VISVESVARAYA TECHNOLOGICAL UNIVERSITY

Jnana Sangama, Belgaum-590018

A PROJECT REPORT **(15CSP85)** ON

"OBJECT DETECTION WITH TEXT & VOCAL REPRESENTATION"

Submitted in Partial fulfillment of the Requirements for the Degree of

Bachelor of Engineering in Computer Science & Engineering

By

PUNYASLOK SARKAR(1CR16CS121)

BIJON BORA(1CR16CS404)

SHASHI SAGAR (1CR15CS197)

VIKASH NARA(1CR16CS433)

Under the Guidance of,

MRS. ANJALI GUPTA

Asst. Professor, Dept. of CSE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CMR INSTITUTE OF TECHNOLOGY

#132, AECS LAYOUT, IT PARK ROAD, KUNDALAHALLI, BANGALORE-560037

CMR INSTITUTE OF TECHNOLOGY

#132, AECS LAYOUT, IT PARK ROAD, KUNDALAHALLI,BANGALORE-560037

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

Certified that the project work entitled **"OBJECT DETECTION WITH TEXT & VOCAL REPRESENTATION"** carried out by **Mr. Punyaslok Sarkar**, USN **1CR16CS121**, **Mr.Bijon Bora**, USN **1CR16CS404**,**Mr**. **Shashi Sagar**, USN **1CR15CS197, Mr.Vikash Nara**, USN **1CR16CS433,** bonafide students of CMR Institute of Technology, in partial fulfillment for the award of **Bachelor of Engineering** in Computer Science and Engineering of the Visveswaraiah Technological University, Belgaum during the year 2019-2020. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the Report deposited in the departmental library.

The project report has been approved as it satisfies the academic requirements in respect of Project work prescribed for the said Degree.

Mrs. Anjali Gupta	**Dr. Prem Kumar Ramesh**	**Dr. Sanjay Jain**
Asst. Professor	**Professor & Head**	**Principal**
Dept. of CSE, CMRIT	**Dept. of CSE, CMRIT**	**CMRIT**

External Viva

Name of the Examiners	Signature with Date
1.	
2.	

DECLARATION

We, the students of Computer Science and Engineering, CMR Institute of Technology, Bangalore declare that the work entitled "**OBJECT DETECTION WITH TEXT & VOCAL REPRESENTATION**" has been successfully completed under the guidance of Prof. Anjali Gupta, Computer Science and Engineering Department, CMR Institute of technology, Bangalore. This dissertation work is submitted in partial fulfillment of the requirements for the award of Degree of Bachelor of Engineering in Computer Science and Engineering during the academic year 2019 - 2020. Further the matter embodied in the project report has not been submitted previously by anybody for the award of any degree or diploma to any university.

Place:

Date:

Team members:

PUNYASLOK SARKAR (1CR16CS121) _____

BIJON BORA (1CR16CS404) _____

SHASHI SAGAR(1CR15CS197) _____

VIKASH NARA (1CR16CS433) _____

ABSTRACT

The YOLO design enables end-to-end training and real-time speeds while maintaining high average precision**.** In present industry, communication is the key element to progress. Passing on information, to the right person, and in the right manner is very important, not just on a corporate level, but also on a personal level. The world is moving towards digitization, so are the means of communication. Phone calls, emails, text messages etc. have become an integral part of message conveyance in this tech-savvy world. In order to serve the purpose of effective communication between two parties without hindrances, many applications have come to picture, which acts as a mediator and help in effectively carrying messages in form of text, or speech signals over miles of networks. Most of these applications find the use of functions such as articulatory and acoustic-based speech recognition, conversion from speech signals to text, and from text to synthetic speech signals, language translation among various others. In this review paper, we'll be observing different techniques and algorithms that are applied to achieve the mentioned functionalities. Communication is the main channel between people to communicate with each other.

ACKNOWLEDGEMENT

I take this opportunity to express my sincere gratitude and respect to **CMR Institute of Technology, Bengaluru** for providing me a platform to pursue my studies and carry out my final year project

I have a great pleasure in expressing my deep sense of gratitude to **Dr. Sanjay Jain,** Principal, CMRIT, Bangalore, for his constant encouragement.

I would like to thank **Dr. Prem Kumar Ramesh,**Professor and Head, Department of Computer Science and Engineering, CMRIT, Bangalore, who has been a constant support and encouragement throughout the course of this project.

I consider it a privilege and honor to express my sincere gratitude to my guide **Mrs. Anjali Gupta, Asst. Professor,** Department of Computer Science and Engineering, for the valuable guidance throughout the tenure of this review.

I also extend my thanks to all the faculty of Computer Science and Engineering who directly or indirectly encouraged me.

Finally, I would like to thank my parents and friends for all their moral support they have given me during the completion of this work.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

FCFS	**First Come First Serve**
RR	**Round Robin**
YOLO	**You Only Look Once**
TTS	**Text-to-Speech**
UML	**Unified Modeling Language**
DFD	**Data Flow Diagram**

CHAPTER 1

INTRODUCTION

Over the past few years, Cell Phones have become an indispensable source of communication for the modern society. We can make calls and text messages from a source to a destination easily. It is known that verbal communication is the most appropriate modem of passing on and conceiving the correct information, avoiding misquotations. To fulfil the gap over a long distance, verbal communication can take place easily on phone calls. A path-breaking innovation has recently come to play in the SMS technology using the speech recognition technology, where voice messages are being converted to text messages. Quite a few applications used to assist the disabled make use of TTS, and translation. They can also be used for other applications, taking an example: Siri an intelligent automated assistant implemented on an electronic device, to facilitate user interaction with a device, and to help the user more effectively engage with local and/or remote services makes use of Nuance Communications voice recognition and text-to-speech (TTS) technology. In this paper, we will take a look at the different types of speech, speech recognition, speech to text conversion, text to speech conversion and speech translation. Under speech the recognition we will follow the method i.e. pre-emphasis of signals, feature extraction and recognition of the signals which help us in training and testing mechanism. There are various models used for this purpose but Dynamic time warp, which is used for feature extraction and distance measurement between features of signals and Hidden Markov Model which is a stochastic model and issued to connect various states of transition with each other is majorly used. Similarly, for conversion of speech to text we use DTW and HMM models, along with various Neural Network models since they work well with phoneme classification, isolated word recognition, and speaker adaptation. End to end ASR is also being tested as of late 2014 to achieve similar results. Speech synthesis works well in helping convert tokenized words to artificial human speech.

Relevance of the Project

It is widely used in computer vision tasks such as face detection, face recognition, video object co-segmentation. It is also used in tracking objects, for example tracking a ball during a football match, tracking movement of a cricket bat, or Tracking a person in a video.

Every object class has its own special features that helps in classifying the class – for example, all circles are round. Object class detection uses these special features. For example, when looking for circles, objects that are at a particular distance from a point (i.e. the center) are sought. Similarly, when looking for squares, objects that are perpendicular at corners and have equal side lengths are needed. A similar approach is used for face identification where eyes, nose, and lips can be found and features like skin color and distance between eyes can be found.

Scope of the Project

In this project we are using general purpose and, a unified model for object detection object (YOLO). The object is detected using the Yolo algorithm and objects name is converted from text to speech.

Chapter Wise Summary

Normally, a blind person uses cane as a guide of him to protect him from obstacles. Most of area of surrounding is covered by the cane, especially the area near to his legs like stairs etc. But certain areas such as near to his head, especially when he is entering or leaving the door which is short in height. This system is specially designed to protect the area near to his head. The product is designed to provide full navigation to user into the environment. It guides the user about obstacles as well as also provides information about appropriate or obstacle free path. We are using buzzer and vibrator, two output modes to user. Logical structure: The logical structure of our system is can be divided into three main parts: the user control, sensor control, and the output to the user. Logical Structure the user control includes the switches that allow the user to choose project's mode of operation. There

are basically two modes of operation, Buzzer mode and Vibration mode. These modes are provided to user for taking output on his portability. Sometimes, he is not comfortable in getting the output in one mode. Vibration mode always not comfortable, can irritate him. Similarly, when there is a lot of noise in environment the buzzer mode is not portable. Another switch is controlled by the user, called initializing switch. The initializing switch is pressed when the user wants to stop the system. Sensor control determines when to tell the sensor to take a measurement and receives the output from the sensor and normalizes it to control value for the sensors. Basically, we are designing a sensor module. We are using proximity IR sensor for detection and it is mounted on a stepper motor. Stepper motor rotates continuously with an angle of 90 degree. The 90-degree angle is divided into three 30-degree portions. Two 30-degree areas are for indicating left direction or right direction obstacles, and third 30-degree area is for indication front obstacles. The main thing is our system is based on protecting the near head area because walking cane does not protect this area. Output to the user includes the indication of obstacles to user. Basically, we are using two output modes, vibration mode and buzzer mode. User can select any of the two modes in accordance to his convenience. Sometimes vibration mode is portable for him, especially when there is a lot of noise into the environment. Buzzer mode is generally used when the environmental noise is low and sometimes vibration can create irritation to the user. Architecture: The system architecture diagram of our project is given in following Fig 1. There are certain functions accomplished by these blocks. The description of blocks is as following Fig 2-Block Diagram. As per our propose application blind person taking video of the path where he was walking the application will give voice message to that blind person and it will help to that person for identifying he’s path. The object gets detected by the key matching technique which is used in the algorithm. And match that object with the database images to confirm the obstacle that comes into the way. When object is matched with database objects the application gives the voice instruction by using the Speech synthesizer. So, Blind user gets the direction from the application.

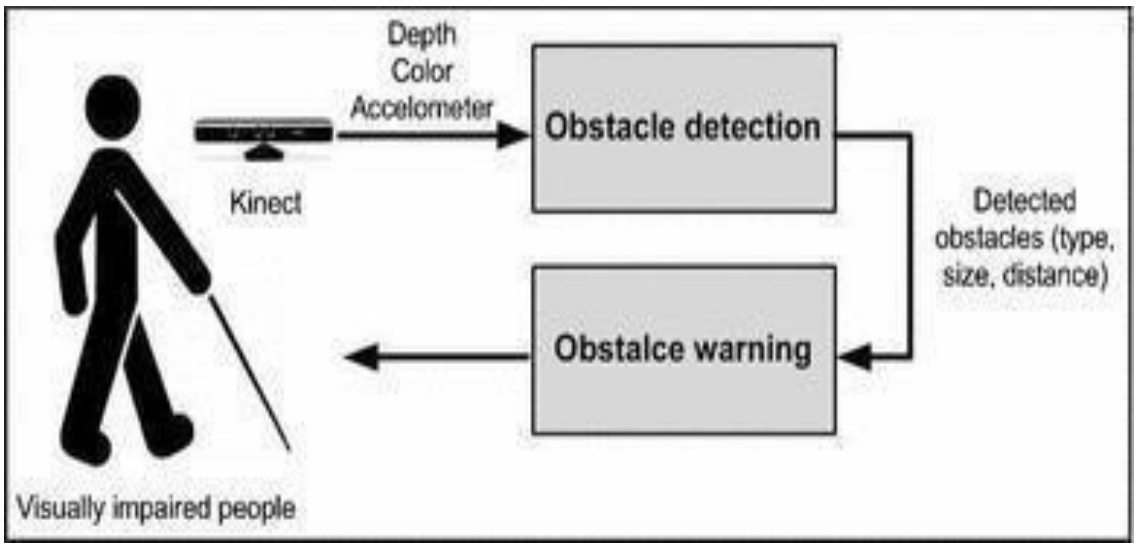


Fig. 1-Object Detection Process

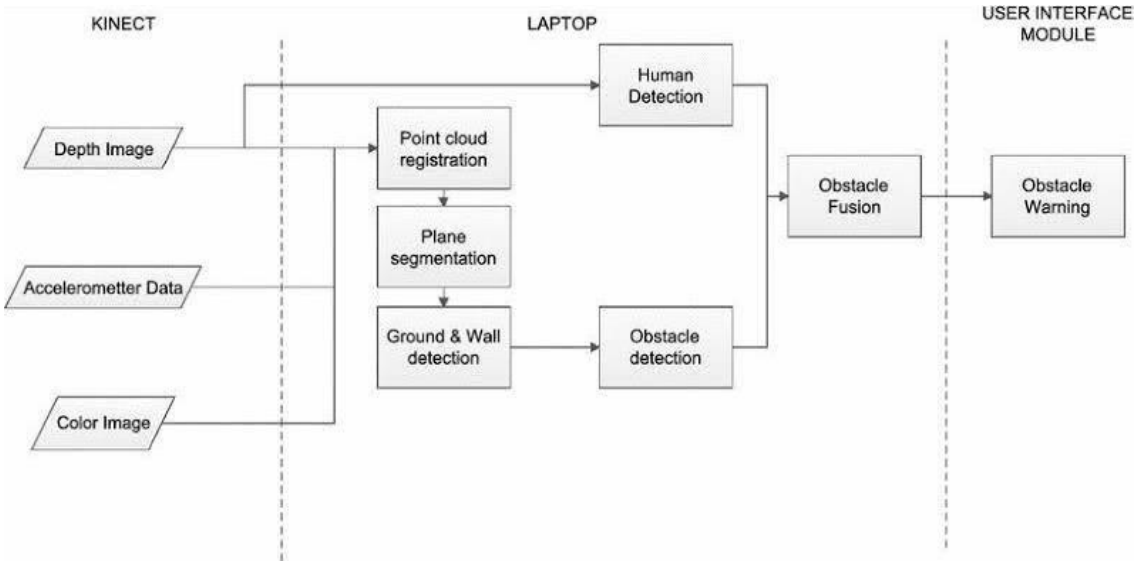


Fig. 2-Block diagram of Object Detection

Merits and Demerits

Reliable: This type of technology Provides good video quality. Difference between various objects like chair and table etc. can be easily differentiated and exact path can will be detected for visually impaired people. Scalable: This application can be run on various operating system. Object will not be stationary so it will capture the ongoing video and process all the developing steps for detection and placement of object. This feature highlights the merit. Efficient cost: The cost will be depending on the smart phones. Open Source: Android application is an open source utility command which is Linux based and released under apache software. It has many versions with extending features and properties. (e.g. lollipop, jellybean, kit Kat etc.) This application is mostly useful for blind person. No need to carry walking stick.

CHAPTER 2

LITERATURE SURVEY

1. The Cross-Depiction Problem: Computer Vision Algorithms for Recognizing Objects in Artwork and in Photographs

The cross-depiction problem is that of recognizing visual objects regardless of whether they are photographed, painted, drawn, etc. It is a potentially significant yet under-researched problem. Emulating the remarkable human ability to recognize objects in an astonishingly wide variety of depictive forms is likely to advance both the foundations and the applications of Computer Vision.

In this paper we benchmark classification, domain adaptation, and deep learning methods; demonstrating that none perform consistently well in the cross-depiction problem. Given the current interest in deep learning, the fact such methods exhibit the same behavior as all but one other method: they show a significant fall in performance over in homogeneous databases compared to their peak performance, which is always over data comprising photographs only. Rather, we find the methods that have strong models of spatial relations between parts tend to be more robust and therefore conclude that such information is important in modelling object classes regardless of appearance details.

2. Histograms of Oriented Gradients for Human Detection

We study the question of feature sets for robust visual object recognition, adopting linear SVM based human detection as a test case. After reviewing existing edge and gradient based descriptors, we show experimentally that grids of Histograms of Oriented Gradient (HOG) descriptors significantly outperform existing feature sets for human detection. We study the influence of each stage of the computation on performance, concluding that fine-scale gradients, fine orientation binning, relatively coarse spatial binning, and high-quality local contrast normalization in overlapping descriptor blocks are all important for good

results. The new approach gives near-perfect separation on the original MIT pedestrian database, so we introduce a more challenging data set containing over 1800 annotated human images with a large range of pose variations and backgrounds.

3. Speech YOLO: Detection and Localization of Speech Objects

In this paper, we propose to apply object detection methods from the vision domain on the speech recognition domain, by treating audio fragments as objects. More specifically, we present Speech YOLO, which is inspired by the YOLO algorithm for object detection in images. The goal of Speech YOLO is to localize boundaries of utterances within the input signal, and to correctly classify them. Our system is composed of a convolutional neural network, with a simple least-mean-squares loss function. We evaluated the system on several keyword spotting tasks that include corpora of read speech and spontaneous speech. Our system compares favorably with other algorithms trained for both localization and classification.

4. Detection and Content Retrieval of Object in an Image using YOLO

It is easy for human beings to identify the object that is in an image. Even if the task is complex, human beings require only a minimal effort. Since computer vision is actually replicating human visual system, the same thing can be achieved in computers when they are trained with large amount of data, faster GPUs and many advanced algorithms. In general terms, Object detection can be defined as a technology that detects instances of object in images and videos by mimicking the human visual system functionalities. The motivation of the paper is making the search process easier for the user i.e., if the object is very new for the user and he has no idea about it, he can upload a picture of that object and the algorithm will detect the object and gives a description about it. The objective of the paper is to detect the object in an image, once the object is detected, the label i.e., the name of the detected object is searched in Wikipedia and few lines of description about that object is retrieved and printed. Also, the label is searched in google and the URL of the top pages with content related to the label are

also displayed. The detection of object in an image is done using YOLO (You Only Look Once) algorithm with pre-trained weights. Previous methods for object detection, like R-CNN and its variations, used a pipeline to perform this task in multiple steps. This can take some time for execution complex optimization may be involved because individual training of components is required. YOLO, does it all fastly with a single neural network. Hence, YOLO is preferred.

5. Real-Time Two-Way Communication Approach for Hearing Impaired and Dumb Person Based on Image Processing

In the recent years, there has been rapid increase in the number of deaf and dumb victims due to birth defects, accidents and oral diseases. Since deaf and dumb people cannot communicate with normal person so they have to depend on some sort of visual communication. Gesture shows an expressive movement of body parts such as physical movements of head, face, arms, hand or body which convey some message. Gesture recognition is the mathematical interpretation of a human motion by a computing device. Sign language provide best communication platform for the hearing impaired and dumb person to communicate with normal person. The objective of this research is to develop a real time system for hand gesture recognition which recognize hand gestures, features of hands such as peak calculation and angle calculation and then convert gesture images into voice and vice versa. To implement this system we use a simple night vision web-cam with 20 megapixel intensity. The ideas consisted of designing and implement a system using artificial intelligence, image processing and data mining concepts to take input as hand gestures and generate recognizable outputs in the form of text and voice with 91% accuracy.

6. Object detection Real-Time Systems

Many research efforts in object detection focus on making standard detection pipelines fast. However, only Sadeghi et al. actually produce a detection system that runs in real-time (30 frames per second or better). We compare YOLO to their GPU implementation of DPM which runs either at 30Hz or 100Hz.While the other efforts don't reach the real-time milestone, we also compare their relative mAP anradeoffs

available in object detection systems. Fast YOLO is the fastest object detection method on PASCAL; as far as we know, it is the fastest extant object detector. With 52.7% map, it is more than twice as accurate as prior work on real-time detection. YOLO pushes map to 63.4% while still maintaining real-time performance. We also train YOLO using VGG-16. This model is more accurate but also significantly slower than YOLO. It is useful for comparison to other detection systems that rely on VGG-16 but since it is slower than real-time the rest of the paper focuses on our faster models. Fastest DPM effectively speeds up DPM without sacrificing much mAP but it still misses real-time performance by a factor of 2. It also is limited by DPM's relatively low accuracy on detection compared to neural network approaches.

7. Fast Detector image detector

Fast and Faster R-CNN focus on speeding up the R-CNN framework by sharing computation and using neural networks to propose regions instead of Selective Search. While they offer speed and accuracy improvements over R-CNN, both still fall short of real-time performance. Many research efforts focus on speeding up the DPM pipeline. They speed up HOG computation, use cascades, and push computation to GPUs. However, only 30Hz DPM actually runs in real-time. Instead of trying to optimize individual components of a large detection pipeline, YOLO throws out the pipeline entirely and is fast by design. Detectors for single classes like faces or people can be highly optimized since they have to deal with much less variation. YOLO is a general-purpose detector that learns to detect a variety of objects simultaneously.

CHAPTER 3

REQUIREMENTS SPECIFICATION

H/W System Configuration: -

- RAM - 8GB (min)
- Hard Disk - 2 GB
- Key Board - Standard Windows Keyboard
- Mouse - Two or Three Button Mouse
- Monitor

Software requirements

- Python 2.7 or higher
- PyCharm
- OpenCV
- Windows-8,10

CHAPTER 4
SYSTEM ANALYSIS & DESIGN

System design is the process of defining the architecture, components, modules, interfaces and data for a system to satisfy specified requirements. One could see it as the application of systems theory to product development. There is some overlap with the disciplines of systems analysis, systems architecture and systems engineering. If the broader topic of product development "blends the perspective of marketing, design, and manufacturing into a single approach to product development," then design is the act of taking the marketing information and creating the design of the product to be manufactured. Systems design is therefore the process of defining and developing systems to satisfy specified requirements of the user.

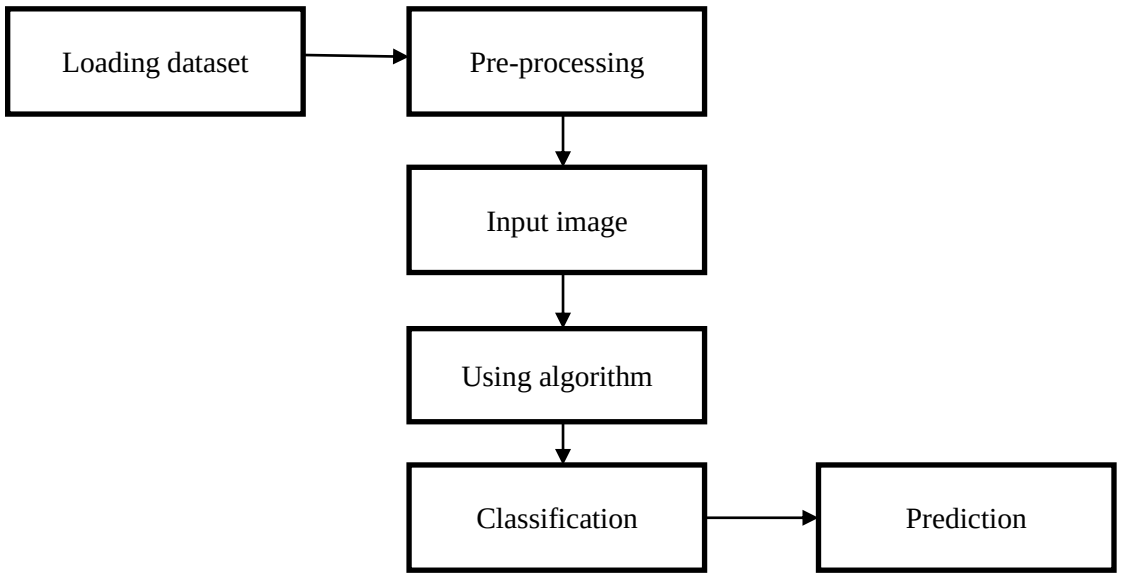


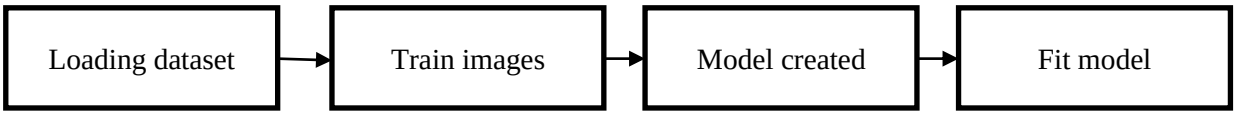
Fig. 3- Training Dataset

DATA FLOW DIAGRAM

A data flow diagram is a graphical representation of the "flow" of data through an information system, modeling its process aspects. Often, they are a preliminary step used to create an overview of the system which can later be elaborated. DFDs can also be used for the visualization of data processing (structured design). The DFD is

also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of the input data to the system, various processing carried out on these data, and the output data is generated by the system.

LEVEL-1



LEVEL-2

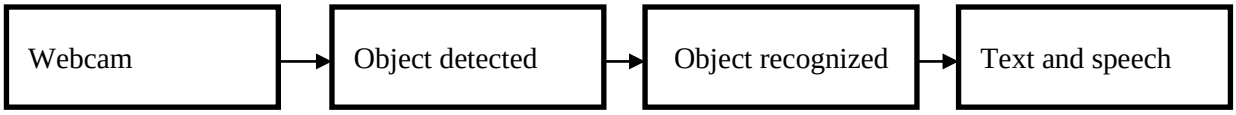


Fig. 4- Training Image

USE CASE DIAGRAM

UML Diagrams

Unified Modeling Language (UML) is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created, by the Object Management Group.

Use Case Diagrams

A use case diagram at its simplest is a graphical representation of a user's interaction with the system and depicting the specifications of a use case. A use case diagram can portray the different types of users of a system and the various ways that they interact with the system.

Flow chart

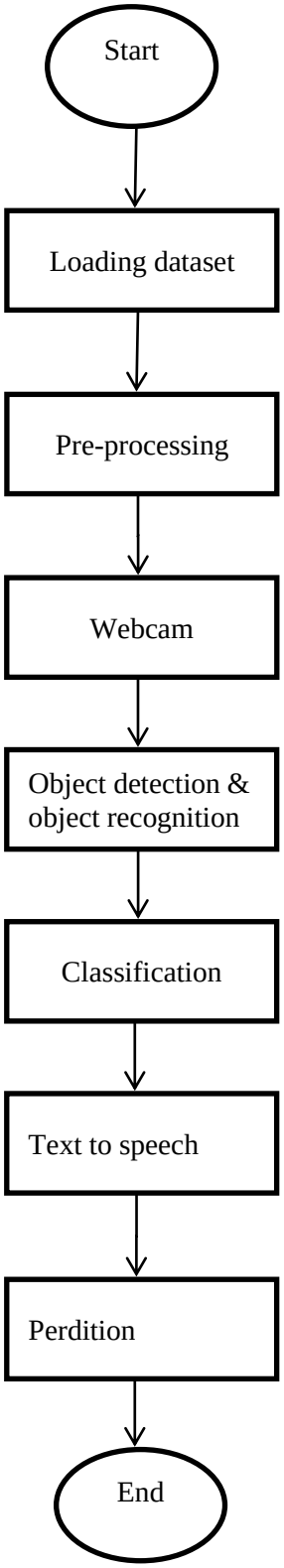


Fig. 5- Flow Chat Diagram

CHAPTER 5

IMPLEMENTATION

Text to speech

Text-to-speech (TTS) is a type of speech synthesis application that is used to create a spoken sound version of the text in a computer document, such as a help file or a Web page. TTS can enable the reading of computer display information for the visually challenged person, or may simply be used to augment the reading of a text message. Current TTS applications include voice-enabled e-mail and spoken prompts in voice response systems. TTS is often used with voice recognition programs. Like other modules the process has got its own relevance on being interfaced with, where Raspberry Pi finds its own operations based on image processing schemes. So once image gets converted to text and thereby it could be converted from text to speech. Character recognition process ends with the conversion of text to speech and it could be applied at anywhere.

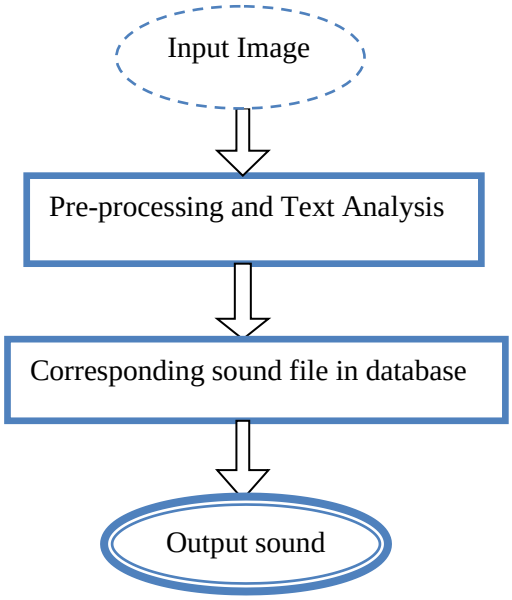


Fig. 6- Text to Sound Conversion

Algorithms based on classification – they work in two stages. In the first step, we're selecting from the image interesting regions. Then we're classifying those regions

using convolutional neural networks. This solution could be very slow because we have to run prediction for every selected region. Most known example of this type of

algorithms are the Region-based convolutional neural network (RCNN) and their cousins Fast-RCNN and Faster-RCNN.

Another method for converting the text into speech can be through the ASCII values of English letters. By using this method, the coding length can be decreased. There are many Text to Speech converters are there but their performance depends on the fact that the output voice is how much close to the human natural voice. For example, consider a name pretty, it can be a name of a person as well as it can be defined as beautiful. Thus, it depends on how the words are pronounced. Many text to speech engines does not give the proper pronunciation for such words thus combining some voice recordings can give more accurate result. The TTS system converts an English text into a speech signal with prosodic attributes that improve its naturalness. There are many systems which include prosodic processing and generation of synthesized control Parameters. The proposed system provides good quality of synthesized speech.

Yolo algorithm

There are a few different algorithms for object detection and they can be split into two groups:

Algorithms based on regression – instead of selecting interesting parts of an image, we're predicting classes and bounding boxes for the whole image in one run of the algorithm. Most known example of this type of algorithms is YOLO (You only look once**)** commonly used for real-time object detection.

Before we go into YOLOs details we have to know what we are going to predict. Our task is to predict a class of an object and the bounding box specifying object location. Each bounding box can be described using four descriptors:

1. center of a bounding box (**bxby**)

2. width (**bw**)

3. height (**bh**)

4. Value **c** is corresponding to a class of an object (car, traffic lights…).

We've got also one more predicted value pc which is a probability that there is an object in the bounding box, I will explain in a moment why do we need this.

Like I said before with YOLO algorithm we're not searching for interested regions on our image that could contain some object. Instead of that we are splitting our image into cells, typically its 19×19 grid. Each cell will be responsible for predicting 5 bounding boxes (in case there's more than one object in this cell). This will give us 1805 bounding boxes for an image and that's a really big number!

Working of Yolo

YOLO trains and tests on full images and directly optimizes detection performance. YOLO model has several benefits over other traditional methods of object detection like the following.

- First, YOLO is extremely fast. Since frame detection in YOLO is a regression problem there is no need of complex pipeline. We can simply run our neural network on any new image at test time to make predictions.
- Second, YOLO sees the entire image during training and testing unlike other sliding window algorithms which require multiple iterations to process a single image.
- Third, YOLO learns generalizable object representations. When trained on real time images and tested, YOLO outperforms top detection methods like DPM and R-CNN.

YOLO network uses features from the entire image to predict each bounding box. It also predicts all bounding boxes across all classes for an image simultaneously.

This means our network reasons globally about the full image and all the objects in the image. The YOLO design enables end-to-end training and real time speeds while maintaining high average precision.

- First YOLO divides the input image into an S × S grid as shown in fig. 7.

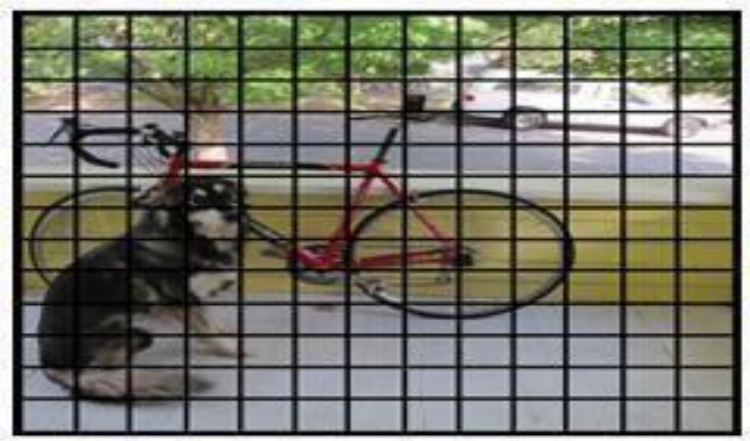


Fig. 7- Divide the image into S × S grid

- If the center of an object falls into a grid cell, that grid cell is responsible for detecting that object.
- Each grid cell predicts B bounding boxes and confidence scores for those boxes as shown in fig. 8.
- This confidence scores reflect how confident the model is that the box contains an object. If no object exists in that cell, the confidence scores should be zero

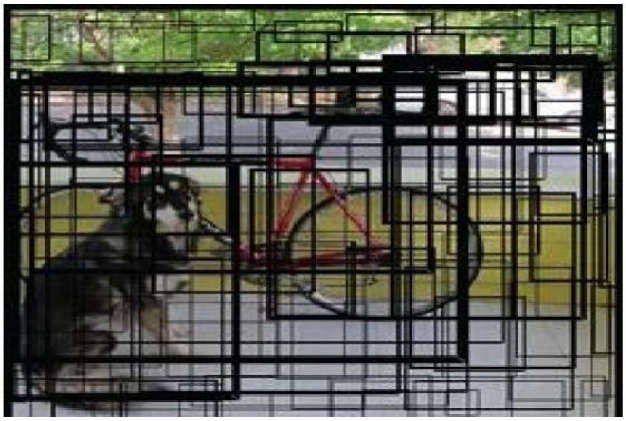


Fig. 8- Calculate Bounding Boxes and confidence score for each box.

- Each grid cell also predicts conditional class probabilities.

- Finally, we multiply the conditional class probabilities as shown in fig. 9 and the individual box confidence predictions which gives us class-specific confidence scores for each box as shown in fig.10.

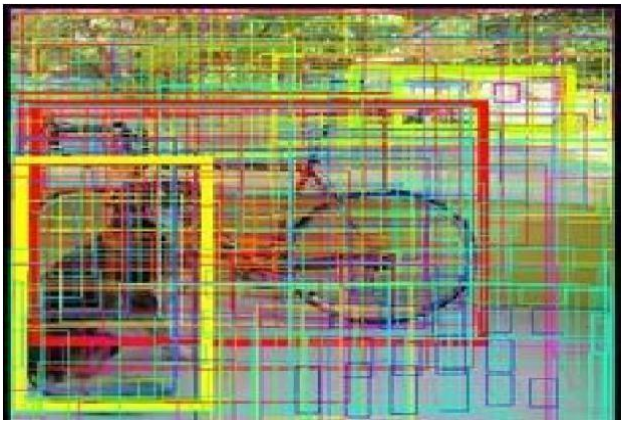


Fig. 9- Multiply Probability and Confidence Scores.

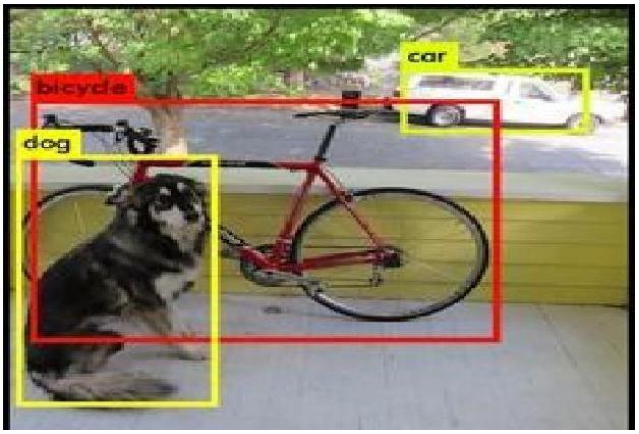


Fig. 10- Final Output

WEB SCRAPING AND TEXT TO SPEECH CONVERSION

Web scraping is a technique that is used to retrieve the content from websites. It consists of two phases namely fetching the web page and later extracting the required content from it. Here two types of web scraping are done one is extracting the content from

Wikipedia and other is top google search links for that label. The required modules are installed to system using pip.

A. Content Retrieval from Wikipedia

After detecting the object from image will use that labelled class to retrieve data from Wikipedia. It is a free encyclopedia in web. So, by extracting data from Wikipedia helps the user to get a idea about what the object is and its uses. Wikipedia is a python library that will help to access and extract data from Wikipedia. In that module with a help of a predefined function Summary (), label (object name) and filter (no of lines from Wikipedia) are arguments for this function and returns a string that contains the extracted data.

B.URL Retrieval from Google

By using the label (object name) will extract top google URL's from google with the help of python module Google search. By using pre-defined function called Search () will extract the required URL's. In this function we can pass arguments like label (object name), no of links need to be extracted etc. With these links they can refer more about the object other than Wikipedia content.

C. Text to Speech Conversion

This step will convert the label (object name) and Wikipedia content to voice so that everybody can understand better. The module used for text to speech conversion is pyttsx which is platform independent and it can convert in offline too. But pyttsx is supported only in python 2.x versions so pyttsx3 module can used in both python 2.x and 3.x versions. In order to use pyttsx3 init () function need to be called to initialize the process and use a predefined method say() with argument text which is platform independent and it can convert in offline too. But pyttsx is supported only in python 2.x

versions so pyttsx3 module can used in both python 2.x and 3.x versions. In order to use pyttsx3 init () function need to be called to initialize the process and use a predefined method say() with argument text which needs to be converted to voice. Finally use run and Wait() to run the speech.

Performance Analysis

To analyze the performance of YOLO, it compared with algorithms like RCNN, fast R-CNN, faster R-CNN on various performance measures like time taken, accuracy

and the frames per second. When analysis was done based on time taken by the algorithm to detect the objects as listed in table 1, it is found that R-CNN takes around 40 to 50 seconds, fast R-CNN takes 2 seconds, faster R-CNN takes 0.2 seconds, and YOLO takes just 0.02 seconds. From this analysis it can be inferred that,

YOLO performs 10 times quicker that faster R-CNN, 100 times quicker than fast RCNN and more than 1000 times quicker than R-CNN.

Algorithm	**Time taken** (in sec)
R-CNN	40-50
Fast R-CNN	2
Faster R-CNN	0.2
YOLO	0.02

Table 1: Performance Evaluation Based on Time Taken

When analysis was done based on the number of frames per second, YOLO performs far better than all other algorithms as shown in Fig. 11, with 48 fps whereas, RCNN processes 2 fps, fast R-CNN processes 5 fps and faster R-CNN processes 8 fps.

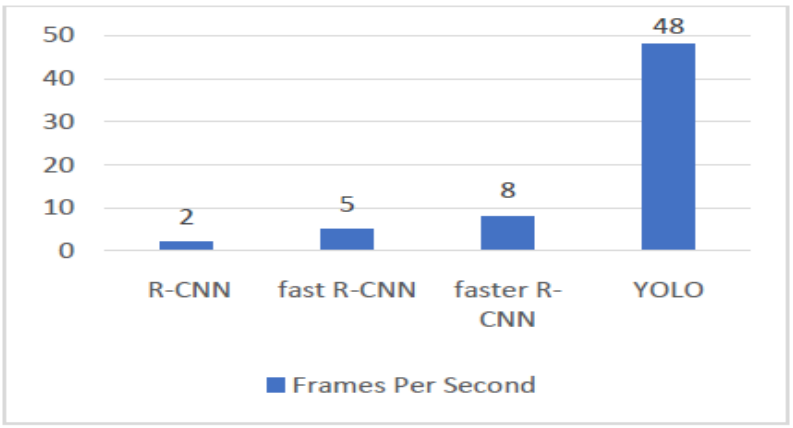


Fig. 11-Performance Analysis Based on frames per second

When analysis was done based on the accuracy it is found that YOLO has lesser accuracy than the other three algorithms as shown in fig.12. So, it is not recommended to use YOLO for applications in which accuracy is the major concern.

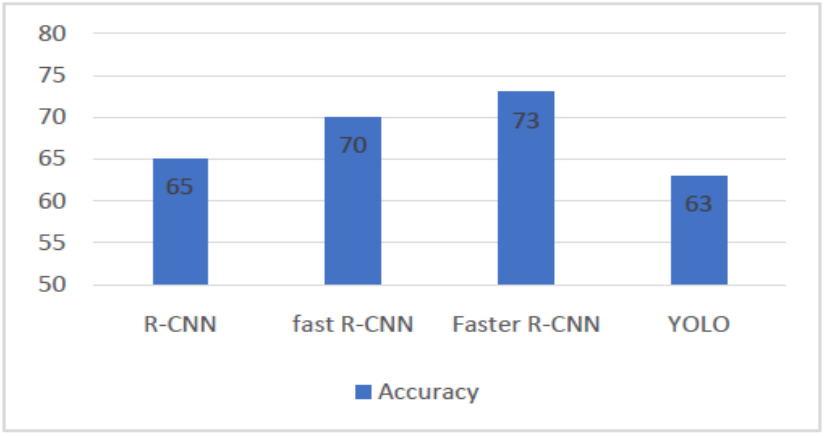


Fig. 12-Performance Analysis based on Accuracy

The model can be used in tracking objects for example tracking a ball during a football match, tracking movement of a cricket bat, tracking a person in a video, Video surveillance, Smart Class for students, Instructor for blind people to get details about unknown objects. It is also used in Pedestrian detection.

Properties

- **Face detection:** An example of object detection in daily life is that when we upload a new picture in Facebook or Instagram it detects our face using this method.
- **People Counting:** Object detection can be also used for people counting, it means that it is used for analyzing store performance or crowd statistics during festivals where the people spend a limited amount of time and other details .This type of analysis is little difficult as people move away from frame.
- **Vehicle detection:** When the object is a vehicle such as a bicycle or car or bus, object detection with tracking can prove effective in estimating the speed of the object. The type of ship entering a port can be determined by object detection based on the shape, size etc. This method of detecting ships has been developed in certain European Countries.
- **Manufacturing Industry:** Object detection is also used in industrial processes to identify products. If we want our machine to detect products which are only circular, we can use Hough circle detection transform can be used for detection
- **Online images:** Apart from these object detections can be used for classifying images found online. Obscene images are usually filtered out using object detection.
- **Security:** In the future we might be able to use object detection to identify anomalies in a scene such as bombs or explosives (by making use of a quadcopter).
- **Medical Diagnose:** Use of object detection and recognition in medical diagnose to detect the X-Ray report, brain tumors.

CHAPTER 6

RESULTS & DISCUSSION

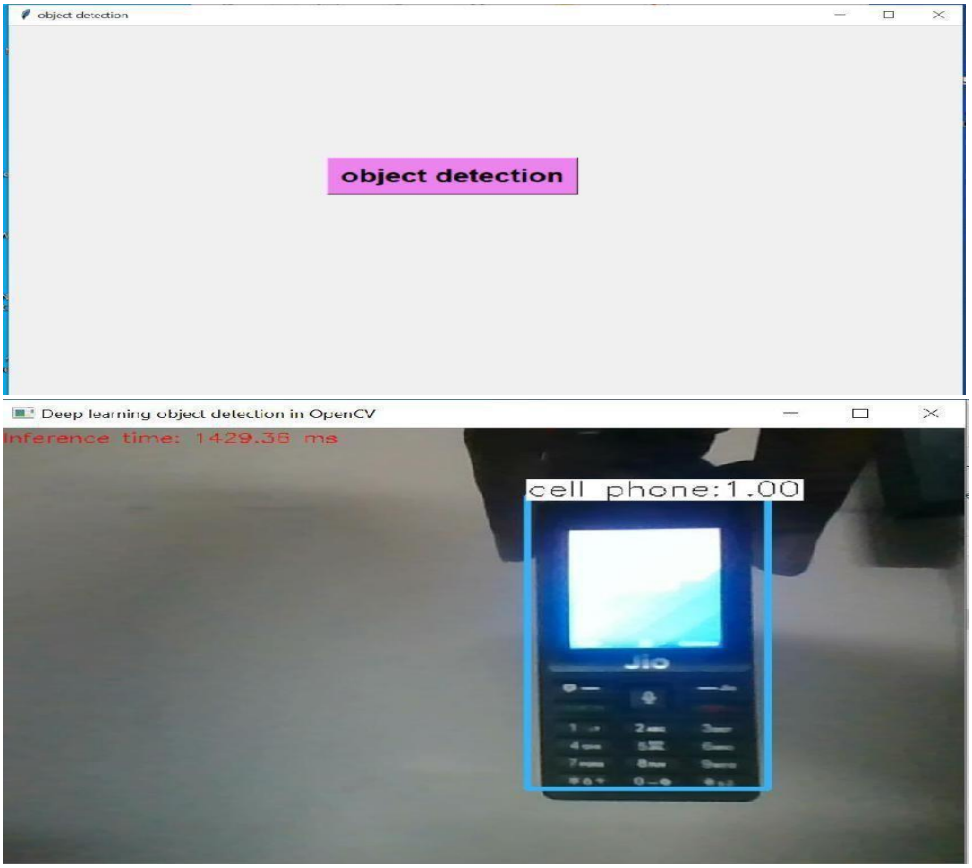


Fig. 13- Output Image 1

Fig. 14- Output Image 2

The above Fig.13 shows that a cellphone is detected by YOLO with a full confidence of 1 and the text is shown (Fig.14) which is then converted to speech voice.

CHAPTER 7

TESTING

Test Case ID	Unit Test Case 1
Description	Object detection page
Input	Images
Expected output	Detection of the object
Actual Result	Got the expected output

Test Case ID	Unit Test Case 2
Description	Using voice-based object detection
Input	Webcam on
Expected output	Vocal representation of the object that is given as an input data through webcam
Actual Result	Got the expected output

Test Case ID	Unit Test Case 3
Description	Using voice-based object detection
Input	Webcam on
Expected output	Test representation of the object that is given as an input data through webcam
Actual Result	Got the expected output

Table 2-Test Cases

Levels of Testing

Testing can be done in different levels of SDLC. They are:

Unit Testing

The first level of testing is called unit testing. Unit testing verifies on the smallest unit of software designs-the module. The unit test is always white box oriented. In this, different modules are tested against the specifications produced during design for the modules. Unit testing is essentially for verification of the code produced during the coding phase, and hence the goal is to test the internal logic of the modules.

It is typically done by the programmer of the module. Due to its close association with coding, the coding phase is frequently called “coding and unit testing.” The unit test can be conducted in parallel for multiple modules.

Integration Testing

The second level of testing is called integration testing. Integration testing is a systematic technique for constructing the program structure while conducting tests to uncover errors associated with interfacing. In this, many tested modules are combined into subsystems, which are then tested. The goal here is to see if all the modules can be integrated properly.

There are three types of integration testing:

- **Top-Down Integration:** Top down integration is an incremental approach to construction of program structures. Modules are integrated by moving downwards throw the control hierarchy beginning with the main control module.
- **Bottom-Up Integration:** Bottom up integration as its name implies, begins Construction and testing with automatic modules.
- **Regression Testing:** In this contest of an integration test strategy, regression testing is the re execution of some subset of test that have already been conducted to ensure that changes have not propagated unintended side effects.

CHAPTER 8

CONCLUSION

We introduce YOLO, a unified model for object detection. Our model is simple to construct and can be trained directly on full images. Unlike classifier-based approaches, YOLO is trained on a loss function that directly corresponds to detection performance and the entire model is trained jointly. Fast YOLO is the fastest general-purpose object detect-tor in the literature and YOLO pushes the state-of-the-art in real-time object detection. YOLO also generalizes well to new domains making it ideal for applications that rely on fast, robust object detection and text to speech is also done.

Future Scope

Object detection is a key ability for most computer and robot vision system. Although great progress has been observed in the last years, and some existing techniques are now part of many consumer electronics (e.g., face detection for auto-focus in smartphones) or have been integrated in assistant driving technologies, we are still far from achieving human-level performance, in particular in terms of open-world learning. It should be noted that object detection has not been used much in many areas where it could be of great help. As mobile robots, and in general autonomous machines, are starting to be more widely deployed (e.g., quad-copters, drones and soon service robots), the need of object detection systems is gaining more importance. Finally, we need to consider that we will need object detection systems for nano-robots or for robots that will explore areas that have not been seen by humans, such as depth parts of the sea or other planets, and the detection systems will have to learn to new object classes as they are encountered. In such cases, a real-time open-world learning ability will be critical.

REFERENCES

[1] Shinde, Shweta S., Rajesh M. Autee, and Vitthal K. Bhosale. "Real-time two-way communication approach for hearing impaired and dumb person based on image processing." Computational Intelligence and Computing Research (ICCIC), 2016 IEEE International Conference on.IEEE, 2016.

[2] Shangeetha, R. K., V. Valliammai, and S. Padmavathi. "Computer vision-based approach for Indian Sign Language character recognition." Machine Vision and Image Processing (MVIP), 2018 International Conference on. IEEE, 2018.

[3] Sood, Anchal, and Anju Mishra. "AAWAAZ: A communication system for deaf and dumb." Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2016 5th International Conference on. IEEE, 2016.

[4] Ahire, Prashant G., et al. "Two Way Communicator between Deaf and Dumb People and Normal People." Computing Communication Control and Automation (ICCUBEA), 2017 International Conference on. IEEE, 2017.

[5] Ms R. Vinitha and Ms A. Theerthana. "Design And Development Of Hand GestureRecognition System For Speech Impaired People.",2018