection is considered valid. To localize the position of the detected objecta boundary box is drawn around it. The scale of the boundary box is specified in equation 1 and equation 2.

Then the class label is converted to a voice output using Google Text To Speech API and the corresponding audio file is loaded and played each time an object is detected.

Algorithm 1: ObjectLabel_to_Audio

Input:
Mobilenet SSD caffe model.prototext
Video frame
Output:
Object detection and localization in video frame
Audio output of object class label.
Begin
{
D: is a Numpy array, denoting detections
h : height of video frame
w:width of video frame
b:video frame resized to 300 X 300 pixel
algorithm ObjectLabel_to_Audio
{
begin
b : resize input video frame
d : feed to network
for each d**do**
c: retrieve confidence of each detection
if c >0.2
i : class label of detected object
b: draw bounding box

$s_k = s_{\min} + \frac{s_{\max} - s_{\min}}{m-1}(k-1), \quad k \in [1,m]$ //default

$a_r \in \{1,2,3,\frac{1}{2},\frac{1}{3}\}$ $(w_k^a = s_k\sqrt{a_r})$ $(h_k^a = s_k/\sqrt{a_r})$ //non square

$s'_k = \sqrt{s_k s_{k+1}}$ //square

Write class label above bounding box

for eachi do
a: convert i to audio using Google Text To Speech
if a not in system path do
save a
load a
play a
endfor
end for
end
}
endObjectLabel_to_audio

Fig 2. Object Label to audio output

Default bounding box equation 1:

$$s_k = s_{\min} + \frac{s_{\max} - s_{\min}}{m-1}(k-1), \quad k \in [1,m] \qquad (1)$$

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 7, Issue 1, January 2020

Non square bounding box equation 2 :

$$a_r \in \{1, 2, 3, \tfrac{1}{2}, \tfrac{1}{3}\} \quad (w_k^a = s_k\sqrt{a_r}) \quad (h_k^a = s_k/\sqrt{a_r}) \qquad (2)$$

Square bounding box equation 3:

$$s_k' = \sqrt{s_k s_{k+1}} \qquad (3)$$

III. RESULTS AND DISCUSSION

The System is designed using Python and OpenCv library for image processing. The system runs on Windows OS with 4GB RAM,100 GB hard disk with a clock speed of 2 GHZ. The network was pre- trained using COCO dataset and fine-tuned using PASCAL VOC [13].The input is a realtime video captured using a 15 MP mobile phone camera. A sample output is shown in Fig 3, where a bottle is captured as a class label and converted into the corresponding audio output.

The Table I shows the number of objects in the frame and the number of objects detected by the algorithm for the PASCAL VOC dataset.

Fig 3.Sample output

TABLE I: Accuracy of detected objects

Test Data	No. of Objects in Frame	No. of Objects Detects	Accuracy (%)
1	7	4	57
2	3	3	100
3	6	5	83
4	4	4	100
5	10	7	70

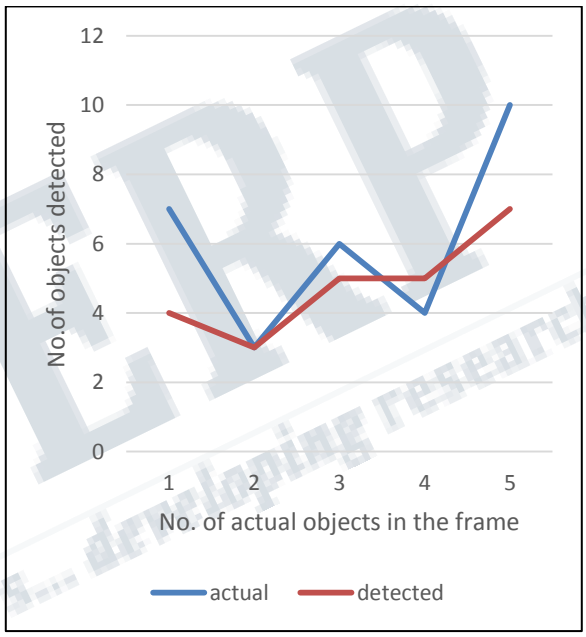

Fig..4. Visualization of performance of SSD

The Fig.4 visualizes the performance of SSD, comparing the actual number of objects in the frame (x-axis) and the number of objects detected (y axis) and localize and localized for audio output to the visually impaired.

From the graph depicted, the average accuracy for the inputted PASCAL VOC dataset is approximately 82%.Thus the object labels notified via audio to the visually challenged is almost close to the actual objects captured by the video frame and labeled using Mobilenet SSD.

The key feature of the model is the use of multi-scale convolution bounding box outputs attached to multiple feature maps at the top of the network. This representation allows us to efficiently model the space of possible box shapes. This model also allows accurate and fast detections

on a real time video input compared to other models. Unlike MultiBox, every feature map cell is associated with a set of default bounding boxes of different dimensions and aspect ratios. This allows any type of input, without requiring a pre-training phase for prior generation.

IV. CONCLUSION

Addressing the day to day activities of physically challenged is the need of the hour. Motivated by these findings, the proposed work focuses on developing a virtual assistant to help the visually challenged in their day to day activities. This work encompasses the concepts of deep learning along with neural net-Mobilenet. The single shot detector algorithm is used to classify objects and their class label is fed as input to Google Text To Speech API to convert the class label to a voice output. The system can be enhanced to assist in the navigation of the visually impaired.

REFERENCES

[1] India has largest blind population https://timesofindia.indiatimes.com/india/India-has-largest-blind population/articleshow/2447603.cms
[2] Rodrigo Verschae1 and Javier Ruiz-del-Solar, "Object Detection: Current and Future Directions", Frontiers in Robotics and AI, pp.1-7,Vol 2, 2015.
[3] Object detection: one-stage methods.https://www.jeremyjordan.me/object-detection-one-stage/
[4] Object Localization and Detectionhttps://leonardoaraujosantos.gitbooks.io/artificialinteligence/content/object_localization_and_detection.html
[5] .Skilton Mark and Hovsepian Felix, "The 4th Industrial Revolution Responding to the Impact of Artificial Intelligence on Business", First Edition, Palgrave Macmillan, 2018.
[6] Frontier of visual recognition technologyhttps://ai.googleblog.com/2016/08/improving-inception-and-image.html
[7] CloudvisionAPIhttps://cloud.google.com/vision/
[8] Esraa Elhariria, Nashwa El-Bendaryb, Aboul Ella Hassanienc and Vaclav Snaself ,"An Assistive Object Recognition System for enhancing seniors quality of life", in the proceedings of the International Conference on Communication, Management and Information Technology , pp. 691-700,Vol.65,2018.
[9] Rohini Chavan and Sachin R. Gengaje ,"Multi object detection techniques in video surveillance application",in the proceddings of IEEE International Conference on Power, Control, Signals and Instrumentation Engineering, pp. 11-23,2017.
[10] Ajeet Ram Pathaka, Manjusha Pandeya and Siddharth Rautaraya , "Application of Deep Learning For Object Detection", in the proceedings of International Conference on Computational Intelligence and Data Science, pp. 1706-1717,Vol.132,2018.
[11].Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto and Hartwig Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications", arXiv:1704.04861v1, 2017.
[12] Wei Liu, Dragomir Anguelov, DumitruErhan, Christian Szegedy,Scott Reed, Cheng-Yang Fu, and Alexander C. Berg, "SSD: Single Shot MultiBox Detector", in the proceedings of the European Conference on Computer Vision, Pg.21-37,2016.
[13] Mark Everingham ,S. M. Ali Eslami,Luc Van Gool,Christopher K. I.Williams,John Winn and Andrew Zisserman,"ThePASCAL VisualObject Classes Challenge:A Retrospective",International Journal of Computer Vision, pp.98-136,Vol.111, No.1,2015